Design Science

AI-driven FMEA: integration of large language models for faster and more accurate risk analysis

Ibtissam El Hassani[1,2], Tawfik Masrour[1,2], Nouhan Kourouma[1] and Jože Tavčar[3]

[1]*Laboratory of Mathematical Modeling, Simulation and Smart Systems (L2M3S), University of Moulay Ismail, ENSAM, Meknes, Morocco*
[2]*Mathematics, Computer Science and Engineering Department, University of Quebec at Rimouski, Rimouski, Canada*
[3]*Design Sciences, Innovation, Lund University, Lund, Sweden*

Abstract

Failure mode and effects analysis (FMEA) is a critical but labor-intensive process in product development that aims to identify and mitigate potential failure modes to ensure product quality and reliability. In this paper, a novel framework to improve the FMEA process by integrating generative artificial intelligence (AI), in particular large language models (LLMs), is presented. By using these advanced AI tools, we aim to streamline collaborative work in FMEA, reduce manual effort and improve the accuracy of risk assessments. The proposed framework includes LLMs to support data collection, pre-processing, risk identification, and decision-making in FMEA. This integration enables a more efficient and reliable analysis process and leverages the strengths of human expertise and AI capabilities. To validate the framework, we conducted a case study where we first used GPT-3.5 as a proof of concept, followed by a comparison of the performance of three well-known LLMs: GPT-4, GPT-4o and Gemini. These comparisons show significant improvements in terms of speed, accuracy, and reliability of FMEA results compared to traditional methods. Our results emphasize the transformative potential of LLMs in FMEA processes and contribute to more robust design and quality assurance practices. The paper concludes with recommendations for future research focusing on data security and the development of domain-specific LLM training protocols.

Keywords: FMEA–failure mode and effects analysis, Generative artificial intelligence, LLM–large language model, Product quality, Knowledge management

Received 24 October 2024
Revised 18 March 2025
Accepted 20 March 2025

Corresponding author
Jože Tavčar
joze.tavcar@design.lth.se

Des. Sci., vol. 11, e10
journals.cambridge.org/dsj
DOI: 10.1017/dsj.2025.7

1. Introduction

Failure mode and effects analysis (FMEA) has been a cornerstone of the product development process (PDP) for decades, providing a systematic approach to identifying potential failure modes and their effects. By proactively addressing these risks during the engineering design phase, FMEA plays a crucial role in ensuring product quality, reliability and customer satisfaction. Despite its importance, the traditional manual execution of FMEA is labor-intensive, prone to human error and often insufficient for a comprehensive analysis of complex designs, highlighting the need for more efficient and accurate methods.

Recent advances in generative artificial intelligence (AI) offer promising solutions to these challenges. By integrating AI into the FMEA process, it becomes possible to automate the identification of failure modes, streamline risk assessment and improve the overall reliability of PDP. Large language models (LLMs) such as ChatGPT have shown great potential in this domain, demonstrating their ability to extract, process and generate valuable data from diverse sources, including historical FMEA reports, product history files, formal complaints and customer reviews (Zhao *et al.* 2023). These capabilities significantly reduce manual effort, minimize errors and enhance the robustness of designs (Dell'Acqua *et al.* 2023).

While LLMs can efficiently handle knowledge-intensive tasks with prompt engineering, their application in FMEA requires specialized tools and robust data management systems. This research proposes a comprehensive framework that integrates LLMs into the FMEA process and includes a process model and an information system model that supports data collection, pre-processing, risk identification and decision-making. The authors of this paper prepared the framework following the guidelines of Gericke *et al.* (2020), who argue that new methods should be developed to a point where the industry can use them alongside existing methods without requiring the active involvement of method creators.

The key contributions of this research are as follows:

- Development of the framework: A novel process and information system model that integrates LLMs into FMEA, enhancing automation and collaboration between AI and human expertise.
- Case study validation: Validation of the framework through a comparative study of GPT-3.5, GPT-4, GPT-4o and Gemini 1.5 FLASH, demonstrating significant improvements in analysis speed, accuracy and reliability.

Our findings highlight the transformative potential of LLMs in improving FMEA processes and contribute to more robust design and quality assurance practices. The proposed model involves a human-in-the-loop approach, where the results generated by LLMs are validated and used as input for corrective actions. These models were compared with expert analysis and previous proof-of-concept results using GPT-3.5 (El Hassani *et al.* 2024), demonstrating that LLMs provide scalable, fast and effective semantic analysis for large datasets. Fully automated processing would require further testing to ensure the reliability of the results. While the case study demonstrates the benefits of LLMs in reducing manual effort and improving accuracy, it also highlights areas that require further refinement and research.

This paper begins with a literature review on the integration of AI into FMEA and PDP, explores the benefits and limitations of using LLMs in this context, and presents a proposed framework for integrating LLMs into FMEA. The framework is validated through a case study, concluding with lessons learned and recommendations for future research.

2. Literature review

2.1. Historical evolution of FMEA

FMEA emerged in the 1950s and 1960s as the aerospace and defense industries prioritized the identification of potential failure modes in complex systems. The US Department of Defense formalized it with the MIL-STD-1629A standard, and

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

NASA integrated it into mission-critical processes to ensure system reliability and safety. These early efforts highlighted the need for cross-industry standardization.

In response, the International Electrotechnical Commission (IEC) published IEC 60812 in 1985, followed by SAE J1739 in 1994, which formalized FMEA guidelines. Today, standards like IATF 16949:2016 mandate their use to ensure product safety and quality, particularly in the automotive industry (Huang *et al.* 2019). Variants such as design FMEA, process FMEA and system FMEA have since evolved to meet specific requirements within PDPs (AIAG and VDA 2019; Soltanali & Ramezani 2023).

Traditional FMEA, though effective, requires meticulous documentation and analysis, making it resource-intensive (Tavčar & Duhovnik 2014; Thomas 2023). The challenges in managing and reusing company-specific knowledge have driven the development of computerized tools and automation techniques to improve efficiency. Despite progress, face-to-face contacts remain critical for knowledge sharing and complement IT tools (Dai *et al.* 2020).

Practical tools such as the Engineering Checksheet improve knowledge reuse, especially for inexperienced engineers, by providing structured and visualized descriptions (Stenholm, Catic & Bergsjö 2019). Recent advances in AI, particularly generative models, have opened up new opportunities for automating FMEA processes that improve risk assessment and reduce human effort (Wu, Liu & Nie 2021).

2.2. AI-driven enhancements for FMEA

Numerous studies have explored AI-driven methods for improving FMEA and the associated risk assessment processes. Wirth *et al.* (1996) were among the early proponents of a knowledge-based approach to FMEA. They suggested that the use of various knowledge bases with controlled vocabularies could improve the accuracy of product descriptions and facilitate the reuse of knowledge acquired during FMEA.

In recent years, advances in AI have opened up new opportunities to improve FMEA. Liu *et al.* (2019) discussed how multi-criteria decision making methods could support risk assessments within FMEA. Soltanali & Ramezani (2023) presented an intelligent FMEA platform that integrates uncertainty quantification, machine learning and multi-criteria decision making to create hybrid FMEA models. Na'amnh *et al.* (2021) showed that risk assessment models using fuzzy inference and neural networks outperform traditional methods, with the fuzzy model having better decision-making capabilities.

Researchers have also focused on data-driven strategies by using machine learning to continuously update and predict risk priority numbers (RPNs) for emerging failure modes (Peddi, Lanka & Gopal 2023). Hassan *et al.* (2023) used historical data and convolutional neural networks (CNNs) to automate the prioritization of contract requirements, while Yucesan, Gul & Celik (2021) applied fuzzy best-worst and fuzzy Bayesian network methods to evaluate risk parameters in FMEA. The benefits of combining data from past maintenance events with employee expertise were highlighted by Filz *et al.* (2021) to improve maintenance planning. Furthermore, Hodkiewicz *et al.* (2021) utilized ontological methods to improve the explicit representation of FMEA concepts.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Sakwe *et al.* present an objective study of an FMEA method for investigating failure risks in high-performance product service systems (PSSs). This method can provide insights into critical failures and provides a basis for design improvements (Sakwe, Pereira Pessoa & Hoekstra 2021).

The application of LLMs, in particular ChatGPT, to the FMEA process has attracted considerable interest. The contextual understanding of ChatGPT and its ability to learn from new data offer potential benefits for FMEA tasks (Thomas 2023). Diemert & Weber (2023) emphasize that the use of ChatGPT in FMEA involves leveraging its core capabilities while integrating company-specific knowledge.

The synergy between AI tools such as ChatGPT and human expertise can improve the FMEA process. However, studies combining FMEA with LLM techniques are still limited. Spreafico & Sutrisno (2023) investigated the use of a chatbot for automated social failure analysis in product sustainability and demonstrated the potential and limitations of the method using three case studies. Finally, the broader impact of digitalization on PDPs is also noteworthy, as it forces companies to adapt to agile and digital workflows, with both operational and managerial consequences (Cantamessa *et al.* 2020).

The literature reviewed emphasizes the importance of integrating AI into FMEA to achieve better results. While significant progress has been made, gaps remain. Many studies have focused on specific AI techniques for FMEA, but a comprehensive framework that systematically applies AI throughout the FMEA process is lacking. Furthermore, the practical challenges of implementing AI in FMEA are often overlooked. Therefore, there is a need for case studies that demonstrate the application of LLMs in different FMEA phases in different contexts.

To address these gaps, this research proposes the development of a framework (comprising a process model and an information system model) that integrates generative AI, specifically LLMs, into the FMEA process and illustrates their practical implementation through a case study.

