

Knowledge-augmented large language models to support automated HAZOP report generation

Ehab Elhosary^{a,c,*}, Osama Moselhi^{a,b}

^a Department of Building, Civil and Environmental Engineering, Concordia University, Canada

^b Director of Center for Innovation in Construction and Infrastructure Engineering and Management (CICIEM), Concordia University, Canada

^c Department of Structural Engineering, Faculty of Engineering, Tanta University, Egypt

ARTICLE INFO

Keywords:

RAG
HAZOP-adapted BERT
P&ID node
LLMs
Role-based CoT
K-Means

ABSTRACT

The Hazard and Operability (HAZOP) study is a widely used method for hazard identification and analysis in process industries. However, its preparation remains highly time-consuming and resource intensive. Recent advances in large language models (LLMs) offer new opportunities for automating HAZOP worksheet generation. Yet concerns persist regarding hallucinations associated with LLMs, low valid scenario ratios, and strong bias toward administrative safeguards. To address these challenges, this study presents a domain augmented LLM method that aims at enhancing the capability of LLMs in generating HAZOP worksheets. First, multimodal LLMs are employed to generate node descriptions from piping and instrumentation diagrams (P&ID) node images. Second, a domain-adaptive pretraining model is developed to enhance hazard-specific semantic representations. Third, a retrieval-augmented generation (RAG) model is conducted to ground LLMs (GPT-4o, GPT-4o-mini, and LLaMa 3.2) in external HAZOP-related knowledge. A total of 6120 past HAZOP records and 1140 incident reports were structured in JavaScript Object Notation (JSON) format, while embeddings of P&ID node descriptions and incident narratives were indexed using Facebook AI Similarity Search (FAISS) to enable similarity-based retrieval. Retrieved knowledge is combined with the P&ID node and its generated description into a role-based chain of thought (CoT) prompt to guide LLMs reasoning. The proposed method was evaluated on a reference P&ID node case and further validated on two additional nodes through an ablation study. The findings demonstrate that RAG-based models incorporating domain-adapted embeddings achieve higher semantic similarity (F1-scores > 92%) compared to 82–87% for non-grounded LLMs. The valid scenario ratio improves by approximately 10–18%, with RAG + GPT-4o + domain-adapted embedding + role-based CoT prompting achieving the highest ratio of 48.2%. In addition, deviation identification is enhanced, reaching a maximum valid ratio of 81.25%. Retrieval grounding further reduces the over-generation of procedural safeguards and increases the proportion of active and passive safeguards, producing distributions closer to the reference worksheet compared to non-grounded models. These findings confirm the effectiveness of domain augmented LLMs for supporting automated HAZOP worksheet generation.

1. Introduction

The hazard and operability (HAZOP) study is the most commonly used method for hazard identification and analysis in process industries (Wang and Gu, 2022). Its objective is to systematically investigate whether hazards are present, to understand how they lead to adverse consequences, and how these consequences could be prevented or mitigated through improved process design (IEC (2016)). Conventional HAZOP analysis is conducted by multidisciplinary teams through collaborative review sessions. The process begins by dividing the piping

and instrumentation diagrams (P&IDs) into process nodes, which are systematically analyzed to identify deviations by combining process parameters with guidewords. The team then investigates the causes and consequences of each deviation and determines appropriate safeguards and recommendations (F. Wang et al., 2023). This process is often time-consuming and resource-intensive, relies heavily on expert judgment, and is susceptible to inconsistency and human error (Joubert et al., 2021; Li and Zhao, 2025).

To address these limitations, intelligent and computer-aided HAZOP approaches have been developed since the 1980s (Single et al., 2019).

* Corresponding author at: Department of Building, Civil and Environmental Engineering, Concordia University, Canada.

E-mail addresses: ehab.elhosary@mail.concordia.ca (E. Elhosary), moselhi@encs.concordia.ca (O. Moselhi).

<https://doi.org/10.1016/j.psep.2026.108997>

Received 6 February 2026; Received in revised form 17 April 2026; Accepted 12 May 2026

Available online 14 May 2026

0957-5820/© 2026 The Author(s). Published by Elsevier Ltd on behalf of Institution of Chemical Engineers. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

These approaches generally fall into three categories: knowledge-based systems, model-based approaches, and data-driven models. Knowledge-based systems, including expert systems, ontologies, and Bayesian networks (BNs), aim to formalize process safety knowledge through structured representations and inference mechanisms. Expert systems emulate human reasoning using rule-based knowledge bases, while ontologies model domain concepts and relationships as directed acyclic graphs (DAGs), and BNs represent probabilistic dependencies among process variables to support causal inference. Model-based approaches integrate HAZOP logic with dynamic simulations and mathematical process models to capture system behavior under abnormal operating conditions. Although knowledge- and model-based methods have contributed to HAZOP support and semi-automation, they remain constrained by the difficulty of acquiring, validating, and continuously updating domain-specific knowledge, which is often subjective, labor-intensive, and time-consuming.

Recently, data-driven models, leveraging machine learning (ML), deep learning (DL), and natural language processing (NLP), have advanced HAZOP automation by learning patterns from historical HAZOP datasets to support safety decision-making and reduce manual effort (Elhosary and Moselhi, 2024). These models have been applied to automate specific HAZOP elements—such as predicting deviations, severity, likelihood, and risk levels, as well as clustering of causes and consequences—but they do not typically enable end-to-end generation of complete HAZOP worksheets. Large language models (LLMs) have significantly transformed natural language understanding and generation, prompting growing interest in their use for process safety applications (Lee et al., 2026).

LLMs have been explored for generating or supporting HAZOP worksheets in different areas. Lee et al. (2026) investigated whether LLMs (GPT-4o, GPT-4o-mini, LLaMa, and Gemini) can generate complete HAZOP worksheets directly from P&IDs, process description, and HAZOP guidelines without human intervention. While high similarity to a reference worksheet (above 86%) was reported, the proportion of semantically valid scenarios remained low (0.19–0.37), and the generated safeguards were heavily biased toward procedural actions over passive or active protections. Moreover, the study constrained the models to predefined deviations adopted from the reference worksheet, limiting their ability to independently identify or discover deviations and focusing primarily on elaborating causes, consequences, and safeguards for a fixed deviation set. Similarly, Kranz et al. (2024) proposed a GPT-4-based copilot that combines guideword reasoning with multi-step prompting to assist hazard identification in human–robot collaboration. However, hazards arising from human-robot interactions were barely recognized, primarily due to the lack of domain-specific training data. In addition, Mukharroo et al. (2025) employed LLMs to evaluate the completeness of safeguard identification in HAZOP studies. Their results show that performance improves with iterative prompting but remains highly sensitive to prompt design and user expertise, while other worksheet elements (deviations, causes, consequences, and recommendations) and other performance attributes (scenario completeness, reliability, and validity) were not assessed.

These studies collectively confirm that LLMs can produce HAZOP-like structured outputs and assist safety experts. However, they also reveal critical limitations, including limited similarity between generated and actual HAZOP worksheets, low valid scenarios ratio, hallucinations, bias toward procedural safeguard, reliance on predefined deviations, strong dependency on prompt design and human oversight, and limited access to domain-specific data. Without explicit grounding in validated process safety knowledge, LLM outputs may include plausible but technically incorrect statements, restricting their applicability in safety-critical decision-making (H. Liu et al., 2025).

These limitations strongly motivate the adoption of retrieval-augmented generation (RAG), in which LLM outputs are grounded in verified external knowledge, such as historical HAZOP records and incident investigation reports. In a RAG pipeline, retrieval plays a

critical role by identifying relevant context that constrains and improves the reliability of generation. However, retrieval effectiveness depends strongly on the embedding model used to represent queries and documents in a shared vector space. Most widely used embedding models, including BERT-based encoders, are pretrained primarily on large-scale general-domain corpora, enabling strong performance on generic NLP tasks but limiting their ability to represent domain-specific process safety semantics (Danish and Chatterjee, 2026). Because HAZOP texts contain specialized terminology and risk-related expressions that may be underrepresented in general corpora, generic embeddings may fail to capture critical safety meaning, thereby reducing retrieval quality and weakening the accuracy of downstream HAZOP worksheet generation. Therefore, domain-adapted embedding models tailored to HAZOP and process safety texts are essential for enhancing retrieval relevance and improving grounded generation. For instance, Feng et al. (2021) demonstrated that further pretraining Chinese wwm-BERT on 290 chemical accident reports and 10 HAZOP reports improved severity classification based on consequence descriptions, highlighting the value of domain-adaptive pretraining for process safety NLP tasks.

In parallel, current AI-assisted HAZOP software tools (e.g., SALUS HAZOP AI) often depend on manual input of process and equipment descriptions, which users have reported to be cumbersome and time-consuming (Elhosary et al., 2024). To reduce this burden, multimodal LLMs, which integrate vision and language capabilities, provide a practical mechanism for generating structured node-level descriptions from P&ID node images. Such descriptions can then be used as queries to retrieve relevant past HAZOP knowledge, enabling RAG-based generation of HAZOP worksheets grounded in domain-specific evidence.

In this context, this study proposes a RAG-based method that establishes a foundation for developing HAZOP-specific generative systems by integrating (i) multimodal LLM-generated P&ID node descriptions, (ii) a domain-adapted embedding model (HAZOP-Adapted-BERT), (iii) role-based chain-of-thought (CoT) prompting for step-wise reasoning, and (iv) retrieval from a HAZOP-specific knowledge base. First, GPT-4o and GPT-4o-mini are used to generate process and equipment descriptions for P&ID nodes under multiple temperature settings and repeated runs to assess consistency and variability. Second, a BERT-based embedding model is further pretrained on HAZOP-related corpora using masked language modeling (MLM) to improve domain representation. Third, the RAG pipeline employs the domain-adapted embeddings and a Facebook AI Similarity Search (FAISS) index to retrieve relevant node descriptions, incident narratives, and their associated HAZOP records. The retrieved content is then used to guide LLM-based worksheet generation using GPT-4o, GPT-4o-mini, and LLaMA3.2 VL. Fourth, role-based CoT prompting is applied to encourage structured reasoning through key HAZOP stages. Finally, the proposed method is evaluated on a reference P&ID node case by comparing grounded and non-grounded generation using semantic similarity metrics, scenario validity assessment, deviation analysis, and safeguard and recommendation distribution analysis. An ablation study is further conducted on two additional P&ID nodes to validate the method and assess the contributions of HAZOP-Adapted BERT and role-based CoT prompting to model performance.

The primary contribution of this study is to support HAZOP teams in generating HAZOP worksheets by systematically leveraging knowledge from previously analyzed projects, including P&ID nodes, associated HAZOP records, and incident investigation reports. Unlike prior studies that rely solely on standalone LLMs—often resulting in nonfactual outputs and hallucinations—the proposed method integrates multimodal LLMs, a domain-adapted embedding model, RAG, and role-based CoT prompting to enable knowledge-grounded reasoning. This integration allows LLMs to access and utilize structured knowledge from node descriptions and incident narratives, along with their associated HAZOP records, thereby improving the accuracy and reliability of generated worksheets. The HAZOP-Adapted BERT model further enhances hazard-related semantic representations, enabling more effective retrieval of

relevant cases from the knowledge base. By embedding and retrieving knowledge from historical HAZOP studies, the proposed method can learn from previously identified hazards and mitigation measures, improving the quality of generated HAZOP worksheets.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method, including multimodal LLM-based node description generation, HAZOP-Adapted-BERT, CoT prompting, and the RAG pipeline. Section 4 describes the data collection and preparation procedures. Section 5 reports implementation details, results, and discussion. Finally, Section 6 concludes the paper, highlighting the contributions of this work and offering insights into future research directions.

2. Literature review

This section focuses on reviewing the evolution of HAZOP automation and the advancements relevant to this study, including, large language models (LLMs), retrieval-augmented generation (RAG), domain-adaptive embedding models, and prompt engineering. It also discusses the current limitations of LLMs and the potential of using RAG to improve their accuracy in HAZOP analysis.

2.1. Evolution of HAZOP automation

The automation of HAZOP studies has been explored for more than three decades through a wide range of intelligent systems. Early efforts in the 1990s and 2000s primarily relied on rule-based expert systems, which encoded domain expertise into structured if-then rules to identify deviations, causes, and consequences, and to trace fault propagation patterns. Qualitative model-based reasoning approaches were developed to infer abnormal system behavior using representations such as signed directed graphs, qualitative differential equations, Petri nets, and Multilevel Flow Models (MFM) (Rossing et al., 2010). Subsequently, quantitative model-based techniques integrated HAZOP logic with dynamic process simulations to provide higher fidelity and more reliable insights under failure conditions. However, these methods typically require detailed plant-specific models and extensive calibration, making them costly and time-intensive to build and maintain (Mokhtarname et al., 2024).

To reduce dependence on manually constructed rule sets and engineering models, knowledge reuse and semantic reasoning methods have also been adopted. Case-based reasoning (CBR) systems retrieve similar historical hazard cases, while ontology-driven approaches support consistent representation of equipment, parameters, deviations, and safety knowledge. Several studies combined ontologies with CBR and ML to enhance reasoning and risk propagation, such as PetroHAZOP (Zhao et al., 2009) and KROSA (Daramola et al., 2013). Recent methods have leveraged knowledge graphs built from heterogeneous process data (e.g., P&IDs, HAZOP tables) and applied graph learning techniques such as graph attention mechanisms to support structured hazard reasoning (Li and Zhao, 2025). In addition, BNs have been introduced to model uncertainty and quantify probabilistic dependencies among risk factors, enabling inference over possible event scenarios based on evidence and prior knowledge (Gao and Li, 2022).

More recently, data-driven models have advanced HAZOP automation by extracting patterns from historical HAZOP datasets to support decision-making and reduce manual effort (Elhosary and Moselhi, 2024). Classification models have been applied to automate specific HAZOP elements, including severity assessment (Feng et al., 2021), deviation prediction (Ekramipooya et al., 2023), consequence prediction (Elhosary and Moselhi, 2025a), and recommendation prediction (Ekramipooya et al., 2024). In parallel, topic modeling techniques have been employed to identify latent themes in HAZOP reports and to cluster causes, consequences, safeguards, and recommendations. They progressed from probabilistic methods, such as Latent Dirichlet Allocation (Wang and Gu, 2022) toward embedding-based models that better

capture semantic similarity (Elhosary and Moselhi, 2025b, 2025c). Although effective for extracting reusable safety knowledge, these data-driven approaches typically address individual HAZOP elements and do not support end-to-end generation of complete HAZOP worksheets.

Despite the contributions of rule-based, model-based, ontology-driven, and data-driven approaches, these systems remain limited by the need for extensive human effort to define rules, construct detailed process models, or manually curate process variables and their relationships. Consequently, they often lack scalability across different plant configurations and may require significant time and expert involvement to adapt to new systems. These limitations have motivated growing interest in large language models (LLMs), which offer flexible natural language reasoning and the ability to generate structured outputs with minimal task-specific engineering.

2.2. LLMs in HAZOP studies

LLMs represent a breakthrough in NLP and are built upon the transformer architecture—a neural network framework introduced by Vaswani et al. (2017)—that utilizes attention mechanisms to capture long-range dependencies and contextual relationship in text. These models often contain billions of parameters and are trained on vast corpora to capture intricate patterns in language (Kojima et al., 2022). Recently, there has been growing interest in applying LLMs to HAZOP studies for automated worksheet generation. Kranz et al. (2024) introduced HAZOP-GPT, a copilot designed to assist safety experts in identifying hazards, causes, and consequences, and proposing mitigation strategies for human-robot collaboration tasks. They applied HAZOP guidewords from the IEC 61882 application guide and adopted a prompt-chaining strategy, where four sequential prompts were used, with each step was building on the output of the previous one. To make it more deterministic, reduce randomness and improve reproducibility, the model was applied by setting the temperature parameter to zero and using consistent prompt phrasing. The results show that the model can identify hazards affecting either the robot or the human individually; however, it struggles to capture hazards emerging from human-robot interactions. This limitation arises because GPT-4 is primarily trained on traditional hazard analysis descriptions and lacks sufficient exposure to detailed reports on hazards caused by interaction between subsystems.