2.3. Generative AI and LLM

Generative AI refers to a subset of AI models that are able to generate new data instances that are similar to the training data. Unlike traditional AI systems that perform classification or prediction tasks based on existing data, generative models create new content, such as text, images or music, by learning the underlying patterns and structures in the training data. Techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and transformer-based models are widely used in generative AI. These models have a wide range of applications, including content creation, data augmentation and simulation of environments for training other AI systems (Kingma & Welling 2013; Goodfellow *et al.* 2014).

LLMs are a type of generative AI specifically designed to understand and generate human language. These models are trained on large amounts of text data and can perform a variety of tasks such as translation, summarization, question answering and text completion. Examples of LLMs include OpenAI's GPT-4, Google's Gemini and the open-source GPT-Neo model. They work by predicting

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

the next word in a sequence so that they can produce coherent and contextually relevant text based on the input received (Brown *et al.* 2020; Rae *et al.* 2022).

There are several techniques to work with LLMs, including prompting, fine-tuning and Retrieval-Augmented Generation (RAG). Here you will find a brief explanation of each technique:

a) Prompt Engineering

Prompt engineering involves the creation of effective prompts to elicit the desired responses from LLMs. Since LLMs respond to the input text provided, the way a prompt is structured can significantly affect the output of the model. Effective prompt engineering can improve the accuracy and relevance of the content generated. Techniques include using specific keywords, providing detailed context and iteratively refining prompts based on the model's responses (Liu *et al.* 2021).

b) Retrieval-Augmented Generation

RAG combines the strengths of retrieval-based and generative models to improve the accuracy and relevance of AI-generated content. In a RAG system, a retrieval component first searches a large corpus of documents to find the most relevant information based on the input query. This information is then passed to a generative model that produces a coherent and contextually appropriate response. This hybrid approach leverages the extensive knowledge stored in specific databases while preserving the generative model's ability to produce fluent and natural language responses (Lewis *et al.* 2020).

c) Fine-Tuning

Fine-tuning is the process of further training a pre-trained generative model on a particular dataset to adapt it to a specific task or domain. This approach utilizes the general language understanding capabilities of large, pre-trained models and refines them to perform better on specific tasks. Fine-tuning involves adjusting the parameters of the model based on the new data, improving its performance and relevance to the intended application (Howard & Ruder 2018; Radford *et al.* 2019).

LLMs are increasingly explored in product development for tasks such as knowledge extraction, idea generation and decision-making areas that align with key objectives of the FMEA process. Their ability to process large amounts of data and derive actionable insights is inspiring ongoing research to improve structured engineering methods such as FMEA. For example, patents and scientific articles, which are rich in design knowledge, serve as valuable resources for building knowledge graphs to extract relevant information (Siddharth *et al.* 2021; Siddharth, Blessing & Luo 2022). While patents use a standardized language suitable for rule-based methods, scientific articles often require a combination of rule-based, ontology-based and supervised techniques for effective knowledge extraction.

Research has highlighted how structured models improve design-specific tasks. For example, Giordano *et al.* (2024) have demonstrated the importance of semantic relationships in engineering design processes, while Wang *et al.* (2023) have shown that structured models, such as function-behavior-structure models, perform better than free-form specifications in generating creative and feasible ideas. These structured approaches resonate with the requirements of FMEA, where the

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

identification of relationships between components and potential failure modes is crucial.

Furthermore, Ehring *et al.* (2024) pointed out the importance of domain-specific training for LLMs to improve classification accuracy and relevance in technical domains – a finding that is relevant for the adaptation of LLMs for FMEA. Similarly, research by Mas'udah & Livotov (2024) and Gomez *et al.* (2024) has shown that well-designed prompting strategies can provide effective solutions to complex engineering challenges, comparable to traditional methods. Sarica & Luo (2024) point out that while AI can expand the technology space by generating new technological concepts, this expansion raises the bar for future inventors by increasing the knowledge required for design originality.

Practical applications have demonstrated the strengths and limitations of LLMs. For example, Girotra *et al.* (2023) showed that GPT-4 can efficiently generate high-quality design concepts in collaboration with human designers. While LLMs show promise in idea generation and optimization, studies such as those by Ege *et al.* (2024) and Meron & Tekmen (2023) emphasize that human supervision remains crucial in tasks that require high levels of creativity or technical accuracy. These findings emphasize the need for strategic integration of LLMs into processes such as FMEA, where automation can complement but not replace human expertise.

The above findings illustrate the potential of LLMs to support structured decision-making, like in FMEA, by extracting and organizing relevant data, assessing risks and suggesting solutions. However, achieving these results requires a tailored approach that leverages the strengths of LLMs while taking into account their limitations.

2.4. Benefits and challenges of implementing LLMs in the FMEA process

Drawing from the general advantages and limitations of LLMs as discussed in existing literature (e.g. Bommasani *et al.* 2021; Hu *et al.* 2023; Thirunavukarasu *et al.* 2023) and combining these insights with the authors' expertise in both LLMs and FMEA, we have compiled a comprehensive set of advantages and limitations specific to the application of LLMs in the FMEA process.

2.4.1. Benefits and contributions of LLMs to the FMEA process

The integration of LLMs into FMEA brings several potential advantages and contributions to risk analysis within the PDP. These benefits include:

- Knowledge and expertise: LLMs can be trained on extensive datasets of technical and engineering information. This enables them to provide accurate and up-to-date knowledge of FMEA methodologies, best practice and industry standards. By utilizing LLM's knowledge base, engineers can gain valuable insight into FMEA concepts, processes and techniques, improving their understanding and application of these methods.
- Data analysis support: LLMs are able to support engineers in analyzing and interpreting data relevant to FMEA. They can assist in pre-processing data, identifying patterns and uncovering correlations within the data. This capability is particularly valuable when it comes to extracting failure-related information

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

from text data, which is crucial for identifying potential failure modes. The ability of LLMs to efficiently process large amounts of data makes them an invaluable tool in the data-intensive FMEA process.
- Continuous improvement and learning: LLMs can continuously learn and improve from new data and feedback, which means that the quality of their results can increase over time. This continuous improvement can lead to more accurate and reliable FMEA analyses as the models adapt to new information and changing conditions.
- Scalability: The ability of LLMs to process large amounts of data quickly and efficiently means that FMEA processes can be scaled to handle more complex systems and larger datasets without a corresponding increase in manual effort. This scalability is particularly beneficial for large organizations with extensive product lines and numerous potential failure modes to consider.
- Cost efficiency: By automating many labor-intensive aspects of FMEA, LLMs can help reduce the costs associated with conducting thorough and accurate risk assessments. These cost efficiencies can make FMEA more accessible and feasible for smaller organizations or projects with limited resources.

By leveraging the capabilities of AI through LLMs, engineers can streamline their FMEA activities, reduce manual effort and improve the overall quality of the analysis. The integration of LLMs not only enhances the efficiency of the FMEA process but also contributes to the development of more robust and reliable products.

2.4.2. Challenges of LLMs in the FMEA process

Although LLMs provide valuable support in the area of FMEA, it is important to recognize their limitations and potential drawbacks. These limitations should be considered when using LLMs in the FMEA process. The following list includes some notable challenges:

- Security concerns: LLMs may have security-related vulnerabilities, such as susceptibility to hostile attacks or privacy issues. Ensuring the security of data and the local LLM model itself is critical when integrating LLMs into FMEA processes, especially as LLMs are often not used locally and information can be transferred to external servers or LLM owners. This requires robust measures to protect the transfer of sensitive data and ensure that external parties, such as LLM owners, have appropriate security practices in place.
- Lack of contextual understanding: LLMs operate based on patterns and associations learned from training data that may not provide a deep understanding of the specific context and nuances of FMEA in different industries or technical domains. This limitation may result in incomplete or inaccurate answers that require careful interpretation and review by subject matter experts (SMEs).
- Potential biases: Like any machine learning model, LLMs can unintentionally produce biased or subjective answers based on the biases present in the training data. These biases can influence the guidance or recommendations provided by LLMs in the FMEA process. It is necessary to critically evaluate these results and compare them with different sources of information to mitigate possible biases.
- Maintenance and updates: LLMs need to be regularly updated and maintained to remain effective. This ongoing requirement can be time-consuming and costly

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

and requires dedicated resources to ensure models remain accurate and up-to--date.
- Over-reliance on automation: Although LLMs can automate many aspects of FMEA, there is a risk of over-reliance on these tools, which can lead to human expertise and judgment being neglected. It is important to find a balance between automated analysis and human oversight.

By recognizing and addressing these challenges, companies can better leverage the strengths of LLMs while mitigating their potential drawbacks in the FMEA process.

3. Methodology

3.1. Systematic framework for integrating LLMs into FMEA

Building on the advantages of integrating LLMs into FMEA, we propose a comprehensive process model and an information system model. This framework is designed to streamline the FMEA process and enhance its effectiveness. The framework was developed based on a comprehensive literature review and the authors' experience. The proposed process model comprises the following steps:

1. **Data collection:** Relevant data is gathered from a variety of sources, including design data, historical failure records and other contextual information. This data forms the foundation for training AI algorithms and provides critical insights for risk analysis.
2. **Data pre-processing**: The collected data undergoes thorough pre-processing to ensure its quality and compatibility with the LLM. This stage may involve data cleaning, normalization, feature extraction and handling of missing values or outliers. The aim is to automate pre-processing with computerized tools to ensure efficiency and accuracy.
3. **Model training:** Various subsets of data (such as previous FMEAs, external reviews, etc.) are labeled with expected outputs (failure modes, effects, risk assessments, corrective actions, etc.).

In this step, we can use prompt engineering, fine-tuning or RAG. In this work, we used prompt engineering techniques to help the LLM generate relevant results. To do this, we had to formulate precise prompts and queries to extract the necessary information from the model. By carefully crafting these prompts, we wanted to improve the model's ability to understand and accurately respond to the specific context and details of the FMEA. Through this process, we evaluated the effectiveness of the prompt engineering in achieving satisfactory accuracy. The goal was to determine if the prompt engineering method alone could make the necessary improvements or if additional steps were needed in the future.

Had fine-tuning or RAG been applied, this would have required formal training of the model, including the use of a loss function (e.g. cross-entropy loss) to guide the optimization process and a parameter tuning strategy (e.g. grid search) to determine the best hyperparameters, like learning rate and batch size. However, since prompt engineering was used in this work, these training-specific components were not required.

After several iterations, the prompt for GPT, shown in Figure 4, was designed. The prompt was the same for all LLMs, but different APIs require some format changes.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

4. **Application – extraction of specific information for FMEA**: Once the system is trained, it is applied to the entire dataset. The LLM then suggests failure modes, effects, risk assessment calculations and corrective actions based on the data.
5. **Integration of LLMs tools into the FMEA process; regular data analyses, improvement of the FMEA process and decision support on the system level:** The generated FMEA information is incorporated into PDP and the Knowledge Management system (KMS). The LLMs tools can produce FMEA reports, visualizations, comprehensive summaries and trends for supporting decision-making. Beyond FMEA, a good overview of data can contribute to the company's quality assurance system, helping to identify competency needs, recurring failure modes, process bottlenecks and more.

The information system model for integrating LLM tools includes data collection, extraction, knowledge management and application within an industrial setting, as depicted in Figure 1. It leverages company-specific knowledge extracted from key documents in the product lifecycle management (PLM) system, such as previous FMEAs, engineering changes (ECs), and product history files (Tavčar, Benedičič & Žavbi 2019). In addition, AI analysis incorporates external sources related to the company's production program and technology-specific information. This is particularly important when developing new products, as external knowledge sources are often crucial. Selected documents are analyzed by AI, and the extracted information is reused in FMEA activities and other PDPs. Systematic data analysis is conducted at regular intervals and leads to the definition of corrective actions for PDP, EC management (ECM), and the FMEA process (Figure 1). LLM tools are exceptionally powerful in analyzing large amounts of data, as they can identify patterns and insights that human analysts might missed. By implementing this framework, organizations can significantly enhance the efficiency and effectiveness

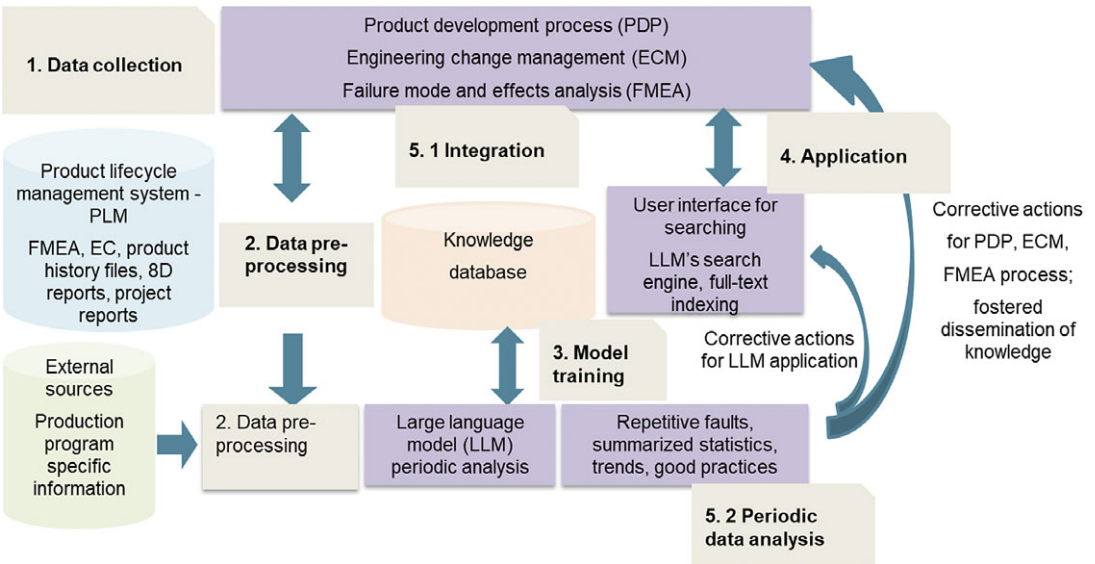

Figure 1. Information system model for LLM application in FMEA (upgraded from El Hassani *et al.* 2024).

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

of their FMEA processes and leverage the strengths of LLMs to improve risk analysis and decision-making.

The implementation of the process model and the information system model within the proposed framework must align with the specific phase of the product life cycle. The maturity level of a product plays a crucial role in determining the optimal support required for the FMEA process. For new products where the company has limited experience, external sources of knowledge are prioritized. These sources provide valuable insights and data that are essential for identifying potential failure modes and accurately assessing risks.

However, for more mature products that have several years of development, manufacturing and sales experience, internally generated knowledge sources become more important. This internally collected data, including historical FMEA reports, ECs and product history files, provides a rich information base that can be used to improve the FMEA process. Utilizing this internal knowledge allows for more accurate and informed risk assessments, as it is based on the company's own experience and lessons learned over time.

3.2. Mapping FMEA challenges to framework solutions

The proposed framework systematically addresses the key challenges in the FMEA process as highlighted in the literature review. Table 1 lists the challenges and the corresponding solutions provided by the framework.

The proposed framework addresses key challenges identified in existing research by using LLMs to automate data pre-processing, failure mode identification and risk

Table 1. Mapping FMEA challenges to solutions.

Challenge	Proposed solution in framework
Manual effort in processing FMEA data	The framework uses LLMs to automate data collection, pre-processing and analysis, significantly reducing the manual workload required in traditional FMEA processes.
Prone to human error	By integrating AI tools, the framework ensures more consistent and accurate identification of failure modes and risk assessments, thereby minimizing the errors associated with human intervention.
Inefficient reuse of knowledge	A dedicated KMS integrated into the framework facilitates the systematic reuse of FMEA results and finding and ensures that past analyses are incorporated into future processes.
Difficult handling of unstructured data	The framework utilizes the advanced natural language processing capabilities of LLMs to handle unstructured data such as customer reviews and informal reports and transform them into structured inputs for FMEA.
Challenges with scalability	The ability of LLMs to process large datasets quickly enables the framework to scale to complex systems and process extensive data inputs without compromising efficiency.
Limited contextual understanding of AI	Through prompt development and fine-tuning, the framework ensures that LLMs deliver contextually relevant and domain-specific results tailored to FMEA tasks.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

analysis to reduce manual effort and minimize human error. To tackle inefficient reuse of knowledge, KMS was integrated to systematically capture and reuse findings from previous analyses. In addition, the framework uses LLMs to effectively process unstructured data and improve data processing and analysis. Scalability and contextual understanding have been improved through prompt development and fine-tuning to ensure adaptability and efficiency in various FMEA tasks. These improvements directly overcome the limitations highlighted in previous research and provide a more robust approach.

4. Practical case study and evaluation

The applicability of the proposed framework is validated by a case study using publicly available data from the automotive industry, in particular data from vehicle reviews by private individuals. This type of data is less structured compared to company FMEA reports and other product-specific documents and therefore poses a greater challenge. This also allows the first steps of the framework process model to be implemented and tested without the security issues that would arise when using proprietary company data. Although this approach has limitations (e.g. the incomplete representation of the company context and the inability to extract corrective actions as it is limited to the design FMEA), it still provides a valuable opportunity to evaluate various aspects such as automatic data preprocessing, model training and information extraction.

In this case study, we aimed to test two key aspects where LLMs can be beneficial for FMEA:

- **Accuracy of information extraction with training:**

The aim was to assess how accurately an LLM can extract relevant information from the data when the model has been trained specifically for the task. For this, we used a model for extracting negative reviews and identifying associated components (the part finder model) as described in Section 4.2 and achieved an accuracy of 98–99%.

- **Accuracy of information extraction without training:**

Here we evaluated the performance of the LLMs in extracting information from the data without specific training. Table 2 shows a summary of the results of a semantic comparison between different LLMs using only prompt engineering. Human-identified error modes (gold standard) were compared with GPT-4o, GPT-4 and Gemini 1.5 FLASH. The comparison included 100 reviews and focused on the similarity scores.

These aspects were tested using the five-step framework process model. First, the preprocessing of the data was tested, then the extraction of the failure modes and finally the relevance of the effects suggested by the LLM. The detailed process flow applied is shown in Figure 2 and the result in Figure 3.

Figure 3 shows an example centered on design FMEA (DFMEA), which is the specific focus of our case study. While DFMEA is a type of FMEA that deals specifically with design-related failures, the underlying principles of risk identification, analysis and mitigation apply broadly to other types of FMEA, including process FMEA (PFMEA). Therefore, the DFMEA example effectively demonstrates the utility of LLMs in systematically analyzing failure modes, which is

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Table 2. Results of semantic comparison between different LLMs and human analysis.