Mukharroo et al. (2025) used LLMs to assist in the automated evaluation of safeguard identification completeness within HAZOP studies. They conducted initial training using synthesized datasets composed of simplified P&IDs and corresponding HAZOP worksheets. Four respondents with varying levels of experience (from 0 to 25 years) interacted with the LLM using few-shot prompting, iteratively refining prompts over multiple “shots”. The trained prompting strategy was applied to a real P&ID case. The LLM was tasked with generating HAZOP worksheet elements, including deviation, cause, consequence, and safeguards. The generated safeguards were then benchmarked against a reference HAZOP conducted by five senior experts. The results showed a positive correlation ($R^2 = 0.8968$) between respondent experience and model accuracy, as well as clear improvement with increasing prompt iterations. Using iterative prompting up to ten shots, the LLM achieved a safeguard completeness accuracy of 86.67%, demonstrating that few-shot prompt refinement can progressively improve performance. However, this study evaluated only safeguard completeness and did not assess other key HAZOP quality attributes, such as scenario completeness, causal correctness, consistency, or the validity of generated deviations, and it did not specify the LLM used.

Lee et al. (2026) examined whether LLMs can autonomously generate complete HAZOP worksheets directly from P&IDs, process descriptions, and HAZOP guidelines without human intervention. Using a zero-shot prompting strategy, the study evaluated GPT-4o, GPT-4o-mini, LLaMa 3.2, and Gemini 2.0 across three temperature settings (0.1, 0.5, and 1.0), with ten repeated runs per model at each temperature. The

120 generated HAZOP worksheets were benchmarked against an expert reference using BERT-based semantic similarity metrics, valid scenario ratio (via principal component analysis (PCA) and K-means clustering), and safeguard-type distribution analysis. Although all models achieved high similarity to the reference worksheet (F1-scores of 86.7% for GPT-4o, 86.5% for GPT-4o-mini and Gemini, and 86.2% for LLaMa), the valid scenario ratio remained low (0.19–0.37). Furthermore, the generated safeguards exhibited a strong bias toward procedural controls over engineering (active/passive) protections, with procedural safeguards comprising 74–81% of the outputs, compared with 65% in the reference worksheet. This overrepresentation of procedural measures indicates a tendency of LLMs to favor text-based, easily generated recommendations, reflecting limited reasoning depth and weak diversity in mitigation strategies. Moreover, the study constrained LLMs outputs to the same predefined guidewords and deviations adopted in the expert reference application. As a result, the models are not evaluated on their ability to independently identify or discover deviations, but rather on their ability to elaborate causes, consequences, and safeguards for a fixed deviation set. Finally, it remains unclear whether the LLMs followed a reliable sequence of HAZOP steps when producing the worksheets, since their internal reasoning process is not observable. Therefore, the authors emphasize the need for future research to adopt a chain-of-thought (CoT) prompting approach, guiding LLMs step-by-step through the HAZOP workflow to better assess their performance at each stage. Table 1 summarizes the application of LLMs for HAZOP automation, highlighting their objectives, main findings, and limitations.

As summarized in Table 1, existing studies demonstrate the potential of LLMs to support automated HAZOP worksheet generation. However, the major gaps related to these developments are constrained deviation generation, hallucinations, weak safeguard diversity, and a lack of domain specific training data. These limitations motivate the development of domain-augmented and retrieval-grounded frameworks that ground outputs in verified domain knowledge and improve scenario validity and safety relevance.

2.3. Retrieval-augmented generation method

Despite their strong natural language generation capabilities, LLMs may produce factually incorrect or unsupported outputs when responding to queries beyond the scope of their training data (Bang et al., 2023). Retrieval-Augmented Generation (RAG) mitigates this by retrieving relevant external knowledge and incorporating it as contextual input for generation, thereby enhancing factual consistency (Y. Wang et al., 2025). Recently, RAG has been increasingly explored within

process safety management and related domains in process industries to enhance the reliability of LLM-based analyses. For example, Yang et al. (2026) applied a RAG-based framework to offshore oil and gas safety management by retrieving relevant context from 1182 Bureau of Safety and Environmental Enforcement (BSEE) incident investigation reports prior to generation. Their results showed performance improvements of approximately 17% in precision and 11% in F1-score compared to LLM-only baselines for root-cause analysis. Other studies have integrated RAG with knowledge graphs and accident databases to support safety reasoning and question answering in industrial safety contexts, demonstrating improved factual grounding and interpretability compared with LLM-only approaches (J. Zhang et al., 2026; Zheng et al., 2025). Across these studies, RAG consistently outperforms standalone LLMs by reducing hallucinated content and improving alignment with historical safety evidence. These findings indicate that RAG has strong potential to be extended to HAZOP studies, where retrieving validated HAZOP records and incident reports may substantially enhance the reliability, completeness, and technical correctness of automatically generated HAZOP worksheets.

RAG frameworks typically consist of two main components: a retriever, which fetches relevant documents from an external knowledge base, and a generator, which synthesizes responses conditioned on the retrieved content. Retrieval is implemented by embedding both the user query and the knowledge base documents into a shared vector space and identifying semantically similar content. In the context of HAZOP studies, the query is the P&ID node description, which encapsulates the node's design intent, operating conditions, process flow, input and output streams, and connectivity. However, in large-scale industrial facilities—where a single plant may contain 400 or more P&ID nodes—manually preparing detailed process and equipment descriptions for each node is highly time-consuming and resource-intensive. Current AI-based HAZOP tools, such as SALUS HAZOP AI, still rely predominantly on manual entry of such descriptions (Salus Technical. 2023). User feedback from more than 150 industry practitioners indicates that this step is time consuming, cumbersome, and susceptible to human error (Elhosary et al., 2024). To address this challenge, multimodal LLMs, which integrate vision and language capabilities, can be leveraged to automatically generate process and equipment descriptions directly from P&ID node images. These generated descriptions can then serve as structured queries for RAG, enabling consistent retrieval of relevant safety knowledge to support downstream HAZOP worksheet generation.

For knowledge embedding, BERT-based encoders, among the most advanced transformer-based architectures, are commonly used to

Table 1
Summary of related studies on LLM-based HAZOP automation.

Study	LLM / Method	Objective	Main Findings	Limitations
(Kranz et al., 2024)	GPT-4 with prompt chaining (4 sequential prompts)	Assist safety experts in identifying hazards in human-robot collaboration.	• Successfully identified robot-only and human-only hazards.	• Limited ability to capture interaction hazards. • lack of domain-specific training data.
(Mukharrom et al., 2025)	Few-shot prompting with iterative refinement	Evaluate safeguard identification completeness.	• Achieved 86.67% safeguard completeness after 10 shots. • Positive correlation ($R^2 = 0.8968$) between respondent experience and model accuracy.	• No quantitative evaluation • Focused only on safeguards. • Did not assess deviations, causes, consequences, or recommendations. • Did not consider scenario completeness, causal correctness, and validity of generated deviations.
(Lee et al., 2026)	GPT-4o, GPT-4o-mini, LLaMA 3.2, Gemini (zero-shot)	Fully automate HAZOP worksheet generation.	• High semantic similarity (BERT F1-scores of 86.7% for GPT-4o, 86.5% for GPT-4o-mini and Gemini, and 86.2% for LLaMa). • Low valid scenario ratio (0.19–0.37)	• LLMs not specified • Constrained LLMs outputs to the same predefined deviations adopted manually. • No deviation discovery. • Strong procedural safeguard bias. • Did not consider recommendations in safeguard analysis. • Used process description generated manually which is time consuming in large scale plants.

encode textual content into high-dimensional vector representations while preserving contextual semantics. In HAZOP studies, bidirectional encoder representation from transformer (BERT) has been widely adopted as a foundational knowledge representation model. For example, Feng et al. (2021) employed a BERT-BiLSTM-Attention architecture to predict severity levels from consequence descriptions, where BERT was used to encode consequence text into contextualized embeddings. Similarly, Ekramipooya et al. (2024) utilized BERT in combination with a multi-layer perceptron (MLP) to predict HAZOP recommendations based on causes and consequences, relying on BERT to transform these textual inputs into vector representations. While these approaches demonstrate the effectiveness of BERT-based embeddings for downstream HAZOP-related tasks, standard BERT models are pretrained primarily on large-scale general-domain corpora. As a result, although they perform well on generic NLP tasks, they often struggle to adequately capture the specialized terminology, causal expressions, and risk-related semantics inherent to process safety texts, limiting their effectiveness in safety-critical applications (Danish and Chatterjee, 2026).

Because HAZOP texts are characterized by highly specialized terminology, causal expressions, and risk-related semantics that are often underrepresented in general-domain corpora, generic embedding models may fail to capture critical safety meaning. This limitation can significantly degrade retrieval relevance in RAG frameworks and weaken the accuracy and reliability of HAZOP worksheet generation. Consequently, domain-adapted embedding models tailored to HAZOP and process safety texts are essential for improving semantic retrieval and enabling more effectively grounded generation. Feng et al. (2021) further pretrained Chinese WWM-BERT on a domain-specific corpus comprising 290 chemical accident investigation reports and 10 HAZOP reports. The domain-adaptive model achieved 88.6% precision, 88.2% recall, and 88.2% F1-score in predicting consequence severity from textual descriptions, outperforming baseline models trained on general corpora. These results underscore the effectiveness of domain-adaptive pretraining in enhancing semantic representation and improving performance in HAZOP studies.

For the generator component, retrieved knowledge is combined with the user query to construct a prompt that conditions the LLM on both the task query and relevant external evidence. The effectiveness of this generation step is strongly influenced by the prompt engineering strategy, which plays a critical role in guiding LLM behavior and output quality. Prompt engineering aims to align model responses with the user intent and task requirements, thereby maximizing the utility of generative LLMs across diverse applications (Liu et al., 2025). Several prompting strategies have been developed, including zero-shot, few-shot, role-based, and chain-of-thought (CoT) prompting. Zero-shot prompting requires the LLM to respond directly to a query without any demonstrations, while few-shot prompting provides example question-answer pairs to guide response structure and content. Role-based prompting explicitly instructs the model to assume the role of a domain expert, encouraging expert-level reasoning and terminology. CoT prompting further enhances performance on complex reasoning tasks by eliciting structured, step-by-step reasoning, enabling the model to decompose problems into intermediate logical steps (Wei et al., 2022). Prior studies have shown that role-based and CoT prompting are particularly effective for inference-driven tasks when input data is implicit, leading to improved reasoning consistency and output reliability (Chen et al., 2025).

As RAG has not yet been applied to HAZOP studies, this research investigates the performance of LLMs grounded in external HAZOP knowledge compared to non-grounded LLMs for HAZOP worksheet generation. The proposed method, detailed in the following section, aims to support HAZOP teams during study sessions by improving the reliability and completeness of automatically generated HAZOP outputs.

3. Proposed method

This section describes a novel method, developed for automated generation of HAZOP worksheets. It consists of three main stages, performed sequentially. In the first, multimodal LLMs are employed to generate structured node descriptions directly from P&ID node images. In the second, a domain-adapted language model, HAZOP-Adapted BERT, is used to derive effective representations of hazard-related semantics. In the third, retrieval-augmented generation (RAG) is incorporated to enhance the interpretability and accuracy of the LLM outputs by grounding them in external HAZOP knowledge and historical incident reports. Role-based CoT prompting is applied to LLMs—including GPT-4o, GPT-4o-mini, and LLaMA 3.2 VL—to enable step-by-step reasoning during HAZOP worksheet generation. Fig. 1 shows the essential components of the proposed method.

3.1. Multimodal LLMs for P&ID node description

The first stage of the proposed method employs multimodal LLMs to generate P&ID node descriptions to facilitate the retrieval of similar nodes from the knowledge base. Multimodal LLMs—such as GPT, DeepSeek, Grok, and Gemini—have demonstrated strong performance across various domains, including hydrology and water science (Hosseini and Pourzangbar, 2026), urban environments (Belaroussi, 2025), software engineering (Peng et al., 2023), and transportation systems. However, their application in process industries, particularly to highly symbolic engineering representations such as P&IDs, remains relatively unexplored. In this study, GPT-4o and GPT-4o-mini were implemented, representing recent advancements in language and vision capabilities, and were evaluated to assess their capability to generate accurate descriptions from P&ID node images.

To query these models in a consistent way, a role-based prompt was carefully designed, because it strongly influences LLM outputs in terms of relevance, depth, tone, and structure. A standardized three-part prompt (see Appendix A) was applied, consisting of: (i) a contextual view describing the P&ID and HAZOP setting, derived from actual HAZOP reports and literature; (ii) a role-based instruction directing the model to respond as a HAZOP expert and provide a structured, domain-specific output; and (iii) a task-specific query requesting the node's main function and operating conditions, inlet and outlet streams, and process flow and connectivity. The prompt was iteratively refined through several trials until it consistently produced clear, precise, and technically sound responses. The finalized prompt was then applied unchanged to all LLMs to ensure fairness and reproducibility in the comparative evaluation.

In addition to prompt design, the temperature parameter was used to control output variability and assess model consistency. Higher temperature values increase response diversity by sampling less probable tokens, whereas lower values promote more deterministic behavior (Peeperkorn et al., 2024). Each LLM was evaluated at three temperature settings (0.0, 0.2, and 0.4) to examine the impact of randomness on the generated descriptions. To standardize output length, the maximum response was capped at 512 tokens for all models. Furthermore, each model generated ten descriptions per temperature setting, and the standard deviation across runs was computed to assess consistency and uniformity of the generated content. The overall workflow for generating and evaluating P&ID node descriptions is illustrated in Fig. 2.

Each P&ID node image, paired with the engineered prompt, is input into each model. The computer vision component identifies symbolic and spatial features, while the LLM generates the textual description of the node's function, operating conditions, inlet/outlet streams, and process connectivity. To enable robust model evaluation, the descriptions generated by each model were systematically compared with 33 benchmark node descriptions extracted from the literature using both quantitative and qualitative evaluation methods. Quantitative evaluation included BERT-based similarity metrics and cross-model similarity

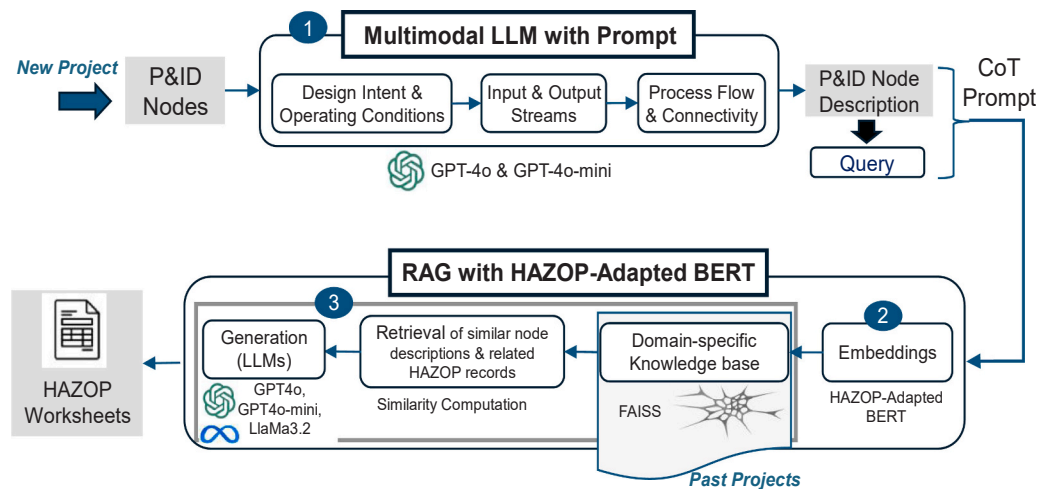


Fig. 1. Overview of multimodal LLM-RAG workflow for HAZOP worksheet generation.

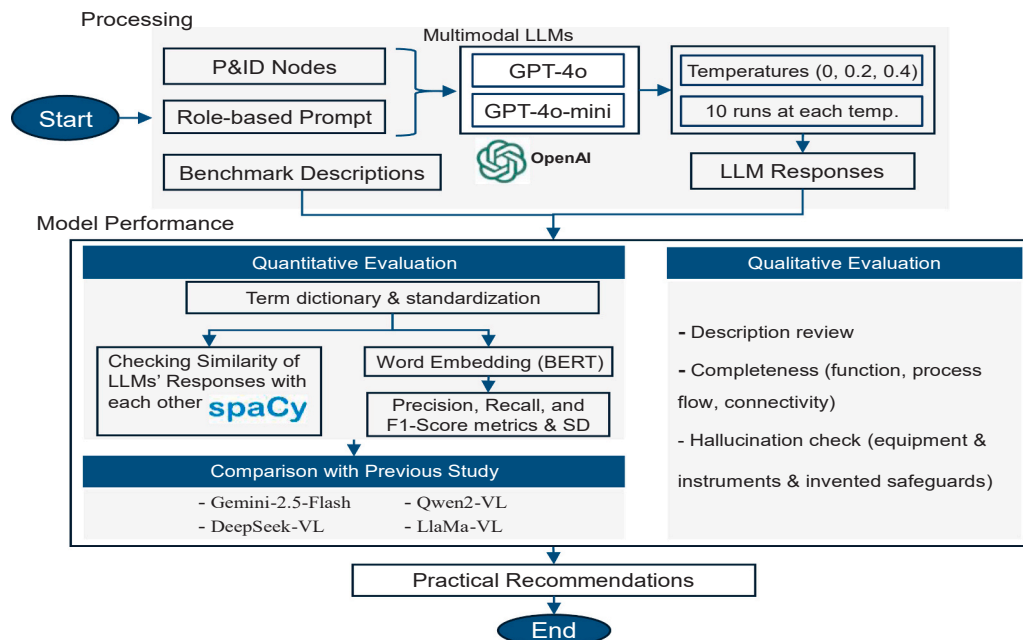


Fig. 2. Flowchart of the proposed method for LLM-based node description and evaluation.

analysis, while qualitative assessment involved manual review of the outputs. In addition, the results were compared with our previous study, which employed four multimodal models—Gemini-2.5-Flash, DeepSeek-VL, Qwen2-VL, and a LLaMa-based vision model—to generate P&ID node descriptions for computing functional similarity between current and past P&ID nodes.

Before computing quantitative evaluation metrics, terminology standardization was applied to reduce lexical variation across model outputs. This step was guided by the terminology used in the literature benchmarks, ensuring domain-consistent phrasing and minimizing inconsistencies during evaluation. Table 2 presents representative examples of this preprocessing step.

After standardization, both generated and benchmark descriptions were encoded into contextual token embeddings using a BERT-based model. Unlike traditional static word embeddings, BERT produces context-dependent representations for each token, allowing semantically equivalent expressions to be matched even when wording differs. BERT-based precision (P_{BERT}), recall (R_{BERT}), and F1-score ($F1_{BERT}$) were computed to provide a fine-grained assessment of semantic

Table 2

Examples of node description terminology standardization prior to Embedding.

Before Modifications	After Modifications	Before Modifications	After Modifications
- excessive pressure - high pressure - over-pressurization	Overpressure	- pressure relief valve. - relief valve - safety relief valve	Pressure relief valve.
- Overheating - temperature too high	High temperature	- check valve. - non-return valve	Check valve.
- Overcooling - temperature too low	Low temperature	- air cooler - air-cooled exchanger	Air cooler
- operator response manual intervention - field operator action	Operator intervention	- Compressor - gas compressor - centrifugal compressor	Centrifugal compressor

alignment between generated and benchmark descriptions by matching contextual token embeddings.

- R_{BERT} measures how well the content of the benchmark description is covered by the generated description.
- P_{BERT} measures how well the content of the generated description is aligned with and supported by the benchmark description.

$F1_{BERT}$ is the harmonic mean of precision and recall, balancing coverage and relevance. (T. Zhang et al., 2019). In this study, $F1_{BERT}$ is used as the primary metric because it simultaneously accounts for both precision and recall, thus provides a more robust measure of overall semantic similarity. Low $F1$ -score values indicate that the generated description diverges from the benchmark, suggesting potential hallucinations or missing functional details. These values are computed using the following equations (Zhang et al., 2019)

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \quad (1)$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \quad (2)$$

$$F1_{BERT} = \frac{2(P_{BERT} * R_{BERT})}{P_{BERT} + R_{BERT}} \quad (3)$$

Where x represents the set of tokens in the benchmark sentence, $\hat{x}$ represents the set of tokens in the generated sentence, and $x_i^T \hat{x}_j$ is the cosine similarity between individual tokens.

In addition, cross-model semantic similarity was computed to evaluate agreement among the LLM-generated descriptions themselves. This analysis quantifies whether different LLMs produce consistent functional narratives for the same node, offering an additional reliability signal beyond benchmark-based evaluation. The spaCy library was used to compute document-level cosine similarity between model outputs and to visualize pairwise agreement across models.

Finally, a qualitative manual review was conducted to identify errors, omissions, and inconsistencies that token-level metrics may not fully capture. The review assessed alignment with the key aspects defined in the prompt—main function, operating conditions, process flow and connectivity, and inlet/outlet streams—and helped distinguish acceptable elaboration from potential hallucination. Hallucinations were identified when the generated content introduced equipment, instruments, control actions, or safeguards that could not be reasonably inferred from the P&ID node image, or when functional relationships and process behaviors were inconsistent with the depicted symbols and connections.

3.2. HAZOP-adapted-BERT model

Although general BERT models have acquired extensive knowledge from broad domains and demonstrated exceptional performance on generic tasks, they face significant challenges in hazard perception capability when applied to highly specialized and safety-critical domains such as HAZOP studies. This limitation hinders effective representation of hazard-related semantics. Consequently, the objective of this stage is to develop adaptive BERT models specifically tailored for this domain to enhance hazard-related semantic representations. It is noteworthy that domain-adaptive pre-training (a process involving further unsupervised pre-training of a model on domain-specific corpora) has been proven as an effective method for infusing domain knowledge into pre-trained BERT models, thereby enhancing their performance on specialized tasks (Danish and Chatterjee, 2026). Developing a HAZOP-Adapted BERT model enables the model to learn HAZOP-related textual associations, thereby improving embeddings for downstream tasks such as content retrieval, similarity assessment, and

HAZOP worksheet generation. Data preprocessing is first performed to build high-quality HAZOP-related chunks, followed by further pre-training using a masked language modeling (MLM) model. The results are evaluated against the base BERT model using cross-entropy loss and the perplexity index.

3.2.1. Data preprocessing

Given the specialized terminology and complex structures of HAZOP studies, constructing a domain-specific preprocessing pipeline is essential to accurately capture deep hazard patterns and enhance downstream task reliability. Fig. 3 illustrates the proposed preprocessing pipeline for HAZOP-related text data. This pipeline ensures data quality and prepares the dataset for subsequent BERT pre-training. The compiled dataset, collected from multiple sources, was first extracted from PDF files and web-based resources (e.g., websites and HTML pages), in addition to structured files such as Excel sheets. The extracted text was then segmented into raw paragraphs or logical sections, and all content was merged into a unified corpus (e.g., in JSON format with associated metadata). Next, a cleaning and normalization stage was applied to lowercase and normalize the text, remove noise, and eliminate duplicate content to avoid bias in the training data. After that, the corpus was processed using BERT WordPiece tokenization, where the sequence length was defined, and long documents were split into multiple fixed-length chunks, while very short chunks were removed. Finally, the resulting preprocessed HAZOP text chunks were prepared for model training by splitting the dataset into 90% training and 10% testing sets.

3.2.2. MLM pretraining

HAZOP-adaptive pre-training was implemented using the Masked Language Modeling (MLM) task by masking portions of the input and training the model to recover the missing tokens from surrounding context. After preprocessing, the HAZOP corpus was split into 90% training data and 10% testing data to enable both learning and held-out evaluation. For MLM, 15% of tokens in each input sequence are randomly selected as prediction targets. Among these selected tokens, 80% are replaced with the special [MASK] token, 10% are substituted with random tokens, and the remaining 10% are kept unchanged to improve robustness. During training, the model outputs a probability distribution over the vocabulary for each masked position, and the learning objective is to maximize the likelihood of the original (ground-

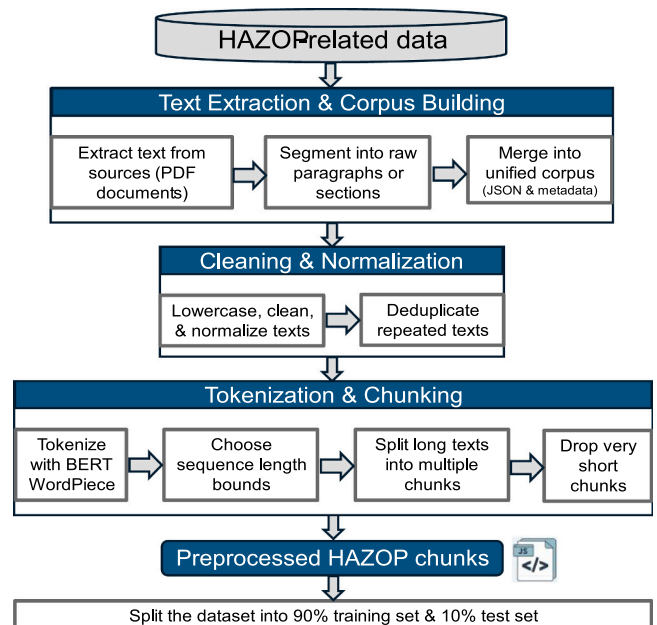


Fig. 3. HAZOP data preprocessing pipeline.

truth) tokens. To optimize performance, the cross-entropy loss function is used to measure the discrepancy between the model's predicted distribution and the ground truth. Performance is then assessed on the held-out test split using both the MLM loss and its exponential form, the perplexity index, where lower perplexity indicates better domain adaptation and stronger predictive ability on HAZOP-specific language patterns. The MLM loss and perplexity are formulated as follows (Lan et al., 2020):

$$l_{MLM} = -\frac{1}{N} \sum_{i=1}^N \log P(t_i | C_i) \quad (4)$$

$$Perplexity = \exp(l_{MLM}) \quad (5)$$

Where N represents the number of masked tokens; t_i denotes the i -th masked token; and C_i indicates the context of the i -th masked token. $P(t_i | C_i)$ represents the predicted probability distribution produced by the model. The adapted pre-training model was compared to the base BERT model to assess its performance in predicting masked tokens, as shown in Fig. 4.

By employing a MLM strategy for domain-adaptive pre-training of the BERT model, HAZOP-related knowledge can be effectively infused into the pre-trained model to enhance its hazard perception capabilities. This process leads to the development of HAZOP-Adapted BERT, a specialized model designed for enhancing the embedding representation of the hazard-related context. The HAZOP-Adapted BERT model is used to enhance the embedding representation of P&ID node descriptions, enabling the retrieval of the most similar nodes from the vector knowledge base to support LLMs in generating HAZOP worksheets.

3.3. HAZOP knowledge augmented LLM

Pretrained LLMs have demonstrated strong performance across a wide range of general-purpose applications. However, their effectiveness in domain-specific tasks is particularly critical for practitioners operating in specialized fields. Although fine-tuning is commonly used to adapt LLMs to specific domains, acquiring sufficiently large, diverse, and well-annotated datasets remains a significant challenge (H. Liu

et al., 2025). Consequently, rather than permanently modifying the model parameters, this study investigates the use of RAG combined with role-based chain-of-thought (CoT) prompt engineering to inject domain knowledge directly at inference time. This approach enhances the capability of LLMs to generate structured and technically grounded HAZOP worksheets.

RAG enables LLMs to access external and private corpora as a local knowledge base, thereby supporting more factually grounded and domain-specific responses (Lewis et al., 2020). Rather than relying solely on the LLM's internal parameters, the RAG method retrieves relevant documents at query time and dynamically augments the model input with contextual evidence. To enable efficient and semantically meaningful retrieval, both P&ID node descriptions and incident narratives are encoded into a shared vector space using the HAZOP-Adapted BERT model, and similarity search is performed over a FAISS index. In this study, the knowledge base comprises embedded node descriptions and incident narratives, along with their associated HAZOP elements stored as metadata. Specifically, node description entries are linked to deviations, causes, consequences, and countermeasures, while incident narrative entries are associated with causes, consequences, and recommendations.

As illustrated in Fig. 5, the proposed RAG method comprises a multi-stage workflow designed to support evidence-grounded HAZOP worksheet generation. First, historical HAZOP worksheets and Bureau of Safety and Environmental Enforcement (BSEE) district investigation reports are collected and preprocessed to extract P&ID node descriptions, incident narratives, and their associated HAZOP elements. Second, the extracted text is cleaned and segmented into smaller, contextually meaningful chunks to facilitate efficient indexing and retrieval. Third, all processed text chunks are encoded into high-dimensional vector representations using the proposed HAZOP-Adapted BERT model. Fourth, the resulting embeddings are indexed using a dense vector database (FAISS), enabling fast similarity search, while the corresponding HAZOP elements are stored as structured metadata linked to each embedding. Fifth, for a new P&ID node, its generated description is embedded using the same HAZOP-Adapted BERT model and used as a query to retrieve the top-k most relevant node descriptions and incident narratives, along with their associated elements from the knowledge

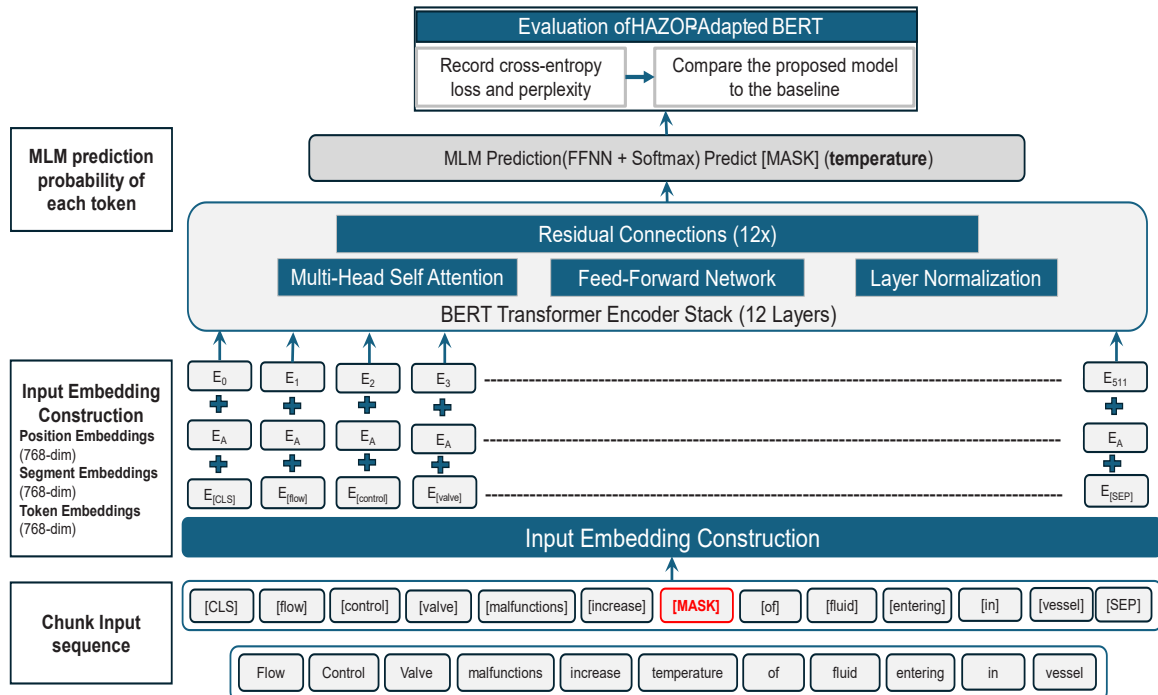


Fig. 4. HAZOP-adaptive pretraining using MLM.