LLM (published)	Average similarity level	Standard deviation	No. of cases similarity = 1; 2	No. of cases similarity = 6; 7; 8; 9; 10
GPT–4o (May, 2024)	8.07	2.34	6	87
GPT–4 (March, 2023)	8.35	2.04	4	91
Gemini 1.5 FLASH (May, 2024)	7.13	2.73	11	80

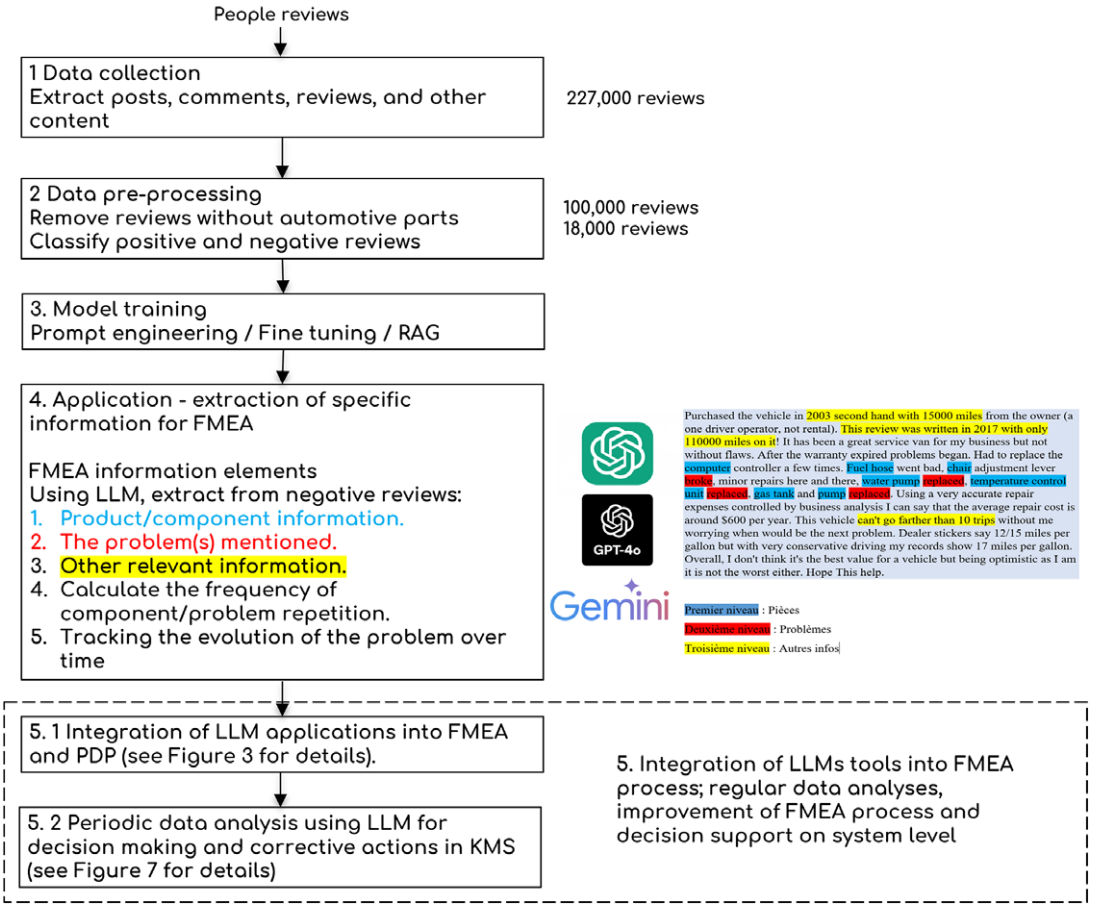

Figure 2. Process flow of the automated FMEA framework.

representative of FMEA activities in general. We believe that this example provides valuable insights that are applicable to a wider range of FMEA scenarios.

For the case study, we initially used GPT-3.5 as a proof of concept to evaluate the feasibility of the approach (El Hassani *et al.* 2024). After validating this concept, we moved on to using GPT-4, GPT-4o and Gemini 1.5 FLASH to enhance the

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Oct. 22, 2023, 3:01 p.m.

Review

We bought our 2014 Santa Fe Limited, brand new, and I was really hoping that this was going to be the vehicle for the long run. It has great features, technology, and storage. Unfortunately, we've had nothing but reliability problems fromthe start . We've put 25k miles on it and owned it for 2 years. The firstissue we had was a crack in the black edge of the windshield which we noticed two or threedays aft erdriving it off the lot . The dealership did right by us and replaced it. Not a big deal, but a year laterthe re placement also developed, what I can only assume to be, a stress crack as well and needs to be replaced. Next, the hood had to be repainteddue to multiple rust spots forming on the corners. Afterthat , the blind spot detection systemfailed and re quiredthe passenger side mirrorto be replaced. Now, during the last snow storm we noticed that the car is leaking anti-freeze and fromtime to time the navigation will go blank and require the car to be turned off and back on again to resolve it. For the most part the warranty has coveredthe issues and out of pocketcost will be the anti -freeze leak and new windshield but these are toomany issues to o fast. We may be the unlucky ones who got the badone out of the bunch but I can only imagine what we'll be dealing with at 100k miles.

Relevant informations

Description	Value
Vehicle Model	2014 Santa Fe Limited
Purchase Condition	Brand New
Mileage	25,000 miles
Duration of Ownership	2 years
Reliability Issues	Multiple problems since the beginning
Windshield Issue	Crackedblack edge , replaced once, developed stress crack again
Rust Spots on Hood	Repainteddue to rust spots forming on the corners
Blind Spot Detection System	Failed, requiredre placement of passenger side mirror
Leaking Anti-freeze	Not coveredby warranty , out of pocketcost
Navigation Issues	Occasionally goes blank, requiresre start of the car

DFMEA TABLE

Component	Potential Failure Modes	Potential Effects	Potential Causes	Current Controls	Severity (S)	Occurrence (O)	Detection (D)	Risk Priority Number (RPN)
Windshield	Cracks, Stress cracks	Reduced Visibility, Safety Concerns	Possible manufacturing defect, Stress	Replacement, Inspection	5	6	8	240
Hood	Rust spots, Corrosion	Aesthetics, Potential Damage	Possible manufacturing defect, Exposure to Moisture	Repainting, Rust prevention measures	4	5	7	140
Blind Spot Detection System	Failure to detect vehicles, Malfunction	Increased risk of accidents, Inaccurate warnings	Possible manufacturing defect, Sensor malfunction	Diagnostic tests, Replacement	7	4	6	168
Anti-freeze System	Leakage	Potential engine damage, Overheating	Component failure, Wear and tear	Regular inspection, Repair/replacement	8	6	5	240
Navigation System	Intermittent failures, Display issues	Loss of navigation functionality	Software glitch, Electrical issue	Software updates, Diagnostic tests	5	3	7	105

DELETE

Figure 3. Example of an automatically generated FMEA table.

performance. We compared these models based on their accuracy and overall effectiveness. The summarized results can be found in Table 2.

The steps involved in the case study are as follows:

1. Data collection

We used a dataset of car reviews from the Kaggle platform (www.kaggle.com), which included fifty different datasets of customer reviews for 50 vehicle brands (AnkurJain 2019), totaling about 227,000 reviews.

2. Data pre-processing

This process was carried out in two main steps:

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

a) Filtering out reviews without identified parts (part finder model)

First, we performed a string comparison to filter out reviews without identified parts. The data was cleaned, formatted and merged into a single dataset. As the FMEA focused on vehicle parts, reviews that contained no references to parts had to be excluded. To facilitate this, we compiled a list of parts by scouring websites such as Wikipedia ("List of auto parts," 2023) and List Explained (Kan 2022). Using a string comparison, ratings without identified parts were removed. However, this method had its limitations (e.g. spelling errors, incomplete data and use of different languages). Therefore, we developed a model using LLMs to extract reviews with parts more accurately. The model was trained with GPT-3 (text-avinci-002) and achieved 98–99% accuracy, including reviews in French and Spanish. This refined dataset included about 100,000 reviews.

To perform fine-tuning, we first formatted our data according to the requirements for fine-tuning a **text-davinci-003** model. Our dataset was streamlined into two columns: a "Prompt" column containing the reviews and a "Completion" column containing the names of the extracted parts determined via string comparison. We added special tokens "- >" at the end of each prompt to indicate that the prompt has concluded and that the model can begin the completion and "END" to signal that the completion is finished and that the model should stop.

Subsequently, we converted our dataset into a JSONL file, which is the required format for fine-tuning an OpenAI model. After preparing the JSONL file, we uploaded it to OpenAI's fine-tuning platform using the OpenAI CLI. We configured specific training parameters, such as the number of epochs and batch size, optimizing them to prevent overfitting. During the fine-tuning process, the model adapted to our dataset, enhancing its performance.

Finally, we evaluated the fine-tuned model by testing it on a separate validation set, measuring its accuracy and observing its behavior on unseen prompts. Upon completion of the fine-tuning process, we obtained a results file from the OpenAI API containing detailed evaluation metrics, including loss, accuracy and other key indicators. These metrics enabled us to assess the effectiveness of the fine-tuning, determining how well the model had adapted to the dataset and whether it was suitable for the task or required further adjustments.

b) Extracting negative reviews (sentiment analysis model)

Next, we focused on identifying negative reviews, as they often contain valuable information about potential problems and concerns related to automotive components and systems. The reviews were divided into negative and positive ratings using an existing labeled dataset (Maas *et al.* 2011) for training. Several deep learning algorithms were tested. TensorFlow's CNN (Abadi *et al.* 2016), which uses bidirectional short-term memory (BiLSTM), achieved an accuracy of 87%. Fine-tuning with GPT-3 (Curie model) improved the accuracy to 97%. The model obtained by GPT-Curie was then used to classify the reviews into negative and positive categories. Only the negative reviews were used to extract failure modes for the FMEA. This pre-processed dataset included about 18,000 reviews, all of which were negative and contained names of automotive parts.

For training the TensorFlow models, we split the dataset into training, testing and validation sets. Specifically, 80% of the data was used for training, with 20% of the training set reserved for validation, while 20% was allocated for testing. We employed cross-validation to optimize hyperparameters. Throughout training, we

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

monitored loss curves to compare validation against training, mitigating overfitting. For certain models, TensorFlow's EarlyStopping method was implemented to stop training early, preventing overfitting or excessive adjustment to the training data.