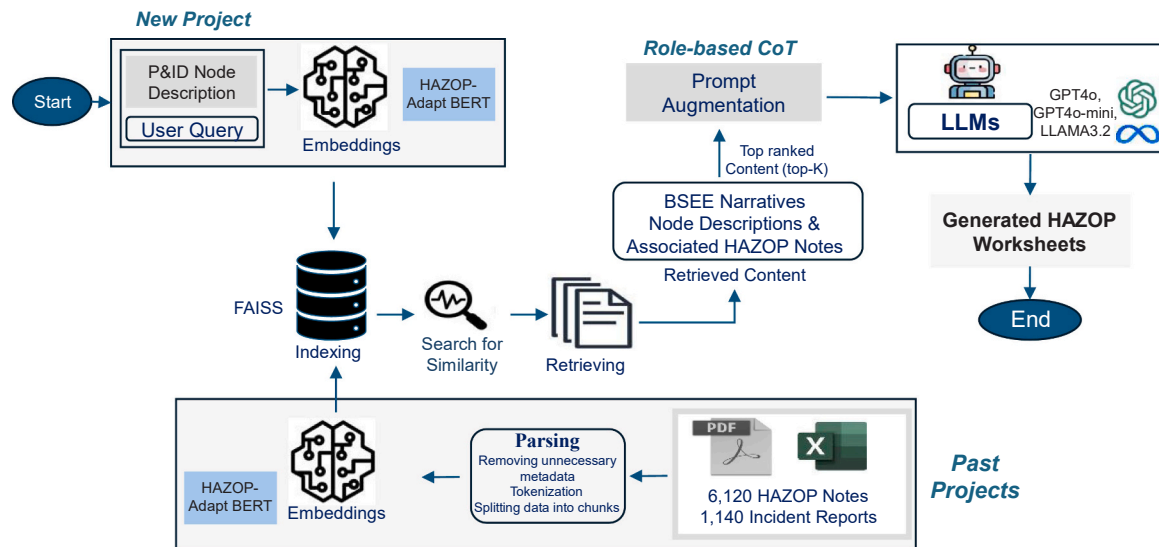


Fig. 5. Workflow of the RAG method to generate HAZOP worksheets.

base.

Finally, the retrieved contextual evidence is combined with a structured role-based chain-of-thought (CoT) prompt to guide the LLM during generation. This prompt encourages the model to (i) leverage past node descriptions, incident knowledge, and their associated elements to improve reasoning accuracy and domain relevance; and (ii) systematically identify deviations from design intent, along with their potential causes, consequences, existing safeguards, and recommended risk-mitigation measures. In general, prompts play a critical role in guiding LLMs toward the task being investigated. The structure, tone, and level of detail of a prompt directly influence the formality, depth, and organization of the generated outputs. In safety-critical applications such as HAZOP studies, carefully designed prompts are therefore essential to ensure systematic reasoning, technical relevance, and consistency across the generated results. In this study, a standardized prompt structure was developed to support automated HAZOP worksheet generation and was applied uniformly across all models and runs. The prompt combines: (i) a role-based instruction, (ii) a controlled chain-of-thought (CoT) reasoning directive, and (iii) retrieval-augmented contextual grounding. Specifically, the role-based instruction directs the LLM to act as a senior HAZOP facilitator, thereby encouraging the use of domain-appropriate terminology and structured analytical reasoning. The CoT component explicitly instructs the model to reason internally in a step-by-step manner by analyzing the P&ID node image, identifying the design intent, systematically applying standard HAZOP guidewords, and evaluating each resulting deviation in terms of causes, consequences, safeguards, and recommendations, while suppressing the disclosure of intermediate reasoning steps in the final output.

To reduce hallucination and improve traceability, the prompt further constrains the model to rely only on information inferred from the P&ID node image, the generated node description, and the retrieved evidence pack, while prohibiting the invention of unsupported equipment, instruments, or safeguards. This uniform prompt structure was iteratively refined through preliminary trials and then fixed for all experiments to minimize prompt-induced bias and ensure a controlled evaluation of model performance. An example of this prompt is provided in Appendix B.1. To assess consistency and robustness, each LLM generates the HAZOP worksheet across 10 runs at three temperature settings (0.1, 0.2, and 0.4), and the generated results are subsequently evaluated against the corresponding expert-validated HAZOP worksheet. Unlike node description generation, which includes a temperature setting of 0.0, a value of 0.1 was used for HAZOP worksheet generation to allow a small degree of variability in reasoning, rather than enforcing fully

deterministic outputs. This choice provides the models with limited flexibility to better analyze the node image, description, and retrieved content, thereby enabling the identification of potential deviations, causes, consequences, as well as appropriate safeguards and recommendations for mitigation. This workflow demonstrates how dense retrieval, domain-adapted embeddings, role-based CoT prompting, and generative inference are integrated into a unified method for automated HAZOP worksheet generation.

In this study, three state-of-the-art LLMs—GPT-4o, GPT-4o-mini, and LLaMa 3.2—are augmented using the proposed RAG method with past HAZOP worksheets and incident investigation reports to enhance their capabilities in generating HAZOP worksheets for new projects. These models, which represent recent advances in language and multimodal reasoning capabilities, are evaluated to investigate the feasibility of automating the HAZOP process.

3.4. HAZOP-augmented LLM evaluation

The proposed HAZOP-knowledge-augmented LLMs are evaluated against baseline LLMs operating without external knowledge in generating HAZOP worksheets. The evaluation comprises two complementary components: model-based similarity assessment and HAZOP-based assessment. Following the evaluation measures used by Lee et al. (2026), textual similarity, the number of generated scenarios, the valid scenario ratio, and safeguard analysis are used to evaluate model performance. The same evaluation measures are applied to both the augmented and baseline LLMs in this study. In addition, the generated deviations are explicitly compared against the reference deviations identified in the real HAZOP worksheet, and the generated recommendations are further analyzed alongside the generated safeguards to assess their relevance and consistency.

The 180 generated HAZOP worksheets, comprising 90 generated by RAG-based models and 90 by baseline LLMs, follow a consistent tabular structure consisting of guideword, parameter, deviation, possible causes, possible consequence, existing safeguards, and recommendations. However, as the worksheets are expressed in natural language, they cannot be directly analyzed using numerical methods. Therefore, the text is first preprocessed and then transformed into numerical representations using word embedding techniques. Given the diversity and ambiguity in HAZOP documentation, terminology standardization is performed prior to embedding. Specifically, synonyms and variant expressions across worksheets are replaced with standardized terms based on the reference worksheet, thereby ensuring uniformity and

minimizing inconsistencies during evaluation. Table 3 lists representative examples of this preprocessing procedure for the selected case study.

3.4.1. Model-based similarity assessment

Model performance is evaluated based on the semantic similarity between the generated and reference HAZOP worksheets, providing an overall indication of each LLM's effectiveness. Following terminology standardization, a three-step process is adopted: (1) row-level embedding using HAZOP-Adapted BERT; (2) contextual row alignment; and (3) similarity measurement between aligned contents. These steps are illustrated in Fig. 6.

Row alignment is performed to match semantically equivalent rows between the reference and generated worksheets. This step is necessary because generated outputs may present the same content in a different row order. Therefore, semantic similarity is used to identify the most appropriate correspondence between rows. First, each row in both the reference and generated worksheets is embedded using the HAZOP-Adapted BERT model to obtain high-dimensional semantic representations (768 dimensions). The semantic similarity between rows is then computed using cosine similarity (Lahitani et al., 2016), defined as:

$$\text{Cosine similarity} = \frac{1 + \cos(r_i, g_j)}{2} \quad (6)$$

$$\cos(r_i, g_j) = \frac{\sum_{k=1}^n r_{i,k} g_{j,k}}{\sqrt{\sum_{k=1}^n r_{i,k}^2} \sqrt{\sum_{k=1}^n g_{j,k}^2}} \quad (7)$$

Where r_i denotes the embedding of the i -th reference row, and g_j denote the embedding of the j -th generated row. The variable n represents the embedding dimension, while $r_{i,k}$ and $g_{j,k}$ represent the k -th components of the corresponding embedding vectors. The normalization term ensures that the similarity measure is independent of vector magnitude. The cosine similarity is rescaled to the range [0,1] for interpretability. For each reference row r_i , similarity scores are computed against all generated rows. The generated row exhibiting the highest similarity score is selected as the best semantic match:

$$g^* = \text{argmax}_{Sim} (r_i, g_j) \quad (8)$$

Where g^* is the generated row with the highest similarity to the reference row r_i . This alignment rule ensures that each reference row is paired with the most semantically equivalent generated row, regardless of differences in row ordering or phrasing.

Following the row alignment step, BERT-based precision, recall, and F1 metrics are computed as defined in Eqs. (1), (2), and (3). These metrics have been widely validated against human references across

various domains (Bui et al., 2025; Tikayat Ray et al., 2023). They provide a fine-grained evaluation by comparing individual tokens between the reference and generated worksheets. The BERT F1 Score is adopted as the primary evaluation metric, as it integrates both precision and recall and provides a robust measure of overall semantic similarity.

In addition, a paired t -test is conducted to compare the F1-scores on a run-by-run basis between the RAG-augmented models and their corresponding baseline models, to verify that the observed performance improvements are attributable to the incorporation of external knowledge rather than random variation. The test produces a p -value, which represents the probability that the observed difference between the two models could occur by chance under the null hypothesis of no true performance difference. A lower p -value indicates stronger evidence against the null hypothesis, suggesting that the observed improvement is unlikely to be due to randomness.

3.4.2. HAZOP-based assessment

HAZOP-based performance is evaluated based on three core objectives: (1) scenario identification, (2) deviation analysis, and (3) assessment of generated safeguards and recommendations.

3.4.2.1. Scenario analysis. This metric evaluates the ability of both augmented and baseline LLMs to generate potential HAZOP scenarios. Consistent with the evaluation framework proposed by Lee et al. (2026), scenario analysis is measured using two indicators: the total number of generated scenarios and the valid scenario ratio. In addition, a complementary manual evaluation is conducted to further validate the scenario ratio.

The total number of scenarios represents the volume of HAZOP scenarios generated across 30 repeated runs at three temperature settings for each model using a consistent role-based CoT prompt. This indicator captures the model's capacity to generate diverse scenarios, including deviations, causes, and consequences, thereby reflecting its generative comprehensiveness.

The valid scenario ratio evaluates the semantic alignment between generated and reference HAZOP scenarios. Following the methodology adopted in prior work, scenario embeddings generated using the HAZOP-Adapted BERT model are first reduced in dimensionality using Principal Component Analysis (PCA) and subsequently clustered using the K-means algorithm. The cluster containing the majority of reference scenarios is identified, and the proportion of generated scenarios within this cluster relative to the total number of generated scenarios is computed as the valid scenario ratio. This metric quantifies the extent to which model outputs are semantically consistent with expert HAZOP analyses.

K-means clustering is applied to the PCA-reduced embeddings to uncover latent semantic groupings. As an unsupervised machine learning algorithm, K-means partitions unlabeled data into K clusters by minimizing intra-cluster variance using Euclidean distance (Wagstaff et al., 2001). The clustering objective is defined by the within-cluster sum of squares (WCSS) (G. Liu et al., 2021). Although increasing k generally reduces WCSS, large values may lead to overfitting and fragmented clusters. Therefore, the elbow method is employed to identify the optimal number of clusters for each model. This method locates the point at which additional clusters yield diminishing returns in reducing WCSS (Bholowalia and Kumar, 2014).

To further validate the reliability of the clustering-based valid scenario ratio, a complementary manual evaluation is conducted by randomly selecting a set of generated worksheets for each model and comparing the generated scenarios against the reference scenarios.

3.4.2.2. Deviation analysis. Deviation analysis focuses on evaluating the accuracy and completeness of model-generated deviations, as models were allowed to freely explore potential hazards associated with the investigated P&ID node. It examines the ability of both augmented and

Table 3
Examples of HAZOP terminology standardization prior to Embedding.

Before Modifications	After Modifications	Before Modifications	After Modifications
More water level in V-106 High level	high level	Pipeline Blockage Blockage in the pipeline Blockage in the line to crude oil pipeline	blockage in pipeline.
Less Water Level in V-106 Low level	low level	Increased Crude oil pressure. Overpressure	Overpressure.
More Pressure in the discharge of P-102 High pressure:	high pressure	Pipeline failure due to corrosion. piping failure	piping failure
Less flow from LV-102 to V-106 Low Flow Less Flow	less flow	Overfilling Excessive water supply	overfilling

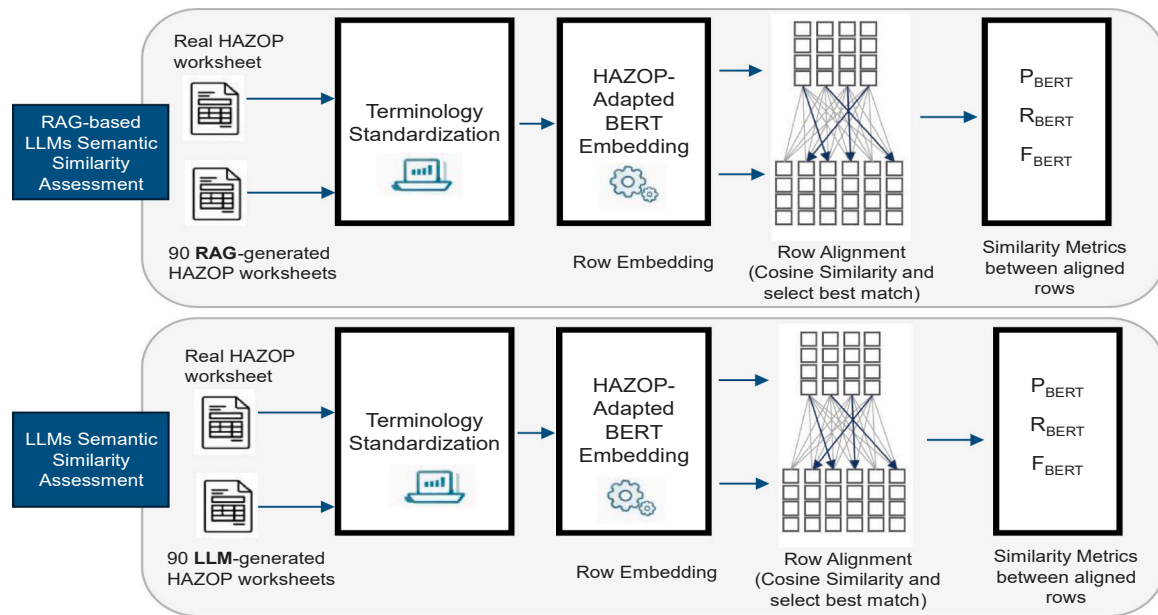


Fig. 6. Similarity evaluation procedures of generated worksheets by RAG and baseline LLMs.

baseline LLMs to accurately reproduce expert-validated deviations across 30 repeated runs and varying temperature settings for each model. Four indicators are employed: (1) the number of unique deviations, (2) the number of correct deviations, (3) the valid deviation ratio, and (4) model efficiency.

The number of unique deviations represents the distinct deviations generated after removing duplicate entries. The number of correct deviations corresponds to those that match the deviations identified in the reference HAZOP worksheet after terminology normalization. The valid deviation ratio is defined as the proportion of correct deviations relative to the total number of deviations in the reference worksheet, providing a normalized measure of deviation accuracy. Model efficiency is defined as the proportion of correct deviations relative to the total number of unique generated deviations, reflecting how effectively the model converts generated deviation diversity into accurate, expert-aligned outputs.

This analysis is conducted using exact string matching. Following terminology standardization and normalization, deviations are represented as unique sets for each worksheet after removing duplications. Correct deviations are identified by computing the intersection between each generated deviation set and the reference deviation set using exact string matching. This procedure enables consistent and transparent identification of correctly generated deviations while minimizing ambiguity caused by linguistic variability.

3.4.2.3. Safeguards and Recommendations distribution analysis. Beyond scenarios and deviation analyses, a fundamental objective of HAZOP is to define appropriate safeguards and recommendations to effectively mitigate risks to personnel, equipment, the environment, and surrounding communities (Almoussa et al., 2025; Park and Pasman, 2018). Building on the classification framework proposed by Lee et al. (2026), which focused primarily on safeguards, this study extends the analysis to evaluate the quality, diversity, and practical relevance of both safeguards and recommendations generated by RAG-based and baseline models.

Safeguards are generally categorized into three primary groups: inherently safer design and passive safeguards, active safeguards, and procedural safeguards (CCPS, 1998; Lee et al., 2026). Inherently safer design and passive safeguards aim to eliminate or reduce hazards at the source through design-oriented measures, such as minimization, substitution, moderation, and simplification, as well as physical protection

elements that operate without human intervention (e.g., rupture disks and containment systems). Active safeguards involve automated detection and response mechanisms, including sensors, alarms, control valves, interlocks, and emergency shutdown systems, which rely on instrumentation and control logic. Procedural safeguards consist of administrative and organizational controls, such as standard operating procedures, operator training, inspection routines, and communication protocols, and are generally more vulnerable to human error.

Each generated safeguard and recommendation is assigned to its corresponding category to facilitate systematic comparison across models. Passive safeguards in this study include equipment installation and design modifications (e.g., non-return valves), while active safeguards comprise instrumentation upgrades, pressure relief devices, alarm and warning systems, and emergency interlocks. Procedural safeguards encompass monitoring practices, labeling, maintenance and inspection activities, operational verification, and personnel training. This classification enables both qualitative and quantitative analyses of safeguard and recommendation distributions, allowing assessment of the extent to which augmented LLMs prioritize inherently reliable protection measures over administratively dependent controls. The resulting distribution patterns provide insights into the feasibility, maturity, and current limitations of RAG-based HAZOP automation from a process safety engineering perspective.

4. Data collection and preparation

This study compiled a multi-source dataset to support domain-adaptive pre-training and RAG-based HAZOP worksheet generation. The data were organized according to their intended use, either for further pre-training or for constructing a knowledge base to support information retrieval.

4.1. Corpus development for HAZOP-adapted BERT model

The development of a domain-adaptive language model for HAZOP studies requires a comprehensive and diverse textual corpus that captures the linguistic patterns, technical terminology, and contextual variations encountered in process industries. A key challenge lies in collecting, standardizing, and integrating data from heterogeneous sources with varying formats, structures, and accessibility constraints. This study gathered HAZOP-related data from multiple academic and

industrial sources to construct a unified corpus that reflects both theoretical knowledge and practical experience in process safety and hazard analysis.

A total of 1513 investigation reports and 513 HAZOP-related publications were collected, covering oil and gas operations, refining, and natural gas pipeline systems. The reports include 1174 incident investigation reports from the U.S. Bureau of Safety and Environmental Enforcement (BSEE, 2025), 181 pipeline audit reports from the Canada Energy Regulator (CER, 2025), 117 investigation reports from the U.S. Chemical Safety Board (CSB, 2025), and 30 pipeline investigation reports from the Transportation Safety Board of Canada (TSB, 2025). These reports were obtained from their official websites using an automated web-scraping pipeline (implemented with the requests and BeautifulSoup libraries). In addition, 11 industrial HAZOP reports were obtained through contacting process safety engineers. The 513 HAZOP-related PDF documents, including literature, theses, standards, and technical reference books, were curated through consultations with domain experts and systematic searches of Scopus, Web of Science, and Google Scholar, primarily covering publications from the past two decades. A detailed summary of the compiled datasets is provided in Table 4.

Textual content was extracted from the PDF documents using an automated pipeline primarily based on the PyMuPDF library. For documents in which PyMuPDF failed to extract meaningful content—such as scanned, protected, or poorly formatted files—the pdfplumber library was employed as a secondary extraction method. This fallback library maximizes text recovery across heterogeneous document formats. For

each page, raw textual content was retrieved and subsequently normalized to remove line-break artifacts, hyphenation, and excessive whitespace. Paragraph-level segmentation was then applied by identifying logical breaks between blocks of text.

To improve data quality, short, low-information, and boilerplate segments (e.g., headers, footers, and legal disclaimers) were removed. Duplicate paragraphs were eliminated using hash-based matching, in which each normalized paragraph was converted into a unique digital fingerprint and repeated entries were discarded. All cleaned paragraphs were stored in a structured JSONL (JSON Lines) format and linked to detailed metadata, including document source, file path, and page number, to ensure traceability. The resulting corpus comprised 29,061 unique paragraphs, totaling 13,462,690 words, with an average length of 463.26 words per paragraph. Since the baseline embedding model (bert-base-uncased) supports a maximum input length of 512 tokens, paragraphs exceeding this limit required further segmentation. Accordingly, 9998 paragraphs longer than 512 tokens were identified and divided into smaller, non-overlapping chunks, each constrained to a maximum length of 512 tokens. Chunks containing fewer than 25 tokens were discarded to eliminate low-information segments. This chunking procedure produced a final training corpus of 40,212 text segments, with a mean length of 334 tokens per chunk. The dataset was subsequently partitioned into 90% training data (36,191 chunks) and 10% testing data (4021 chunks) to support domain-adaptive pretraining and model evaluation.

4.2. Corpus development for RAG-based HAZOP worksheet generation

For RAG-based HAZOP worksheet generation, a structured dataset consisting of P&ID node descriptions and their associated HAZOP elements, including deviations, causes, consequences, safeguards, and recommendations, is required. This structure enables the RAG pipeline to retrieve semantically similar node descriptions and their corresponding HAZOP elements to support accurate worksheet generation.

Due to security restrictions, access to complete industrial HAZOP reports is limited. Consequently, most HAZOP records were manually extracted from published literature, which typically contains partial case studies presented in heterogeneous formats and reporting styles. The extracted HAZOP records include certain deviations while omitting others. Manual extraction was therefore necessary to ensure consistency across the dataset. In total, 6120 structured HAZOP records linked to 278 P&ID nodes were extracted from the 513 HAZOP-related publications and 11 complete industrial HAZOP reports. It should be noted that the limited availability of actual complete HAZOP reports (only 11 reports collected from industry) restricts the diversity and comprehensiveness of the knowledge base and may consequently affect the performance of the generation process. This limitation may also impact the model's ability to identify rare or complex hazard scenarios that are underrepresented in the dataset. These records include deviations, causes, consequences, and countermeasures. Since node-level descriptions are rarely available in the literature, corresponding P&ID node descriptions were generated using the GPT-4o-mini multimodal model.

In addition, the 1140 BSEE district incident investigation reports follow a semi-structured template and include dedicated sections describing incident narratives, causal factors, consequences, and corrective actions. This structured information provides valuable domain knowledge that can enhance the learning and generation capabilities of LLMs. All PDF documents were initially processed using the Unstructured library to convert each report into an ordered sequence of text segments while preserving the original reading order. Incident narratives were then extracted using a rule-based procedure based on explicit section anchors and regular expressions. The beginning of each narrative was identified by searching for headings matching patterns such as "INVESTIGATION FINDINGS" or "DESCRIBE IN SEQUENCE HOW ACCIDENT HAPPENED". The end boundary of the narrative was

Table 4
Overview of the dataset used for HAZOP-Adapted BERT training.

Dataset	Data Source	Scope	Time Range	Data Volume
BSEE panel reports	U.S. Bureau of Safety and Environmental Enforcement (BSEE)	Major offshore incidents involving fatalities, serious injuries, or major pollution.	2008–2024	34 reports
BSEE district reports	U.S. Bureau of Safety and Environmental Enforcement (BSEE)	Less severe incidents that require fewer resources than panel investigations.	2003–2025	1140 reports
CER reports	Canada Energy Regulator (CER)	Review of pipeline incidents and audit findings under certain operating conditions.	2012–2023	181 reports
CSB reports	U.S. Chemical Safety Board (CSB)	Detailed reports on significant oil and gas and chemical accidents.	2000–2025	117 reports
TSB reports	Transportation Safety Board of Canada (TSB)	Natural gas Pipeline accident investigations to advance transportation safety.	1996–2013	30 reports
HAZOP actual reports	Collected through contacting Process safety engineers	Actual industrial HAZOP reports in oil and gas facilities.	2022–2024	11 reports
HAZOP literature, guides, and reference books	Consultations with domain experts and searches of Scopus, Web of Science, and Google Scholar	Case studies, technical guides, theses, reference books related to HAZOP application in the process industries	2000–2025	513 HAZOP related data

determined by subsequent section headers such as “LIST OF PROBABLE CAUSE(S)”. This strategy was adopted because the “INVESTIGATION FINDINGS” section often contains multiple subsections (e.g., Incident Summary, Sequence of Events, BSEE Investigation, and Conclusions). Therefore, defining explicit start and end anchors ensures reliable extraction of complete narratives. Causes were extracted using section headings such as “LIST THE PROBABLE CAUSE(S) OF ACCIDENT”, “LIST THE CONTRIBUTING CAUSE(S) OF ACCIDENT”, and “SPECIFY VIOLATIONS DIRECTLY OR INDIRECTLY CONTRIBUTING. NARRATIVE.” Consequences were identified using headings including “PROPERTY DAMAGED” and “NATURE OF DAMAGE”, while protective actions were extracted from sections such as “LIST THE ADDITIONAL INFORMATION” and “RECOMMENDATIONS TO PREVENT RECURRENT NARRATIVE.” It is worth noting that the BSEE reports do not explicitly incorporate deviations, which may limit the model’s ability to learn deviation-specific patterns and could affect the completeness of generated HAZOP worksheets.

During data preprocessing, encoding artifacts, formatting noise, and irregular characters were removed. The extracted text was normalized and segmented into contextually coherent chunks with a maximum length of 512 tokens. Longer narratives were divided into semantically connected sub-chunks. This procedure resulted in 278 node descriptions chunks and 1813 narrative segments. During embedding, representations of the sub-chunks were aggregated using mean pooling to preserve the semantic meaning of complete incident narratives rather than isolated fragments.

For FAISS-based vector indexing, P&ID node descriptions and incident narratives were embedded using the proposed HAZOP-Adapted BERT model to serve as primary semantic representations, while associated elements—including deviations, causes, consequences, and countermeasures—were stored as structured JSONL metadata linked to each indexed entry. This design enables efficient retrieval of semantically relevant cases together with their corresponding HAZOP elements, thereby enhancing the accuracy, consistency, and reliability of downstream RAG-based HAZOP worksheet generation.

To ensure data quality, we verified that every HAZOP record is correctly attached to its corresponding node and that the number and ordering of embedded node-description vectors exactly match the node-level metadata entries used during FAISS retrieval. We also quantified

node coverage and inspected the distribution of records per node (ranging from 8 to 440) and checked for duplicate HAZOP records within each node. For the BSEE corpus, we validated that each narrative chunk retains non-empty associated elements (causes, consequences, and recommendations) in the stored metadata, ensuring that retrieved incidents provide complete contextual grounding for generation.

To evaluate the proposed method, a P&ID node was randomly selected from the literature to provide a proof of concept, for which a conventional HAZOP study had been conducted (see Appendix B.2). The selected node represents a water injection system in which vessel V-106 functions as a buffer between the potable water intake and the crude oil pipeline injection point. The P&ID node configuration is illustrated in Fig. 7 (Marhavi et al., 2019).

5. Results and Discussion

Building on the methodology described in Section 3, this section presents the results of evaluating the performance of multimodal LLMs in generating P&ID node descriptions, HAZOP-Adapted BERT in enhancing hazard related semantic representations, and RAG and baseline LLMs in generating HAZOP worksheets. The following subsections report and discuss the detailed results for each aspect. It should be noted that the node description and associated HAZOP records of the selected P&ID nodes were excluded from the retrieval corpus to ensure an unbiased generation and evaluation.

The experiments were performed in Google Colaboratory Pro using Python 3.12.12 and PyTorch with CUDA support. Computations were accelerated using an NVIDIA A100-SXM4 GPU equipped with 80 GB VRAM. The system was configured with CUDA 12.4 and NVIDIA driver version 550.54.15, providing sufficient memory and computational capacity for large-scale multimodal model evaluation. The LLaMA 3.2 model was accessed through the Hugging Face platform, while GPT-4o and GPT-4o-mini were accessed via the OpenAI API. In addition, the RoBERTa-large model was employed to generate node description embeddings for comparative analysis with a previous study that adopted the same embedding model. SpaCy library was used to calculate cross-model similarity. Supporting libraries, including NumPy, Matplotlib, and Seaborn were used for visualization, and image preprocessing.

The top-K retrieved content used to support LLM reasoning was set to

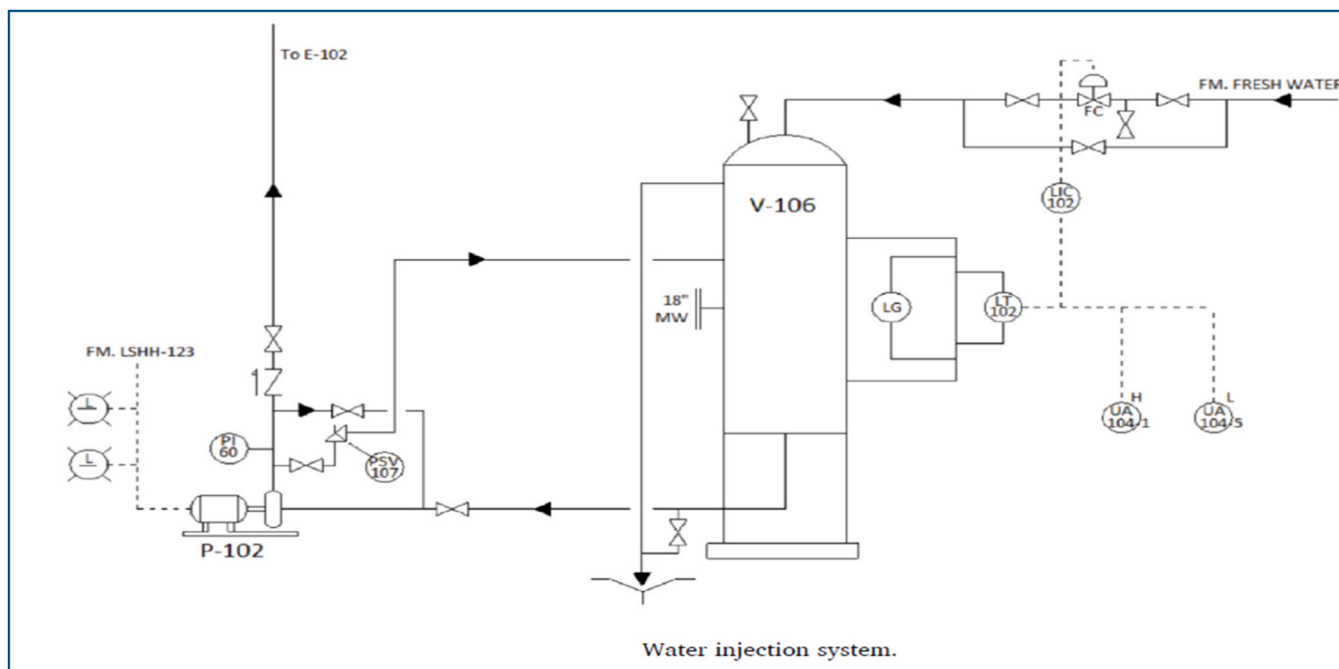


Fig. 7. The reference P&ID node (Node #1) (Marhavi et al., 2019).

10, comprising five nodes retrieved from the P&ID node descriptions and their associated HAZOP records, and five items retrieved from the BSEE incident narratives along with their associated causes, consequences, and recommendations. Table 5 summarizes the experimental configurations of both baseline and RAG-based models, highlighting the differences in input, prompt settings, retrieval sources, and embedding model. The maximum number of tokens was set to 1200 to accommodate the length of the reference HAZOP worksheet, which contains approximately 900 tokens.