3. Model training

In this step, we could use prompt engineering, fine-tuning or RAG. In this work, we used prompt engineering techniques to help the LLM generate relevant results. This required us to formulate precise prompts and queries to extract the necessary information from the model. By carefully crafting these prompts, we wanted to improve the model's ability to understand and accurately respond to the specific context and subtleties of the FMEA. Through this process, we evaluated the effectiveness of the prompt engineering in achieving satisfactory accuracy. The goal was to determine if the prompt engineering method alone could make the necessary improvements or if additional steps were needed in the future.

After several iterations, the prompt for GPT, shown in Figure 4, was designed. The prompt was the same for all LLMs, but different APIs require some format changes.

```
{
  "role": "system",
  "content": "You are an assistant tasked with identifying failure modes
and their effects for vehicle parts mentioned in a given review. You
should return a JSON object that contains the part names, failure modes,
and effects for all the parts mentioned in the review. Your response
should follow this format:
{
  "extracted_parts": ["part_1", "part_2", ...],
  "failure_mode": {
    "part_1": ["failure1", "failure2", ...],
    "part_2": ["failure1", "failure2", "failure3", ...],
    ...
  },
  "effect": {
    "part_1": ["effects_failure1", "effects_failure2", ...],
    "part_2": ["effects_failure1", "effects_failure2",
"effects_failure3", ...],
    ...
  }
}

If you are unable to determine the failure mode or effect, please replace
it with 'N/A'. Return the JSON object without any additional comments or
prompts."
},
{
  "role": "user",
  "content": user_prompt
}
```

Figure 4. Prompt for search of failure modes and effects, which is formatted for GPT-4 and GPT-4o.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

4. Extraction of specific information for FMEA and validation of results

A total of 100 reviews were randomly selected using the sentiment analysis model and the part finder model: 20 reviews each for the parts "door," "tire," "seats," "wheel" and "window." The LLM was prompted to extract failure modes as accurately as possible, while also making suggestions for other FMEA information elements. To create the files, we need structured responses (in JSON format) so that we can programmatically retrieve the necessary values and insert them into the Excel file.

Similar to coding methods in qualitative research (e.g. Campbell *et al.* 2013), the failure modes were first analyzed manually and evaluated independently by two experts. They compared their results one-to-one and reached a consensus after discussion to ensure accuracy and minimize bias. As the data were user reviews and not technical reports, the failure modes reflected the users' perspective, which might not always match the technical definitions.

Semantic analysis was used to compare the results of the different LLMs. Semantic analysis is a process in natural language processing (NLP) that focuses on understanding the meaning of words and texts in context. It goes beyond the simple matching of keywords and interprets the actual meaning, relationships and nuances between words, such as synonyms, antonyms or hierarchical relationships. By capturing the context and subtle differences in meaning, semantic analysis helps us to create a more accurate assessment for the FMEA of each LLM.

LLMs outperform traditional NLP methods in semantic analysis because they use deep learning techniques, especially transformer architectures, to dynamically understand context and meaning. Unlike traditional methods that rely on fixed word embeddings and statistical correlations, LLM model words as vectors in a high-dimensional space where their positions are influenced by the words around them, capturing relationships, meanings and nuances more accurately. LLMs achieve this by learning from huge datasets and using attention mechanisms to focus on different parts of a sentence and understand word meanings in context rather than in isolation. This allows them to capture complex language patterns, disambiguate meanings and provide more accurate and flexible interpretations. This is particularly useful when comparing the FMEA provided by each model to the "golden standard," as this requires a deep understanding of nuanced or contextual information (Xu *et al.* 2024).

We compared the results of different LLMs and the failure modes found by experts. The used prompt for the semantic comparison of failure modes is presented in Figure 5.

An example of the numerical semantic analysis is shown in Figure 6, which contains input data, numerical results and reasoning for the evaluation.

The semantic comparison of the failure modes was carried out separately for each review. The results of the semantic analysis were not deterministic. For the upper example from Figure 6, we could obtain a similarity score of 8/10 or 9/10. The variation is within a predictable range and acceptable according to the authors.

Application of LLM for semantic comparison has many advantages. The results of the comparison are understandable and logical for human criteria. The comparison can be performed on a much larger amount of data, which is not possible for humans.

There are several LLMs on the market. One of the objectives was to compare the results between different models for the specific task – the search of failure modes

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

```
 "role": "system",
    "content": "You are an assistant tasked with comparing two sets of
vehicle part failure mode data. Your role is to perform a semantic
comparison between them and provide a score out of 10 based on how
similar they are. You must generate a response in JSON format that
includes a score and an explanation of why that score was given and give
a numerical evaluation from 1 to 10 of how similar they are from only a
semantic point of view, where 1 indicates very different and 10 indicates
very similar in meaning."

    "role": "user",
    "content": "Compare these two paragraphs from only a semantic
perspective and provide a similarity score.\n\n Paragraph1:
{human_failure_modes}\n Paragraph2: {model_failure_modes}"
```

Figure 5. Prompt for semantic comparison of failure modes with GPT-4o.

in reviews. The results of the semantic comparison are summarized in Table 2.

Results of comparison:
Explanation of the semantic comparison parameters:

- **Average similarity level:** Represents how close the failure modes identified by LLM are semantically to the "golden standard" created by humans. This results in an overall performance score for each model.
- **Standard deviation:** Highlights the variability of semantic similarity level between different reviews, where a lower standard deviation reflects greater consistency.
- **Number of cases (similarity = 1; 2):** Indicates instances of significant deviation or poor performance, useful for identifying model weaknesses.
- **Number of cases (similarity = 6; 7; 8; 9; 10):** Indicates the ability of the LLM to achieve high semantic relevance, closely matching the failure modes extracted by humans.

The visual differences when comparing the text output of different LLMs are significant. However, if we perform a deeper comparison on a semantic level, the differences are smaller. GPT-4o and GPT-4 provide more similar results compared to Gemini, and this is in line with expectations. In particular, Gemini represents error modes with different words.

Based on recommendations from the literature (Landis and Koch, 1977) and manual comparisons, a similarity level of 6 or more is considered a substantial agreement. GPT-4 and GPT-4o have a degree of similarity with substantial agreement of 91 and 87%, respectively. Gemini 1.5 is still good with 80%.

Table 2 shows that a large percentage of the ratings have no or only a very low similarity with rating 1 or 2. A manual comparison has shown that no similarity is found if the manual search has found a failure mode but the LLM has not or vice versa.

It can be critical if the failure mode is not recognized in the text. For example, GPT-4o did not recognize a failure mode in the sentence: "No way can you fit 3 car seats in the back row," and the similarity was scored as 1.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Given the comparison between the two paragraphs:

Paragraph 1

{
"engine": ["lack of power"],
"transmission": ["limited to 5 speed"],
"handling": ["poor handling in adverse weather conditions"],
"fuel system": ["small tank capacity"],
"audio controls": ["awkward placement"],
"interior": ["made of plastic"]
}

Paragraph 2

{
"engine": ["low power", "lack of horsepower"],
"transmission": ["lack of 6-speed option"],
"gas tank": ["small capacity"],
"steering wheel": ["audio controls in back"],
"interior": ["all plastic"]
}

Semantic Analysis

• Engine: Both paragraphs describe the engine as having "low power" or a "lack of power," with Paragraph 2 providing additional detail with "lack of horsepower." These are semantically similar.

• Transmission: Paragraph 1 mentions the transmission being "limited to 5 speed," while Paragraph 2 discusses the "lack of 6-speed option." Both essentially describe the same limitation, though in slightly different terms.

• Handling vs. Steering Wheel: Paragraph 1 discusses "poor handling in adverse weather conditions," which is a specific critique of handling. Paragraph 2 mentions "audio controls in back" of the steering wheel, which is a different aspect of the driving experience. These are not semantically similar.

• Fuel System vs. Gas Tank: Both paragraphs mention a "small tank capacity" or "small capacity" in the gas tank, which are directly comparable and semantically identical.

• Audio Controls: Paragraph 1 states "awkward placement" of audio controls, while Paragraph 2 specifies their location as "in back" of the steering wheel. These are somewhat related but not identical.

•Interior: Both paragraphs describe the interior as "made of plastic" or "all plastic," which is semantically identical.

Numerical Evaluation

Given that the two paragraphs describe similar concepts with some slight variations in detail and terminology, they are quite similar overall, though not perfectly identical.

Similarity Score: 8/10

This score reflects a high degree of semantic similarity, as both paragraphs criticize similar aspects of a vehicle, albeit with some differences in emphasis and terminology.

Figure 6. Example of a semantic comparison with numerical evaluation.

On the other hand, in the sentence: "The car does not handle well on rain or snow," the expert did not identify a failure mode, but the LLM did, and therefore the similarity was rated 1. However, from the end user's point of view, it is a problem, and we can agree that the LLM is right.

In some cases, a failure mode was not understandable to the LLM, and the answer was N/A. The expert was able to identify a failure mode.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

GPT-4o recognized a failure mode in the sentence: "There is no window bar between the front and back window," but the expert did not.

A general conclusion is that different LLMs provide useful results. More advanced LLMs such as GPT-4o generally provide better results. However, the results sometimes differ from the expert's pragmatic assessment. This is the reason why the similarity of GPT-4o is worse than the similarity of GPT-4.

The conclusion may be that the search with LLMs is not perfect. However, the search enables a computer-aided search in large datasets. So LLMs enable something that is simply not possible for humans. We can expect better results when searching in professional documents with a more consistent document structure and more precise language.

Finally, the system was asked to generate potential causes, current controls, severity, occurrence and detection for a given failure mode (examples can be found in Figure 3, right). Several suggested values were consistent with the reviews, but no quantitative analysis was performed as there was no specific data to justify the proposed numbers.