5.1. Multimodal LLM performance for node description generation

The performance of GPT-4o and GPT-4o-mini in generating P&ID node descriptions was evaluated using BERT-based similarity metrics by comparing the generated descriptions with literature-based benchmarks for 33 nodes, as summarized in Table 6. Both models demonstrated strong and consistent performance across the three temperature settings (0.0, 0.2, and 0.4) and over ten repeated runs per temperature for each node, indicating stable and reproducible generation behavior. GPT-4o achieved the highest semantic alignment with the benchmark descriptions, attaining a maximum F1-score of 86.88% at temperature 0.2. This result reflects its strong capability to capture both the functional intent and structural characteristics of P&ID nodes. GPT-4o-mini exhibited comparable performance, with its highest F1-score of 86.67% at temperature 0.0, suggesting that the lighter-weight model can still generate high-quality technical descriptions when operated under more deterministic conditions. The low standard deviations observed for all metrics (generally below 0.005) confirm that the generated descriptions remain consistent across repeated runs. This stability is essential for engineering applications, where reproducibility and reliability are critical requirements.

In addition, GPT-4o consistently maintained balanced precision and recall values across temperature settings, indicating its ability to generate content that is both relevant and comprehensive. In contrast, GPT-4o-mini at temperature 0.2 exhibited the lowest precision (85.47%), suggesting a higher tendency to introduce auxiliary or weakly supported details that are not strongly grounded in the benchmark descriptions. At the same time, its lowest recall (86.96%) indicates that some benchmark-specific functional elements may be omitted or expressed more concisely. Overall, the relatively close F1-scores across both models, temperature settings, and repeated runs indicate substantial alignment with the literature descriptions, supporting the feasibility of using multimodal LLM-generated node descriptions as reliable inputs for RAG-based HAZOP worksheet generation.

When compared with the results of our previous study using four multimodal LLMs—Gemini 2.5-Flash, LLaMA-1.5, DeepSeek-VL, and Qwen2-VL—the GPT-based models show a clear improvement in benchmark alignment. The highest F1-score reported in the previous study was 82.89% for LLaMa-1.5 followed by Gemini (82.30%), whereas GPT-4o and GPT-4o-mini achieved F1-scores exceeding 86%, representing an improvement of roughly 4% in semantic similarity. This performance improvement indicates that the newer GPT-4o-based models provide more accurate and semantically faithful P&ID node descriptions, likely due to their enhanced multimodal reasoning capabilities and improved alignment between visual and textual

Table 6

Similarity scores for P&ID node descriptions generated by multimodal LLMs.

Model	Temperature	BERT_P	BERT_R	BERT_F1
GPT-4o	0.0	0.8641	0.8710	0.8676
		± 0.0034	± 0.0031	± 0.0024
		0.8662	0.8713	0.8688
	0.2	± 0.0023	± 0.0021	± 0.0020
		0.8620	0.8730	0.8675
	0.4	± 0.0054	± 0.0050	± 0.0044
GPT-4o-mini	0.0	0.8630	0.8704	0.8667
		± 0.0069	± 0.0026	± 0.0045
		0.8547	0.8696	0.8621
	0.2	± 0.0031	± 0.0019	± 0.0018
		0.8581	0.8715	0.8647
	0.4	± 0.0041	± 0.0023	± 0.0030

representations.

An important aspect of the reliability of the generated descriptions is the degree of agreement among different multimodal LLMs. To assess this aspect, the spaCy library was used to compute document-level cosine similarity between the standardized outputs of all models. The resulting cross-model similarity matrix, shown in Fig. 8, reveals consistently high pairwise agreement, with average similarity values ranging from 0.84 to 0.99. Notably, GPT-4o and GPT-4o-mini exhibit near-identical outputs (0.99), indicating that the lightweight variant closely reproduces the descriptive patterns of the full model. Similarly, strong agreement is observed between GPT-4o and Gemini (0.97), GPT-4o and Qwen (0.96), and among Gemini, LLaMA, and Qwen (0.90–0.93), reflecting a shared interpretation of node functionality and process connectivity.

Overall, the high level of cross-model agreement demonstrates that the generated P&ID node descriptions are not strongly dependent on a specific LLM architecture. Instead, they reflect a stable and reproducible

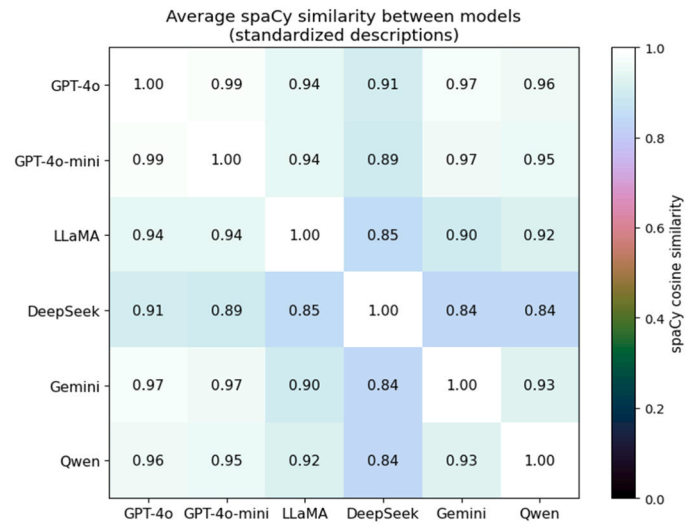


Fig. 8. Cross-model similarity of multimodal LLMs' generated descriptions.

Table 5

Experimental configurations of implementing models for HAZOP worksheet generation.

Models	Input	Prompt	Retrieval sources	Embedding model	Top-K	Runs
RAG-based Models (3 models)	P&ID node image, node description, and retrieved content	Role-based CoT, temperatures (0.1, 0.2, 0.4), maximum tokens (1200)	Node descriptions and associated HAZOP records & BSEE incident narratives and associated elements	HAZOP-Adapted BERT	10	90 runs (30 per model)
Baseline LLMs ((3 models)	P&ID node image and node description	Role-based CoT, temperatures (0.1, 0.2, 0.4), maximum tokens (1200)	–	–	–	90 runs (30 per model)

understanding of the underlying process systems. This finding is consistent with the BERT-based similarity results reported in Table 6. Together, these complementary analyses provide strong evidence of reliability: BERT-based similarity confirms semantic alignment with authoritative descriptions, while spaCy-based agreement demonstrates internal consistency across models.

In addition to the quantitative evaluation, a qualitative manual review was conducted to identify potential errors, omissions, and inconsistencies in the generated descriptions. This review focused on the key aspects specified in the prompt, including design intent, operating conditions, process flow and connectivity, and inlet and outlet streams, in order to assess technical correctness and functional plausibility. Table 7 presents a representative example based on the P&ID node analyzed in this study, in which the GPT-4o-generated description is compared with the literature benchmark in an aspect-wise manner. The results show strong agreement in core functional elements, including the identification of the water injection system, inlet and outlet streams, and the sequential flow from V-106 to P-102 and E-102.

Minor deviations were observed in secondary details, such as the inclusion of pressure protection and monitoring elements that are visible in the P&ID node but not explicitly stated in the literature. These cases were classified as acceptable elaborations rather than hallucinations, as they are supported by diagram evidence. Overall, this qualitative analysis complements the BERT-based metrics and cross-model agreement by providing domain-level validation of technical accuracy and interpretability, thereby strengthening confidence in the reliability of the generated descriptions for subsequent RAG-based HAZOP worksheet generation.

5.2. Performance of the HAZOP-adapted BERT model

The baseline embedding model employed in this study is bert-base-uncased, which follows the original BERT transformer architecture proposed by Devlin et al. (2019). The model consists of 12 transformer layers with a hidden dimension of 768 and 12 self-attention heads, resulting in approximately 110 million trainable parameters. It was pre-trained on large-scale general-domain corpora, including Wikipedia and BookCorpus, comprising approximately 3.3 billion words, using masked language modeling (MLM) and next sentence prediction models. This pre-training enables the model to capture rich contextual and semantic representations.

The proposed HAZOP-Adapted BERT model was further pre-trained using the MLM model with a cross-entropy loss function. Domain-adaptive pre-training was conducted for two epochs with a batch size of 8 and a learning rate of 5×10^{-5} , employing the AdamW optimizer with a weight decay of 0.01.

The effectiveness of domain-adaptive pre-training was evaluated by comparing the baseline model with the adapted model on the held-out test set using cross-entropy loss and perplexity. The baseline BERT model achieved a cross-entropy loss of 2.82, corresponding to a

perplexity of 16.74. In contrast, the HAZOP-adapted model attained a substantially lower loss of 1.66 and a perplexity of 5.24, demonstrating a significant improvement in modeling domain-specific language patterns.

The resulting reduction in entropy loss (1.16) and perplexity (11.50) indicates that the adapted model assigns significantly higher likelihood to HAZOP-related text compared to the generic pre-trained model. This improvement confirms that further pre-training on safety reports, regulatory documents, and technical literature enhances the perception capability for HAZOP knowledge and improves the effectiveness of semantic representation for this knowledge. Consequently, the adapted model exhibits stronger predictive capability and reduced uncertainty when processing domain-relevant sequences.

In addition to quantitative evaluation, a qualitative assessment was conducted using 30 sentences randomly selected from the test set (representative examples provided in Appendix C). The HAZOP-Adapted BERT model outperformed the baseline BERT, achieving higher Top-1 accuracy (0.452 vs. 0.387), Top-5 accuracy (0.903 vs. 0.645), and mean prediction probability (0.267 vs. 0.187). For example, in the sentence “High pressure deviation may damage process [MASK],” the adapted model correctly predicted “equipment” with a higher confidence (0.4347) compared to the baseline model (0.0329). Similarly, for “A high level alarm should [MASK] the emergency shutdown,” the adapted model correctly predicted “initiate” with higher confidence and ranking, while the baseline model assigned a lower rank and probability. These results confirm that domain-adaptive pre-training improves both the accuracy and confidence of masked token prediction, particularly for HAZOP-specific terminology.

5.3. RAG-based semantic similarity performance

The semantic and contextual similarity between the generated HAZOP worksheets and the reference worksheet was evaluated using BERT-based precision, recall, and F1-score metrics, as summarized in Table 8. The results demonstrate a consistent and substantial performance improvement when RAG is employed. Among all models, RAG + GPT-4o achieved the highest performance, with an F1 score of 0.9397, followed closely by RAG + GPT-4o-mini (0.9387). RAG + LLaMA 3.2 VL

Table 8

Similarity scores of RAG-based and baseline LLMs to generate HAZOP worksheets.

Model	BERT_P	BERT_R	BERT_F1
RAG + GPT-4o	0.9383 ± 0.0054	0.9412 ± 0.0040	0.9397 ± 0.0042
RAG + GPT-4o-mini	0.9381 ± 0.0053	0.9393 ± 0.0059	0.9387 ± 0.0052
RAG + LLaMa3.2 VL	0.9226 ± 0.0169	0.9187 ± 0.0220	0.9206 ± 0.0190
GPT-4o	0.8825 ± 0.0091	0.8615 ± 0.0191	0.8718 ± 0.0132
GPT-4o-mini	0.8652 ± 0.0129	0.8698 ± 0.0162	0.8675 ± 0.0135
LlaMa3.2 VL	0.8611 ± 0.0072	0.8565 ± 0.0166	0.8587 ± 0.0094

Table 7

Comparison of literature-based and GPT-4o-generated descriptions for the P&ID node used in this study.

Aspect	GPT-4o Description	Literature Description	Alignment	Notes
Design Intent	Identifies node as a water injection system with vessel V-106 and pump P-102 supplying E-102	V-106 acts as a buffer between potable water intake and crude oil injection point; supports water injection via P-102	Accurate	
Operating Conditions	Mentions level control and overpressure protection	Implied normal water storage and controlled level; no explicit pressure/temperature stated	Partially Accurate	Level control is supported; overpressure protection is not clearly shown in reference
Input Streams	Fresh water via flow control valve	Potable water intake into V-106	Accurate	
Output Streams	Outflow to E-102 via P-102	Water from vessel to P-102, then to E-102	Accurate	
Process Flow & Connectivity	Same sequential flow described	Intake → V-106 → P-102 → E-102	Accurate	
Control, Safety, and Utilities	Mentions LT-102 and LIC-102, P-102 outflow, PSV-107, PI-60, LSHH-123, and UA-1041/1045	LT-102 controls LV-102; P-102 takes suction from vessel; no explicit mention of pressure protection, monitoring, alarms, or utilities	Acceptable	Additional elements are grounded in the P&ID but not stated in the benchmark

also showed strong performance, achieving an F1 score of 0.9206. In addition, the relatively low standard deviations observed for the RAG-based models indicate stable and consistent performance across different temperature settings and multiple trials, suggesting that retrieval grounding reduces the stochastic variability inherent in generative models.

In contrast, baseline models operating without external knowledge exhibited noticeably lower similarity scores. GPT-4o, GPT-4o-mini, and LLaMA 3.2 VL achieved F1 scores of 0.8718, 0.8675, and 0.8587, respectively. This performance gap highlights the limitations of stand-alone LLMs in consistently capturing domain-specific safety knowledge solely from pretrained parameters compared to RAG-based models.

A consistent improvement is observed in both precision and recall for all RAG-based models compared with their baseline counterparts. Higher precision indicates that retrieved knowledge helps reduce the generation of irrelevant or unsupported content, while improved recall reflects enhanced coverage of critical HAZOP elements. These findings confirm that RAG not only enhances contextual fidelity but also reduces the generation of unsupported or extraneous content, leading to more accurate and reliable HAZOP worksheet generation compared to standalone LLMs.

Notably, the performance difference between RAG + GPT-4o and RAG + GPT-4o-mini is marginal, suggesting that effective retrieval can partially compensate for differences in model capacity. This finding indicates that retrieval quality plays a central role in domain-specific generation and may reduce dependence on very large language models in practical deployments.

In addition, statistical significance analysis was conducted to evaluate whether the observed performance improvements are attributable to the augmented external knowledge rather than random variation. A paired *t*-test was performed across the 30 runs for each model, comparing the F1-scores of RAG-augmented models with their corresponding baseline models. The results indicate that all RAG-based models significantly outperform their baseline counterparts, with *p*-values less than 0.05, and in most cases below 0.001. These findings confirm that the observed improvements in semantic similarity are statistically significant and not due to stochastic variation in the generation process.

5.4. HAZOP-based assessment: scenario results

Scenario identification represents a fundamental component of the HAZOP methodology, as it reflects a model's ability to systematically anticipate potential deviations, causes, and consequences within a process system. In this study, scenario identification performance is

evaluated using two complementary indicators: (i) the total number of generated scenarios and (ii) the valid scenario ratio. Together, these metrics provide a quantitative basis for assessing both the generative capacity and the semantic reliability of each model.

5.4.1. Scenario generation capacity

As shown in Fig. 9, the number of generated HAZOP scenarios varies substantially across models and temperature settings (0.1, 0.2, and 0.4). Among the baseline models, GPT-4o exhibits stable and conservative generation behavior, producing a total of 563 scenarios (18.77 per worksheet), which closely approximates the reference HAZOP worksheet containing 17 scenarios. This consistency across temperatures indicates a balanced generation strategy that avoids excessive over-production while maintaining adequate scenario coverage.

GPT-4o-mini generates fewer scenarios overall, with 372 scenarios (12.40 per worksheet). Its median values remain below the reference benchmark, particularly at higher temperatures, suggesting limited exploratory capability and a tendency toward under-generation. This behavior reflects the constrained reasoning capacity of smaller models when operating without external knowledge support. In contrast, LLaMA 3.2 demonstrates substantially higher generative capacity, producing 842 scenarios (28.07 per worksheet). This trend is most pronounced at temperature 0.4 (384 scenarios), where both the median and dispersion increase considerably, frequently exceeding the reference scenario count. While this behavior indicates strong exploratory potential, it also suggests a higher likelihood of redundant, marginally relevant, or weakly grounded scenarios.

The incorporation of external HAZOP knowledge leads to notable changes in scenario generation behavior. For GPT-4o and GPT-4o-mini, RAG produces a moderating effect, yielding tighter distributions and medians closer to the reference benchmark. Specifically, RAG + GPT-4o generates 496 scenarios (16.53 per worksheet), while RAG + GPT-4o-mini produces 411 scenarios (13.70 per worksheet). Both models exhibit reduced variance and improved consistency compared to their baseline counterparts. These results indicate that retrieval grounding constrains scenario generation by anchoring the models to relevant historical HAZOP records and incident narratives, thereby reducing over-generation and improving focus on technically meaningful scenarios. For LLaMA 3.2, RAG integration preserves high generative capacity while enhancing stability. RAG-LLaMA 3.2 produces 910 scenarios (30.33 per worksheet), maintaining medians above the reference level but with reduced dispersion relative to the baseline model. This indicates that retrieval augmentation partially constrains excessive variability while retaining the model's strong reasoning and exploration capabilities.