5. Integration of LLMs tools into the FMEA process, regular data analyses, improvement of the FMEA process and decision support.

5.1 Integration of the LLM application into the FMEA process.

The supporting tools and methods must be well integrated into the key processes. The data elements obtained from the sentiment analysis model and the parts finder model were systematically organized to facilitate the subsequent FMEA. Using a well-formulated prompt, the LLM generated a comprehensive table of relevant information and an FMEA table for each review. The extracted FMEA information is shown in Figure 3. This illustrates the capabilities of the LLM as an integral part of a KMS.

In the case study, we investigated the identification of failure modes from individual review texts. However, in a more realistic scenario, we may need to analyze failure modes for a given component in multiple reviews simultaneously. This more complex task leads to several considerations regarding the approach and response time.

In the context of searching for all potential failure modes associated with a particular component in a large set of reviews, two possible approaches can be considered:

- **Extraction of all data with subsequent filtering**: In this method, the system first extracts all relevant failure modes from the entire set of reviews. Once the data is collected, it is filtered by component within the FMEA table. While this approach ensures comprehensive coverage of all possible failure modes, it can significantly increase processing time as large amounts of data are processed before filtering by component.
- **Component-specific prompts for targeted extraction**: Another, more targeted approach is to design prompts that explicitly request the extraction of failure modes for a specific component across all documents. By specifying the component in the initial prompt, the LLM can limit its search to the relevant data, which could improve response time and reduce unnecessary processing. This approach could increase efficiency when working with large datasets as it

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

directly targets the information that is relevant to the component under investigation.

There are other strategies that can be used to optimize the process:

- **Batch processing of reviews:** Instead of processing all reviews at once, the data could be split into smaller batches, with each batch focusing on a specific component. This would distribute the computational load and could lead to faster response times while maintaining accuracy.
- **Pre-classification of reviews by component:** A pre-processing step could classify reviews based on the components they mention. This classification would allow the LLM to focus only on the subset of reviews that are relevant to a particular component, further reducing the data load and speeding up the extraction process.

The data used in the case study – end-user reviews – is not reliable data to determine effects, failure mechanisms and probability of severity, occurrence and detection. We were not able to assess the quality of the predictions. However, according to the experts, the answers were appropriate for the given context. At this point, we would like to highlight the technical possibilities of integrating applications and the next steps for industrial implementation.

5.2 Periodic data analysis (LLM) for decision making and corrective actions.

The integration of FMEA into the PDP using LLMs can significantly improve decision support by streamlining information extraction and risk management. In this proposed framework (Figure 1), a knowledge base serves as a central repository for the systematic organization and storage of FMEA results and related documents. This integration provides comprehensive access to information on failure modes and associated risk mitigation strategies, supporting different phases of the product development cycle.

Effective knowledge management is based on continuous learning and system improvement. Regular reviews of product development data and FMEA results by senior engineers are crucial, with corrective action taken where necessary (Figure 1). Obtaining key information can be time-consuming but is essential for informed decision-making and the implementation of effective corrective actions. This is where LLMs offer a powerful solution. Their advanced NLP capabilities enable fast and accurate data extraction, significantly reducing the time and effort required for these tasks.

Although regular data analysis using LLMs is beyond the scope of this paper, we have conducted an example of data analysis to demonstrate the practical application of these tools and illustrate the steps described in the framework (Figure 1). Figure 7 shows a visual representation of the statistical data generated by LLM tools and illustrates their efficiency in processing large volumes of documents. The visual representation of data, such as the frequency of parts involved in customer reviews or the frequency of particular failure modes, can be of great help in long-term decision-making. Only systematic and statistical processing of data can identify trends and show the big picture that is not visible in day-to-day work. Traditional manual methods for searching, collecting and summarizing large datasets are not only time-consuming but also costly. LLMs can effectively automate these tasks,

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

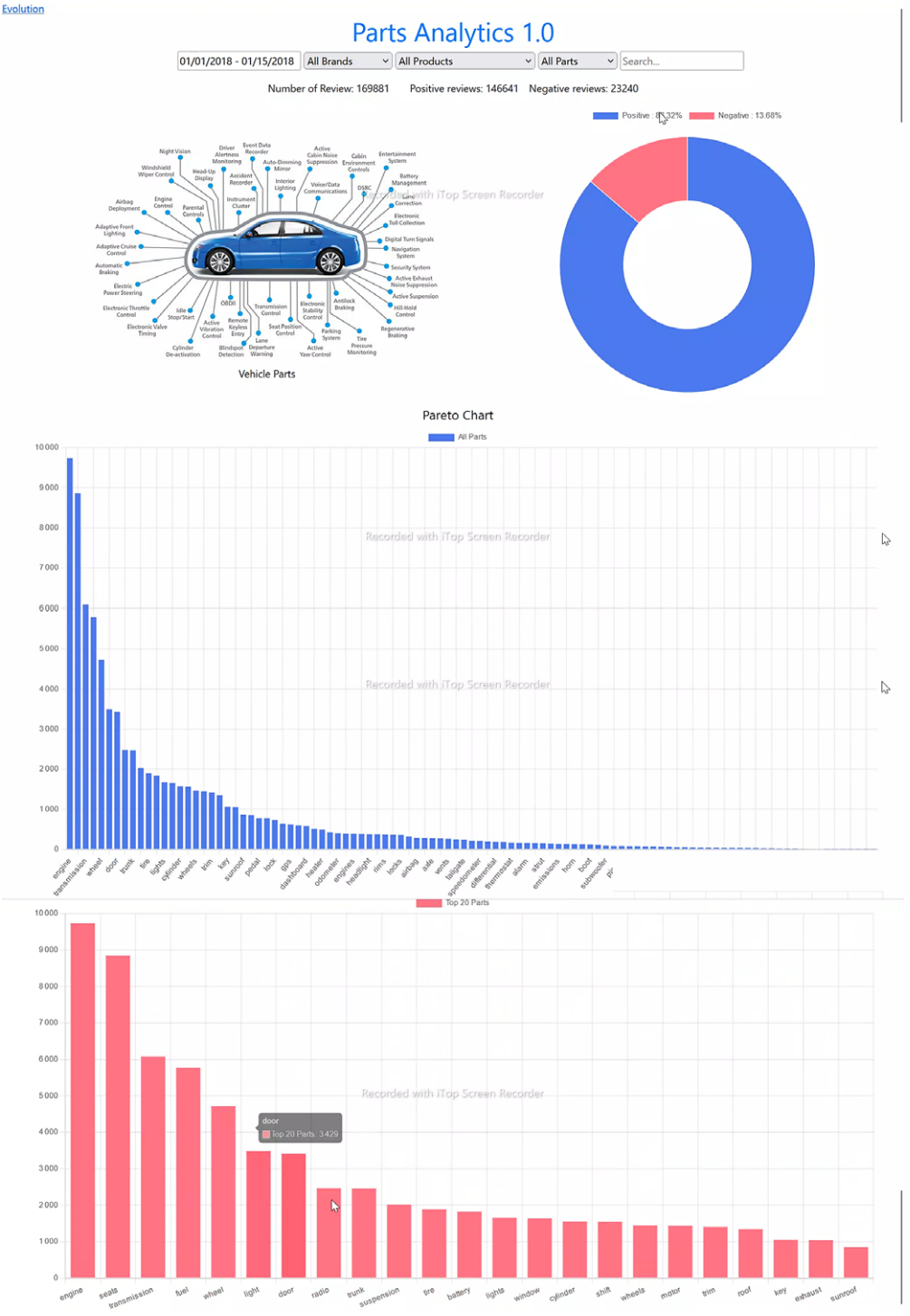

Figure 7. The periodic analysis supported by LLM and visual presentation as a tool for corrective actions.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

enabling continuous data analysis and the improvement of processes and products based on this data.

The ability of LLMs to scale FMEA analyzes across multiple tests and components and integrate the results into a knowledge base underscores their potential to revolutionize engineering processes. By automating these aspects of FMEA, companies can significantly speed up the analysis process and improve decision making. This automation reduces the burden of manual data analysis and allows engineers and decision makers to focus on more strategic and challenging tasks. As a result, overall productivity and innovation are increased as more resources are available for exploring new ideas and optimizing product design.

In summary, the use of LLMs in FMEA and product development offers significant benefits. They enable the rapid extraction of relevant information, support decision-making processes and facilitate the continuous improvement of products and systems. By using these tools, companies can make more informed decisions, implement effective risk mitigation strategies and ultimately drive innovation and quality in product development.

5. Discussion

The proposed framework effectively addresses several key challenges in the FMEA process, as shown in Table 1. Through the use of LLMs, the framework automates labor-intensive tasks such as data collection, pre-processing and failure mode identification, significantly reducing manual effort and minimizing human error. It also includes a KMS to improve the reuse of FMEA findings and ensure that past analyses are incorporated into future processes. In addition, the framework's ability to process unstructured data and scale to complex systems increases efficiency, while customized prompt engineering ensures that outputs are contextually relevant and aligned with FMEA-specific requirements. These solutions demonstrate that the framework is able to overcome traditional FMEA limitations and streamline the overall process.

Additionally, the proposed framework has demonstrated its potential to address several key challenges in the integration of LLMs into the FMEA process,

Table 3. Mapping LLM challenges to potential solutions.

LLM callenge	Potential solution
Security concerns	Use of locally deployed LLMs or encrypted data transfer protocols for sensitive information.
Lack of contextual understanding	Use of advanced prompt engineering techniques and validation to SMEs to improve the relevance of results.
Potential bias	Iterative review and validation of results by SMEs to mitigate bias and ensure objective results.
Maintenance and updating	Structured model maintenance with regular retraining using domain-specific data for relevance.
Over-reliance on automation	Clear division of tasks, with critical decision-making reserved for human experts to complement automation.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

particularly through the application of prompt engineering. As shown in Table 3, the study effectively tackled the challenge of contextual understanding by employing tailored prompt engineering techniques and validation by SMEs. This ensured that the outputs were both relevant and tailored to the specific requirements of the FMEA tasks. In addition, the framework addressed the risk of over-reliance on automation by maintaining a clear division of tasks, reserving critical decisions for human experts to complement the automated analysis.

Despite this progress, some challenges remain partially or fully unaddressed and require further investigation. For example, security concerns, such as protecting sensitive data and mitigating vulnerabilities in external LLM deployments, could be addressed by using locally deployed LLMs or encrypted data transfer protocols. Although contextual understanding was only partially addressed, further studies could improve the accuracy and relevance of the results through domain-specific training. Dealing with potential biases in LLM results requires iterative assessment and the use of different training datasets to ensure fairness and objectivity. Finally, the challenge of maintaining and updating LLMs requires the development of structured maintenance protocols, including regular retraining with relevant data.

Future research should focus on implementing these solutions in real industrial environments to evaluate their effectiveness. This would provide a more comprehensive understanding of the integration of LLMs into FMEA and strengthen the robustness of the proposed framework. Furthermore, investigating these challenges in diverse contexts and industries could uncover new insights and applications and further advance the practical utility of LLMs in structured engineering processes such as FMEA.

6. Conclusion and prospects for future research

This study addresses two major gaps in the literature: the lack of a comprehensive framework for the integration of LLMs into the FMEA process and the limited research on the practical challenges of such integration. The proposed framework combines a process model and an information system model to streamline FMEA activities and demonstrate the benefits of LLMs in improving efficiency, scalability and accuracy.

Key findings

- **Framework contributions:**
  - Development of a systematic framework that uses LLMs to automate data preprocessing, failure mode identification and risk analysis to reduce manual effort and minimize human error.
  - Use of a KMS to improve the reuse of knowledge to ensure that past analyses are incorporated into future FMEA processes.
  - Demonstrate the ability to process unstructured data, such as user reviews, and transform it into structured input for FMEA.
  - Improved scalability through the ability of LLMs to process large datasets quickly and efficiently.
- **Insights from the case study:**

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

  - Minimal programming and no fine-tuning were required, with effective results achieved through prompt engineering in just 2–3 iterations.
  - The generated FMEA tables highlighted the potential of LLMs for organizing, documenting and visualizing information to improve risk assessment and decision-making.
  - Working with company-specific FMEA reports could lead to even greater precision in the extraction of failure modes.
- **Limitations:**
  - **Generalization of outputs:** LLMs occasionally extrapolate failure modes beyond the input data, which raises questions about the desired level of detail for specific FMEA tasks.
  - **Reliance on SMEs:** Outputs had to be validated and refined by SMEs, limiting the scalability of fully automated processes.
  - **Domain-specific focus:** The study was focused on automotive parts, leveraging LLMs' pre-existing knowledge in this area. The extension to less common products or industries has yet to be tested.
  - **Confidentiality of data:** Security concerns, such as risks associated with external LLMs, were not fully addressed in this study.

Future research directions

- **Extension of the framework:** Inclusion of additional FMEA elements, such as the selection of corrective actions and the evaluation of risk mitigation strategies.
- **Improve generalization control:** Explore training and prompting strategies to minimize undesirable generalizations while retaining the ability to infer relevant context.
- **Dynamic data management:** Integrate real-time updates that link extracted information to source documents and dynamically refresh results as new data becomes available.
- **Testing in diverse domains:** Apply the framework to lesser-known products and industries to evaluate its robustness and adaptability.
- **Improve security and privacy:** Deploy locally hosted LLM infrastructures or explore open access models, such as OpenLLaMA (Geng & Liu 2023), to safeguard sensitive proprietary data.
- **Industrial validation:** Conduct large-scale evaluations in real industrial contexts to measure the impact of the framework on efficiency, time savings and accuracy. For example, Dell'Acqua *et al.* (2023) have shown that LLMs improve the quality of tasks by over 40% and reduce task duration by 25%, which is promising for FMEA applications.

This study illustrates the transformative potential of LLMs in automating labor-intensive and time-consuming FMEA processes. Although the framework offers significant benefits, such as faster pre-processing and better data organization, future efforts should focus on addressing the identified challenges to enable broader industrial application. By refining the framework and expanding its scope, LLMs can unlock their full potential in engineering processes and pave the way for more efficient, accurate and innovative solutions.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Acknowledgment

The authors would like to thank Damien Motte from Lund University for his suggestion to use LLM-based semantic analysis.

References

Abadi, M., **Barham, P.**, **Chen, J.**, **Chen, Z.**, **Davis, A.** *et al.* 2016 TensorFlow: a system for large-scale machine learning. *arXiv*, 1605.08695; doi:10.48550/arXiv.1605.08695.

AIAG and VDA. 2019 *Failure Mode and Effects Analysis—FMEA Handbook: Design FMEA, Process FMEA, Supplement FMEA for Monitoring and System Response*. Automotive Industry Action Group and Verband der Automobilindustrie, Southfield, MI.

AnkurJain. 2019 Edmunds-Consumer Car Ratings and Reviews (Version 3). Kaggle. Available at: https://www.kaggle.com/datasets/ankkur13/edmundsconsumer-car-ratings-and-reviews (accessed 4 February 2024).

Bommasani, R., **Hudson, D. A.**, **Adeli, E.**, **Altman, R.**, **Arora, S.** *et al.* 2021 On the opportunities and risks of foundation models. *arXiv*, 2108.07258; doi:10.48550/arXiv.2108.07258.

Brown, T., **Mann, B.**, **Ryder, N.**, **Subbiah, M.**, **Kaplan, J.**, **Dhariwal, P.**, … & **Amodei, D**. 2020 Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901.

Campbell, J. L., **Quincy, C.**, **Osserman, J. and Pedersen, O. K**. 2013 Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* **42**(3), 294–320; doi:10.1177/0049124113500475.

Cantamessa, M., **Montagna, F.**, **Altavilla, S.**, **Casagrande-Senetti, A**. 2020 Data-driven design: the new challenges of digitalization on product design and development. *Design Science* **6**, e27. doi:10.1017/dsj.2020.25.

Dell'Acqua, F., **McFowland, E. III**, **Mollick, E.**, **Lifshitz-Assaf, H.**, **Kellogg, K. C.** *et al.* 2023 Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper No. 24–013, Harvard Business School. Available at: https://www.hbs.edu/faculty/Pages/item.aspx?num=64700 (Accessed 19 December 2024).

Dai, J. X., **Boujut, J. F.**, **Pourroy, F.**, **Marin, P**. 2020 Issues and challenges of knowledge management in online open source hardware communities. *Design Science* **6**, e24. doi:10.1017/dsj.2020.18.

Department of Defense. 1980 *MIL-STD-1629A, Procedures for Performing a Failure Mode, Effects and Criticality Analysis*. Department of Defense, Washington, D.C.

Diemert, S. & **Weber, J. H**. 2023 Can large language models assist in hazard analysis? *arXiv*, 2303.15473; doi:10.48550/arXiv.2303.15473.

El Hassani, I., **Masrour, T.**, **Kourouma, N.**, **Motte, D.** & **Tavčar, J**. 2024 Integrating large language models for improved failure mode and effects analysis (FMEA): a framework and case study. *Proceedings of the Design Society* **4** 2019–2028. doi:10.1017/pds.2024.204.

Ege, D. N., **Øvrebø, H. H.**, **Stubberud, V.**, **Berg, M. F.**, **Steinert, M.**, **Vestad, H**. 2024 Benchmarking AI design skills: insights from ChatGPT's participation in a prototyping hackathon. *Proceedings of the Design Society* **4**, 1999–2008. doi:10.1017/pds.2024.202.

Ehring, D., **Menekse, I.**, **Luttmer, J.** & **Nagarajah, A**. 2024 Automatic identification of role-specific information in product development: a critical review on large language models. *Proceedings of the Design Society* **4**, 2009–2018. doi:10.1017/pds.2024.203.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Filz, M. A., **Langner, J. E. B.**, **Herrmann, C.** & **Thiede, S**. 2021 Data-driven failure mode and effect analysis (FMEA) to enhance maintenance planning. *Computers in Industry* **129**, 103451; doi:10.1016/j.compind.2021.103451.

Giordano V, **Consoloni M**, **Chiarello F**, **Fantoni G**. 2024 Towards the extraction of semantic relations in design with natural language processing. *Proceedings of the Design Society*. **4**, 2059–2068. doi:10.1017/pds.2024.208.

Geng, X. & **Liu, H**. 2023 *OpenLLaMA: An open Reproduction of LlaMA*. Available at: https://github.com/openlm-research/open_llama (accessed 11 November 2023).

Gericke, K., **Eckert, C.**, **Campean, F.**, *et al.* 2020 Supporting designers: moving from method menagerie to method ecosystem. *Design Science* Placeholder Text**6**, e21. doi: 10.1017/dsj.2020.21.

Girotra, K., **Meincke, L.**, **Terwiesch, C.** & **Ulrich, K. T**. 2023 Ideas are dimes a dozen: Large language models for idea generation in innovation; doi:10.2139/ssrn.4526071.

Gomez, A. P., **Krus, P.**, **Panarotto, M.** & **Isaksson, O**. 2024 Large language models in complex system design. *Proceedings of the Design Society* **4**, 2197–2206. doi:10.1017/pds.2024.222.