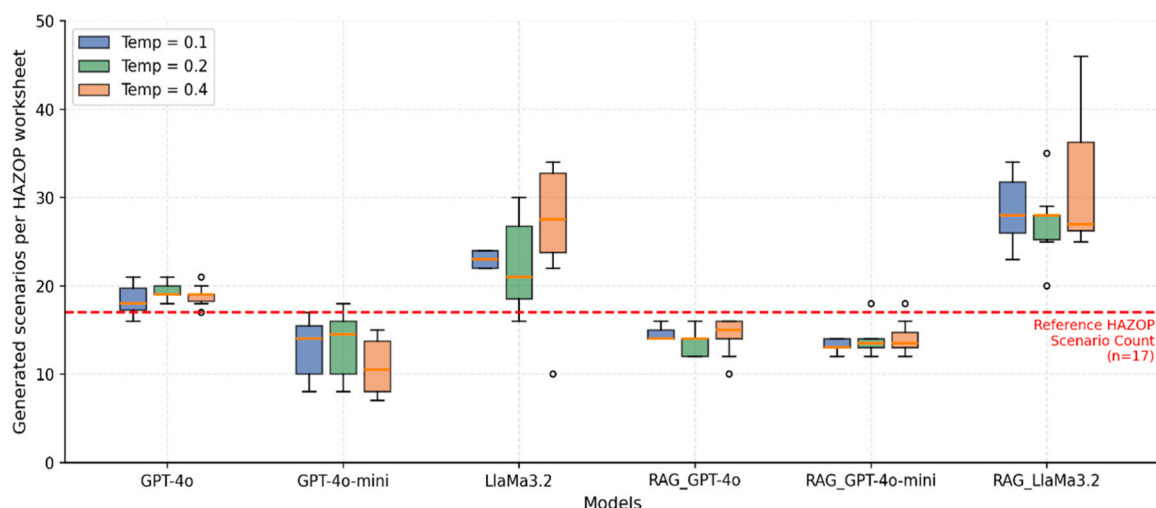


Fig. 9. Scenario generation comparison across models at varying temperatures.

Among all evaluated models, RAG + GPT-4o achieves the closest alignment with the reference worksheet in terms of scenario count, suggesting an optimal balance between coverage and precision. In addition, RAG augmentation enhances the generative capability of GPT-4o-mini, increasing the average number of identified scenarios from 12.4 in the baseline model to 13.7, while maintaining improved consistency and relevance.

Scenario diversity across temperature settings reveals additional performance differences among the evaluated models. Increasing the temperature is generally expected to promote output variability by introducing greater stochasticity during generation. This effect is most evident for LLaMA 3.2, which exhibits wider interquartile ranges at higher temperatures (0.4), indicating increased dispersion in the number of generated HAZOP scenarios. In contrast, GPT-4o and GPT-4o-mini show relatively stable scenario counts across temperatures (0.1–0.4), suggesting limited sensitivity of scenario generation to temperature changes.

RAG-augmented models demonstrate even greater consistency to temperature changes. Their narrower interquartile ranges and scenario counts clustered closer to the reference HAZOP benchmark ($n = 17$) indicate that retrieval-based grounding mitigates temperature-induced variability. This stabilization effect confirms that external knowledge integration constrains random exploration and promotes consistent scenario generation. These results suggest that temperature adjustments primarily influence surface-level linguistic diversity rather than the structured reasoning required for systematic scenario identification.

This observation aligns with prior findings that temperature tuning alone has a limited impact on performance in structured engineering where domain knowledge and reasoning frameworks play a more critical role (Lee et al., 2026; Peeperkorn et al., 2024).

Overall, scenario analysis reveals that baseline models exhibit varying trade-offs between coverage and stability, with larger models tending toward over-generation and smaller models toward under-generation. The integration of RAG consistently improves alignment with expert practices by moderating excessive variability, enhancing consistency, and reducing unsupported outputs. Among all evaluated configurations, RAG + GPT-4o achieves the most balanced performance, combining near-reference scenario counts with low variance. These findings demonstrate that retrieval-augmented architectures are particularly effective for systematic hazard identification tasks, where controlled reasoning and grounding in historical knowledge are essential.

5.4.2. Valid scenario ratio

Following the analysis of scenario generation volume, the semantic validity of generated scenarios was evaluated using clustering-based analysis. High-dimensional embeddings of individual HAZOP scenarios, obtained using HAZOP-Adapted BERT model, were first projected onto two principal components using PCA. This transformation preserves the dominant semantic variance while enabling visual inspection of scenario distributions in a reduced dimensional space (Fig. 10 and Fig. 11). To determine the appropriate number of semantic

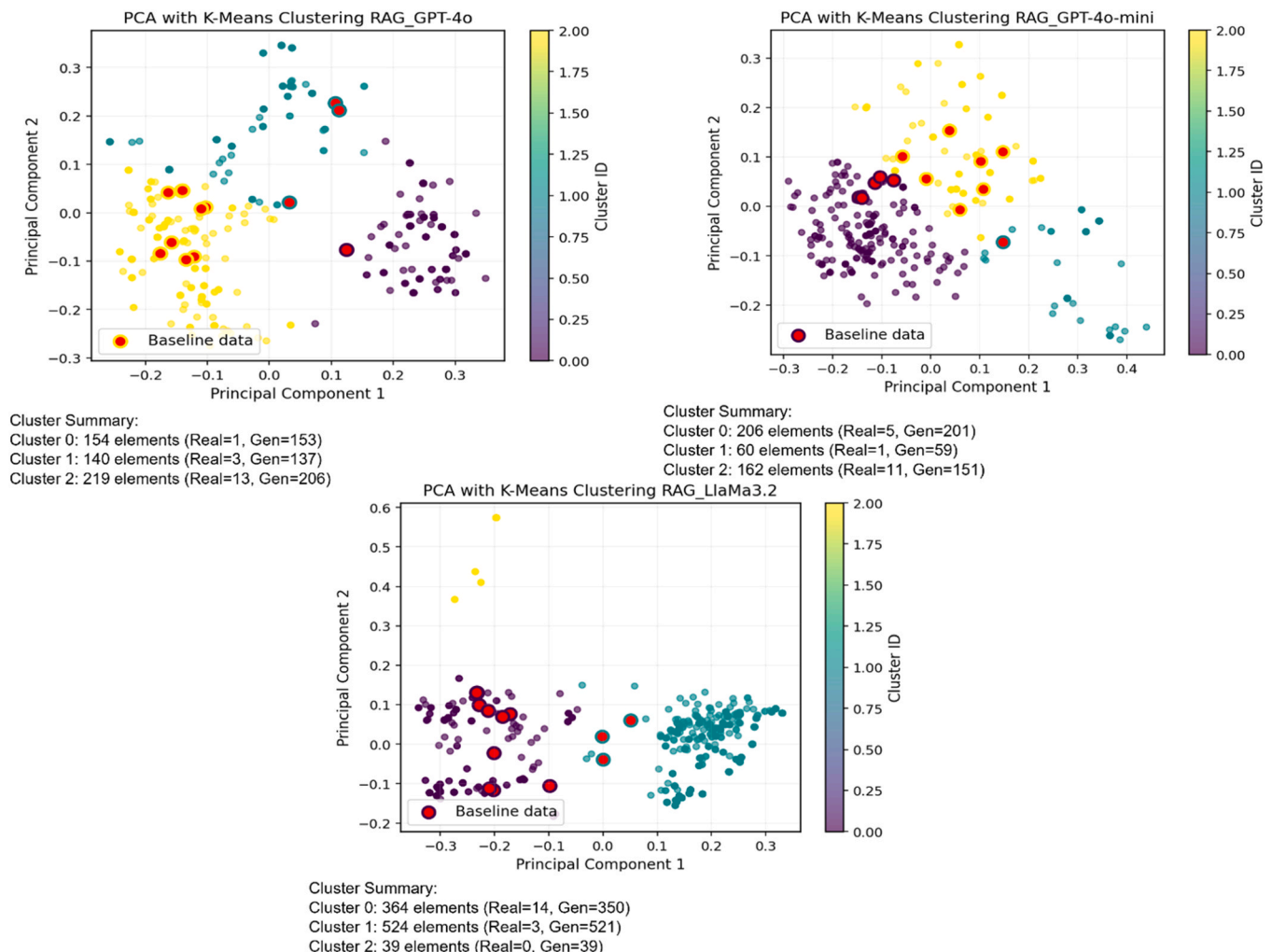


Fig. 10. PCA with K-Means clustering for RAG-based models.

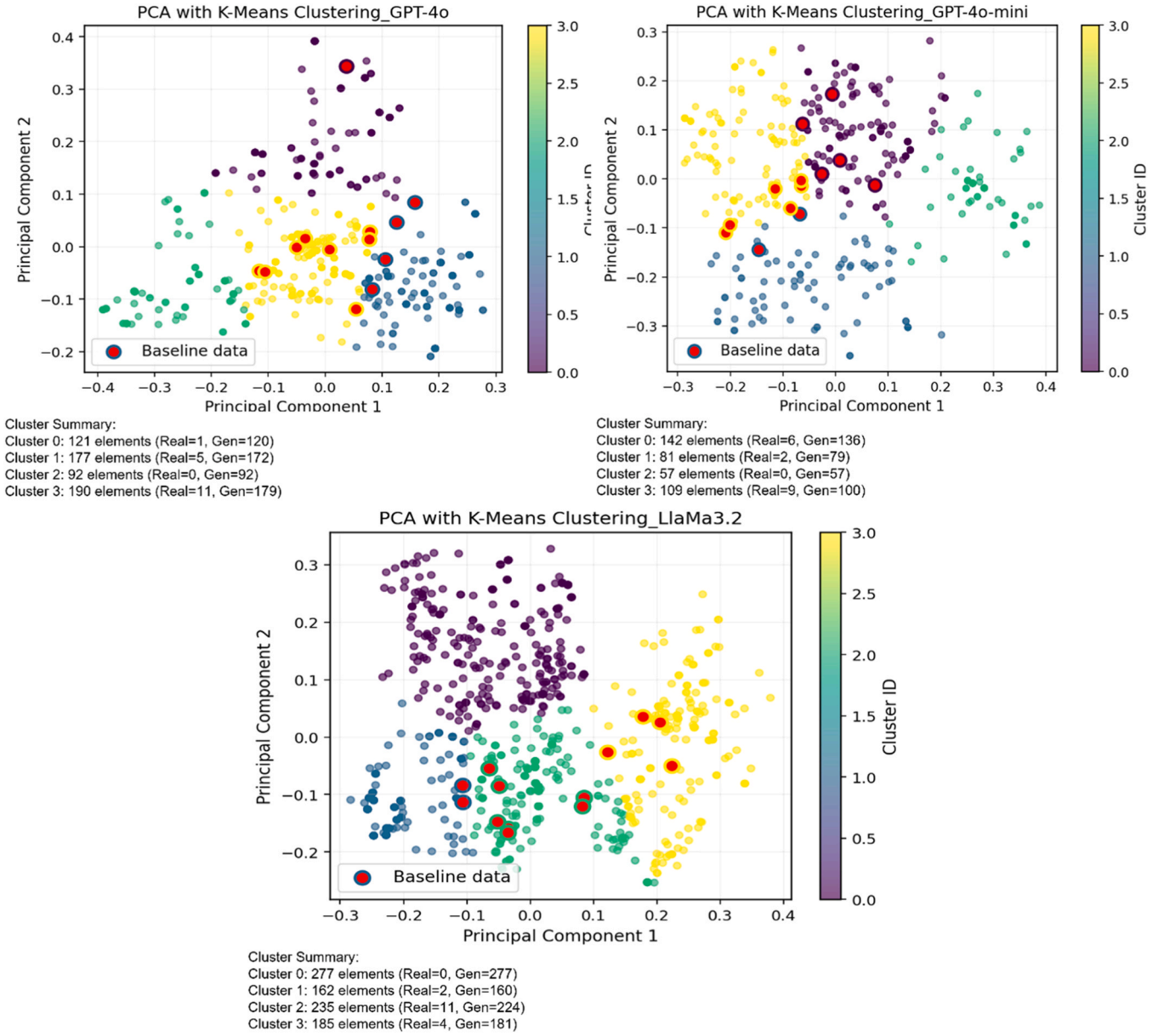


Fig. 11. PCA with K-Means clustering for LLMs-based models.

clusters, the elbow method was applied to the within-cluster sum of squares (WCSS). As illustrated in Fig. 12 and Fig. 13, the elbow criterion consistently identified an optimal value of $k = 3$ for the RAG-augmented models and $k = 4$ for the baseline LLMs. These values reflect the intrinsic semantic structure of the generated scenarios under each modeling paradigm.

Fig. 10 presents the PCA projections with K-means clustering for the RAG-based models. In these visualizations, red markers denote expert-derived reference scenarios, while colored points represent RAG-generated scenarios grouped according to their cluster assignments. For RAG + GPT-4o, the expert scenarios are largely concentrated within a dominant cluster, indicating strong semantic convergence between generated and reference outputs. A similar pattern is observed for RAG + GPT-4o-mini, where expert scenarios form a compact cluster with limited dispersion, suggesting stable semantic grounding despite lower overall scenario counts. RAG + LLaMA 3.2 also exhibits clear cluster formation, with expert scenarios embedded within a major cluster, although accompanied by a broader distribution of generated outputs.

Quantitatively, the clustering results reveal that RAG-based models

achieve substantially higher valid scenario ratios than their baseline counterparts, with improvements of approximately 10–12%. RAG + GPT-4o attained the highest valid scenario ratio of 41.53%, with 206 out of 496 generated scenarios aligned with the expert-dominant cluster. RAG + GPT-4o-mini achieved a ratio of 36.74% (151 out of 411 generated scenarios), while RAG + LLaMA 3.2 reached 38.46% (350 out of 910 generated scenarios), despite producing a significantly larger number of scenarios. These findings indicate that retrieval augmentation improves not only the quantity of generated scenarios but, more importantly, their semantic relevance to expert knowledge.