Goodfellow, I., **Pouget-Abadie, J.**, **Mirza, M.**, **Xu, B.**, **Warde-Farley, D.**, **Ozair, S.** … & **Bengio, Y**. 2014 Generative adversarial nets. *Advances in Neural Information Processing Systems* **27**, 2672–2680.

Hassan, F., **Nguyen, T.**, **Le, T.** & **Le, C**. 2023 Automated prioritization of construction project requirements using machine learning and fuzzy failure mode and effects analysis (FMEA). *Automation in Construction* **154**, 105013; doi:10.1016/j.autcon.2023.105013 .

Hodkiewicz, M., **Klüwer, J. W.**, **Woods, C.**, **Smoker, T.** & **Low E**. 2021 An ontology for reasoning over engineering textual data stored in FMEA spreadsheet tables. *Computers in Industry* **131**, 103496; doi:10.1016/j.compind.2021.103496.

Howard, J. & **Ruder, S**. 2018 Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia: Association for Computational Linguistics, pp. 328–339. https://aclanthology.org/P18-1031/

Hu, X., **Tian, Y.**, **Nagato, K.**, **Nakao, M.** & **Liu, A**. 2023 Opportunities and challenges of ChatGPT for design knowledge management. *Procedia CIRP* **119**, 21–28; doi:10.1016/j.procir.2023.05.001

Huang, J., **Xu, D.-H.**, **Liu, H.-C.** & **Song, M.-S**. 2019 A new model for failure mode and effect analysis integrating linguistic Z-numbers and projection method. *IEEE Transactions on Fuzzy Systems* **29**(3), 530–538; doi:10.1109/TFUZZ.2019.2955916 .

IATF. 2016 *IATF 16949:2016, Automotive Quality Management system standard, International Automotive Task Force.* IATF

IEC. 1985 *IEC 60812, Analysis Techniques for System Reliability – Procedure for Failure Mode and Effects Analysis (FMEA).* International Electrotechnical Commission, Geneva.

Kan, D. 2022, November 7 Car parts that start with A to Z – A thorough exploration. *List Explained.* Available at: https://listexplained.com/car-parts-that-start-with-a-to-z/ (accessed 3 June 2023).

List of Auto Parts. 2023, March 3 *Wikipedia.* Available at: https://en.wikipedia.org/wiki/List_of_auto_parts (accessed March 2023).

Kingma, D. P. & **Welling, M**. 2013 Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Landis, J. R., & **Koch, G. G.** (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**(1), 159–174. https://doi.org/10.2307/2529310.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Lewis, P., **Perez, E.**, **Piktus, A.**, **Petroni, F.**, **Karpukhin, V.**, **Goyal, N.** … & **Riedel, S**. 2020 Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474.

Liu, H.-C., **Chen X.-Q.**, **Duan, C.-H. and Wang, Y.-M**. 2019 Failure mode and effect analysis using multi-criteria decision-making methods: A systematic literature review, *Computers & Industrial Engineering* **135**, 881–897; doi:10.1016/j.cie.2019.06.055.

Liu, P., **Yuan, W.**, **Fu, J.**, **Jiang, Z.**, **Hayashi, H.** & **Neubig, G**. 2021 Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Maas, A. L., **Daly, R. E.**, **Pham, P. T.**, **Huang, D.**, **Ng, A. Y. and Potts, C.** 2011 Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y. & Mihalcea, R. (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 142–150. Stroudsburg, PA. Placeholder TextPlaceholder TextAvailable at: https://aclanthology.org/P11-1015. Dataset derived from original available at: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews (accessed 4 February 2024).

Mas'udah, M. & **Livotov, P**. 2024 Nature's lessons, AI's power: sustainable process design with generative AI. *Proceedings of the Design Society* **4**, 2129–2138. doi:10.1017/pds.2024.215.

Meron, Y. & **Tekmen, A. Y**. 2023 Artificial intelligence in design education: evaluating ChatGPT as a virtual colleague for post-graduate course development. *Design Science* **9**, e30. doi:10.1017/dsj.2023.28.

Na'amnh, S., **Salim, M. B.**, **Husti, I.** & **Daróczi, M**. 2021 Using artificial neural network and fuzzy inference system based prediction to improve failure mode and effects analysis: a case study of the busbars production. *Processes* **9**(8), 1444; doi:10.3390/pr9081444.

Peddi, S., **Lanka, K.** & **Gopal, P. R. C**. 2023 Modified FMEA using machine learning for food supply chain. *Materials Today: Proceedings*, in Press; doi:10.1016/j.matpr.2023.04.353.

Rae, J. W., **Borgeaud, S.**, **Cai, T.**, **Millican, K.**, **Hoffman, J.**, **Song, H. F.** & **Irving, G**. 2022 Scaling language models: Methods, analysis & insights from training Gopher. *arXiv preprint arXiv:2112.11446*.

Radford, A., **Wu, J.**, **Child, R.**, **Luan, D.**, **Amodei, D.** & **Sutskever, I**. 2019 Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 9.

SAE. 1994 *SAE J1739, Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual*. Society of Automotive Engineers, Warrendale, PA.

Sakwe, J. B., **Pereira Pessoa, M.** & **Hoekstra, S**. 2021 A FMEA based method for analyzing and prioritizing performance risk at the conceptual stage of performance PSS design. In *Proceedings of the International Conference on Engineering Design (ICED21)*, Gothenburg, Sweden, 16–20 August 2021. doi:10.1017/pds.2021.9.

Sarica, S. & **Luo, J**. 2024 The innovation paradox: concept space expansion with diminishing originality and the promise of creative artificial intelligence. *Design Science* **10**, e11. doi:10.1017/dsj.2024.10.

Siddharth, L., **Blessing, L.** & **Luo, J**. 2022 Natural language processing in-and-for design research. *Design Science* **8**, e21. doi:10.1017/dsj.2022.16.

Siddharth, L., **Blessing, L. T. M.**, **Wood, K. L.** & **Luo, J**. 2021 Engineering knowledge graph from patent database. *Journal of Computing and Information Science in Engineering* **22**, 021008; doi:10.1115/1.4052293.

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press

Soltanali, H. & **Ramezani, S.** 2023 Smart failure mode and effects analysis (FMEA) for safety–Critical systems in the context of Industry 4.0. In Garg, H. (Ed.) *Advances in Reliability, Failure and Risk Analysis*, pp. 151–176. Springer, SingaporePlaceholder TextPlaceholder Text; doi:10.1007/978-981-19-9909-3_7.

Spreafico, C. & **Sutrisno, A.** 2023 Artificial intelligence assisted social failure mode and effect analysis (FMEA) for sustainable product design. *Sustainability* **15**(11), 8678; doi: 10.3390/su15118678.

Stenholm, D., **Catic, A** & **Bergsjö, D**. 2019 Knowledge reuse in industrial practice: evaluation from implementing engineering checksheets in industry. *Design Science* **5**, e15. doi:10.1017/dsj.2019.10.

Tavčar, J., **Benedičič, J.** & **Žavbi, R**. 2019 Knowledge management support in the engineering change process in small and medium-sized companies. *International Journal of Agile Systems and Management* **12**(4), 354–381; doi:10.1504/IJASM.2019.104587 .

Tavčar, J. & **Duhovnik, J**. 2014 Tools and methods stimulate virtual team co-operation at concurrent engineering. In *Proceedings of the 21st ISPE Inc. International Conference on Concurrent Engineering*, Beijing, China, September 8–11, 2014, pp. 457–466. IOS Press, AmsterdamPlaceholder TextPlaceholder Text; doi:10.3233/978-1-61499-440-4-457.

Thirunavukarasu, A. J., **Ting, D. S. J.**, **Elangovan, K.**, **Gutierrez, L.**, **Tan, T. F. and Ting, D. S. W**. 2023 Large language models in medicine. *Nature Medicine* **29**(8), 1930–1940; doi:10.1038/s41591-023-02448-8.

Thomas, D. 2023 Revolutionizing failure modes and effects analysis with ChatGPT: unleashing the power of AI language models. *Journal of Failure Analysis and Prevention* **23**(3), 911–913; doi:10.1007/s11668-023-01659-y.

Wang, B., **Zuo, H.**, **Cai, Z.**, **Yin, Y.**, **Childs, P.**, **Sun, L.** & **Chen, L**. 2023 A task-decomposed AI-aided approach for generative conceptual design. In *ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, V006T06A009; doi:10.1115/detc2023-109087.

Wirth, R., **Berthold, B.**, **Krämer, A.** & **Peter, G**. 1996 Knowledge-based support of system analysis for the analysis of failure modes and effects. *Engineering Applications of Artificial Intelligence* **9**(3), 219–229; doi:10.1016/0952-1976(96)00014-0.

Wu, Z., **Liu, W.** & **Nie, W**. 2021 Literature review and prospect of the development and application of FMEA in manufacturing industry. *The International Journal of Advanced Manufacturing Technology* **112**(5–6), 1409–1436; doi:10.1007/s00170-020-06425-0

Xu, S., **Wu, Z.**, **Zhao, H.**, **Shu, P.**, **Liu, Z.**, **Liao, W.** … & **Li, X**. 2024 Reasoning before comparison: LLM-enhanced semantic similarity metrics for domain specialized text analysis. *arXiv preprint arXiv:2402.11398*.

Yucesan, M., **Gul, M.** & **Celik, E**. 2021 A holistic FMEA approach by fuzzy-based Bayesian network and best-worst method. *Complex & Intelligent Systems* **7**(3), 1547–1564; doi: 10.1007/s40747-021-00279-z

Zhao, W. X., **Zhou, K.**, **Li, J.**, **Tang, T.**, **Wang, X.** *et al.* 2023 A survey of large language models. *arXiv*, 2303.18223; doi:10.48550/arXiv.2303.18223

https://doi.org/10.1017/dsj.2025.7 Published online by Cambridge University Press