In contrast, baseline LLMs exhibited lower valid scenario ratios despite comparable or higher scenario counts as shown in Fig. 11. GPT-4o achieved a valid ratio of 31.79% (179 out of 563 generated scenarios), while GPT-4o-mini and LLaMa-3.2 recorded 26.88% (100 out of 372 generated scenarios) and 26.60% (224 out of 842 generated scenarios), respectively. The PCA projections show that baseline models tend to distribute generated scenarios across multiple clusters, indicating weaker semantic cohesion and reduced alignment with expert benchmarks. This dispersion is particularly evident for LLaMa-3.2,

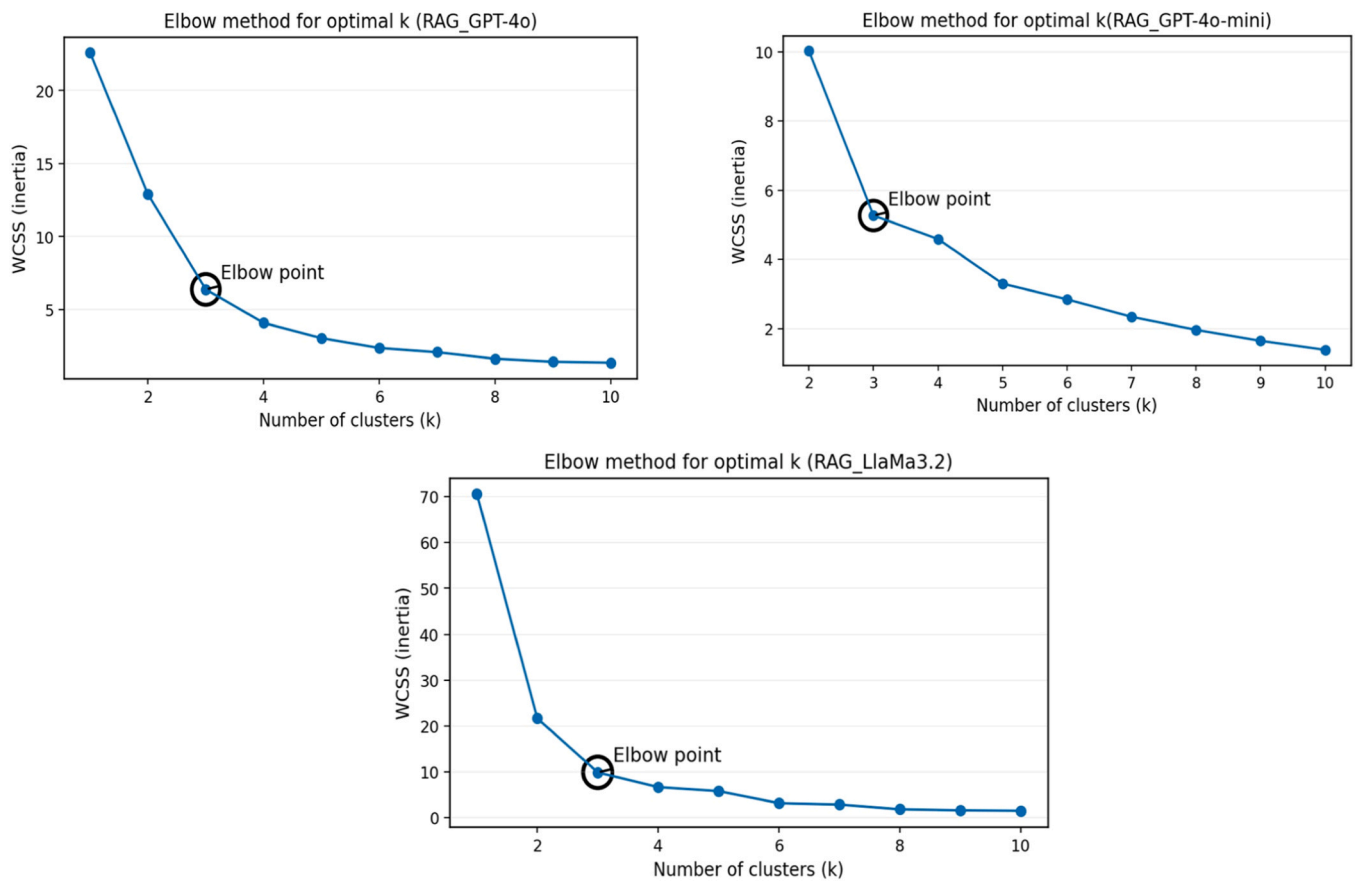


Fig. 12. Optimal number of clusters for RAG-based models determined by Elbow method.

which, although producing the highest volume of scenarios, exhibits substantial semantic spread and lower consistency.

Overall, these findings highlight a clear trade-off between scenario quantity and validity. While baseline models—especially LLaMa-3.2—tend to maximize scenario coverage, they do so at the expense of semantic coherence. In contrast, RAG-augmented models consistently generate scenarios that are more tightly aligned with expert knowledge, reflecting improved contextual grounding and reduced hallucination. Furthermore, the observed valid scenario ratios exceed those reported in prior studies, which typically range between 19% and 37%. This improvement confirms the effectiveness of integrating domain-adapted embeddings, dense retrieval mechanisms, and role-based CoT prompting in enhancing the reliability and interpretability of automated HAZOP scenario generation. Collectively, these findings demonstrate that retrieval grounding plays a critical role in transforming LLMs from high-volume generators into knowledge-aligned decision-support tools suitable for safety-critical engineering applications.

To further validate the reliability of the clustering-based valid scenario ratio, a complementary manual evaluation was conducted. For each RAG-based model, twelve generated HAZOP worksheets were randomly selected and independently reviewed to identify scenarios that were semantically consistent with the reference scenarios. The results indicate that RAG + GPT-4o produced 74 correct scenarios out of 166 generated scenarios, corresponding to validity ratios ranging from 35.7% to 57.1%. RAG + GPT-4o-mini achieved 68 correct scenarios out of 174, with validity ratios between 33.0% and 46.2%. Although RAG + LLaMa 3.2 generated a substantially larger number of scenarios, only 76 out of 270 were identified as correct, resulting in lower validity ratios ranging from 17.4% to 34.8%. Overall, RAG + GPT-4o achieved the highest manual validity ratio, while RAG + LLaMa 3.2 exhibited the lowest performance. These manually derived results are consistent with

the automated PCA-K-means analysis, indicating that the clustering-based method provides a reliable approximation of expert judgment and supports its application for large-scale evaluation of scenario validity.

5.5. HAZOP-based assessment: Deviation analysis results

Table 9 summarizes the deviation analysis results for both baseline and RAG-augmented models in terms of total deviations, unique deviations, correct deviations, and the valid deviation ratio.

Among the baseline models, LLaMA 3.2 generated the highest number of unique deviations (359), followed by GPT-4o (280) and GPT-4o-mini (235). This indicates that LLaMA 3.2 exhibits the strongest exploratory behavior in identifying diverse deviation types. However, higher diversity does not necessarily translate into higher accuracy. GPT-4o achieved a more balanced performance by generating a substantial number of unique deviations while maintaining stronger alignment with expert references. For the RAG-augmented models, RAG + LLaMa 3.2 produced the largest number of unique deviations (377), slightly exceeding its baseline counterpart. In contrast, RAG + GPT-4o and RAG + GPT-4o-mini generated fewer unique deviations (227 and 184, respectively), suggesting that retrieval grounding constrains deviation exploration and reduces redundant or speculative outputs.

Correct deviations represent the subset of generated deviations that match those in the reference HAZOP worksheet. Among baseline models, GPT-4o achieved the highest number of correct deviations (178), substantially outperforming GPT-4o-mini (113) and LLaMa 3.2 (106). This indicates that GPT-4o possesses stronger inherent domain reasoning capabilities for deviation identification. RAG-based models exhibit a marked improvement in correctness. RAG + GPT-4o achieved the highest number of correct deviations overall (195), representing an

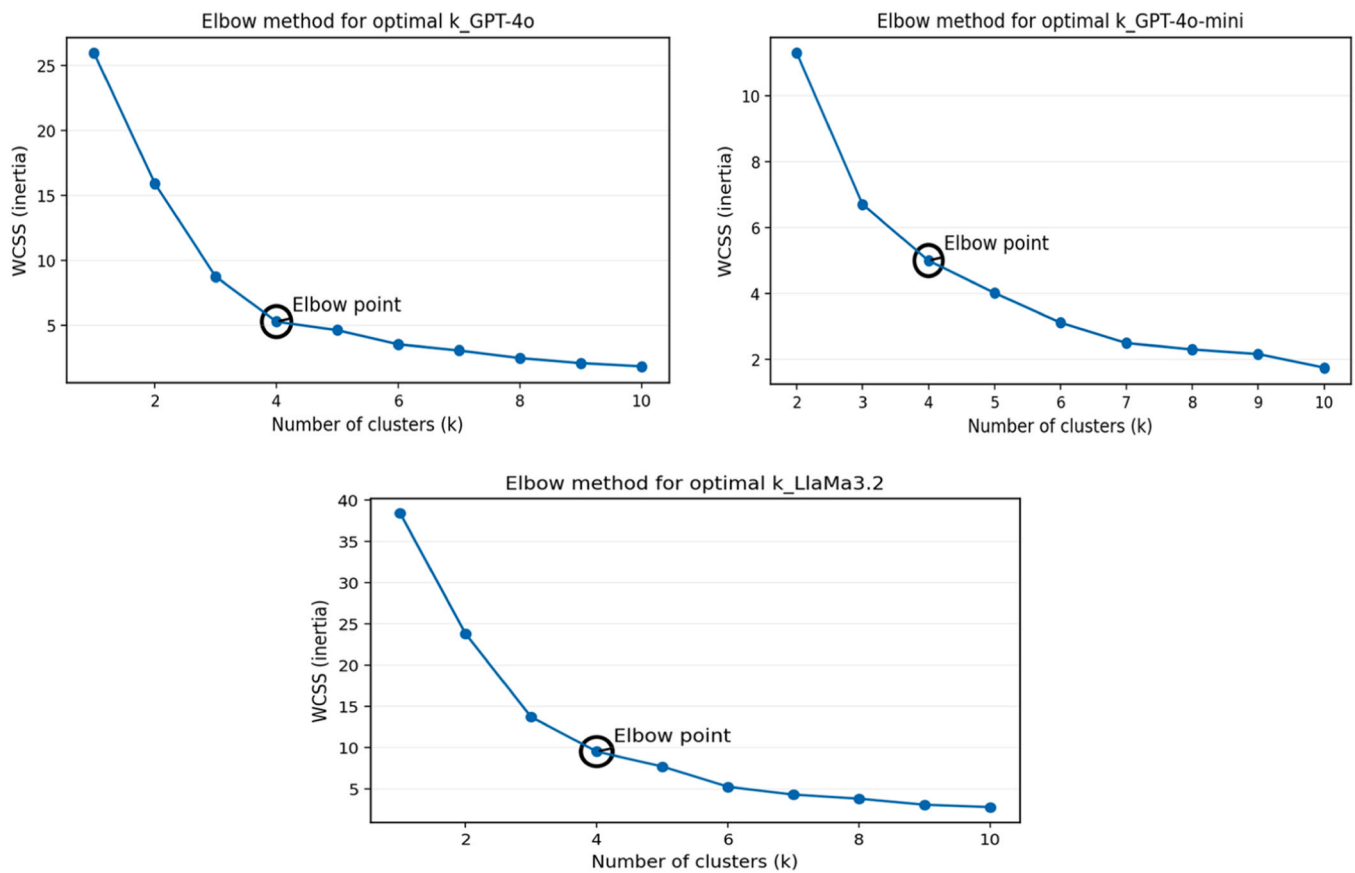


Fig. 13. Optimal number of clusters for LLMs-based models determined by Elbow method.

Table 9

Deviation analysis results for baseline and RAG-augmented models.

Model	Total deviations	Unique deviations	Correct deviations	Average per sheet	Real unique deviations/sheet	Valid deviation ratio	Correct / Unique	Efficiency
GPT-4o	563	280	178	5.93	8	74.13%	178 / 280	63.6%
GPT-4o-mini	372	235	113	3.77	8	47.13%	113 / 235	48.1%
LLaMa3.2	842	359	106	3.53	8	44.13%	106 / 359	29.5%
RAG+ GPT-4o	496	227	195	6.50	8	81.25%	195 / 227	85.9%
RAG + GPT-4o-mini	411	184	148	4.93	8	61.63%	148 / 184	80.4%
RAG + LLaMa3.2	910	377	140	4.67	8	58.38%	140 / 377	37.1%

increase of nearly 10% compared to its baseline version. Similarly, RAG + GPT-4o-mini improved from 113 to 148 correct deviations, reflecting the positive impact of retrieval grounding. RAG + LLaMa 3.2 also improved from 106 to 140 correct deviations, despite maintaining high generative diversity.

The valid deviation ratio reflects the proportion of real deviations successfully identified by each model. Among the baseline models, GPT-4o achieved the highest ratio (74.13%), indicating strong baseline reliability. GPT-4o-mini and LLaMA 3.2 performed substantially worse, with ratios of 47.13% and 44.13%, respectively. RAG augmentation led to systematic improvements across all models. RAG + GPT-4o attained the highest valid deviation ratio (81.25%), demonstrating superior coverage of expert deviations. RAG + GPT-4o-mini increased to 61.63%, while RAG + LLaMA 3.2 reached 58.38%. These gains confirm that retrieval grounding significantly improves deviation validity.

To further evaluate model efficiency, the ratio of correct deviations to unique generated deviations was examined. This ratio reflects how efficiently each model converts generated deviation diversity into correct, expert-aligned outputs. Among the baseline models, LLaMA 3.2

exhibits the lowest efficiency (29.5%), indicating that most of its generated deviations are either redundant or incorrect. GPT-4o demonstrates relatively strong baseline efficiency (63.6%), reflecting better internal filtering mechanisms. RAG-based models show substantial efficiency gains. RAG + GPT-4o achieves the highest efficiency (85.9%), indicating that most unique deviations are correct. RAG + GPT-4o-mini also demonstrates strong efficiency (80.4%), confirming that retrieval grounding substantially reduces noise. Although RAG + LLaMA 3.2 improves over its baseline, its efficiency remains comparatively low (37.1%), suggesting persistent over-generation tendencies.

5.6. HAZOP-based assessment: safeguards and recommendations results

Fig. 14 presents the distribution of safeguards and recommendations across inherently safer design/passive (ISD/Passive), active, and procedural categories for the reference worksheet and all evaluated models. This analysis provides insight into how closely the generated outputs reflect established safety engineering practices. The reference HAZOP worksheet exhibits a balanced safety strategy, with active safeguards

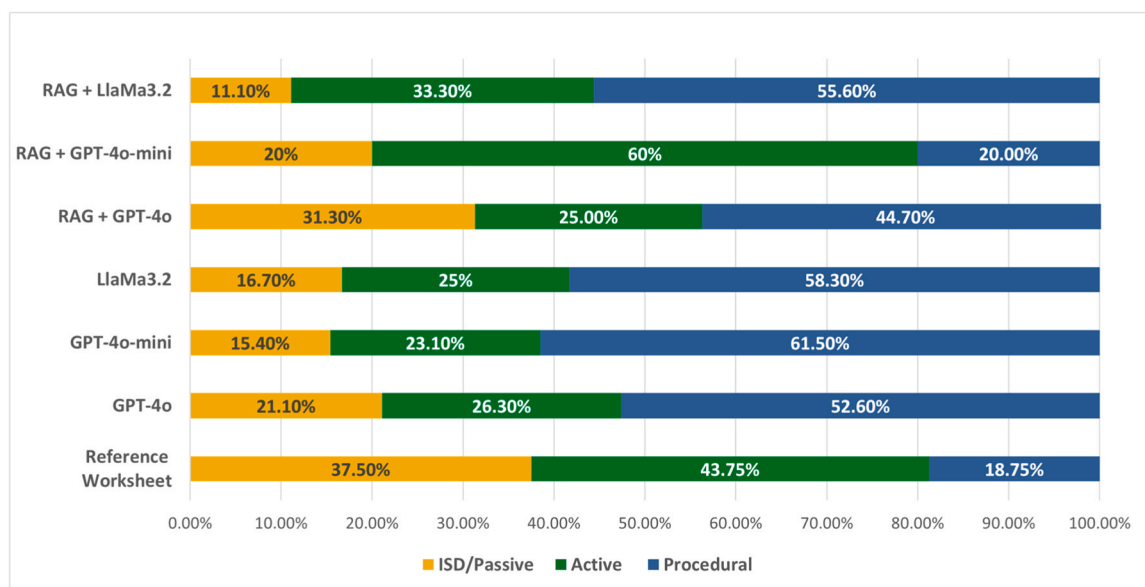


Fig. 14. Safeguard and recommendation distribution analysis across models.

representing the largest proportion (43.75%), followed by ISD/passive measures (37.5%) and procedural controls (18.75%). This distribution reflects industrial practice, in which inherent and engineering controls are prioritized, while administrative measures serve a complementary role.

Among the baseline models, GPT-4o and GPT-4o-mini show a strong reliance on procedural safeguards, accounting for 52.6% and 61.5% of their outputs, respectively. This indicates a tendency to favor inspection, maintenance, and operational instructions over engineering-based protections. LLaMA 3.2 exhibits a similar pattern, with 58.3% procedural safeguards and a relatively low proportion of ISD/passive measures (16.7%). Such distributions suggest that baseline LLMs tend to emphasize human-dependent and administrative controls, which are generally less reliable than physical and automated protections in safety-critical systems. These findings are consistent with prior work, which reported that LLM-generated HAZOP worksheets were dominated by procedural safeguards (approximately 74–81%), reflecting a systematic bias toward maintenance- and instruction-based controls rather than engineered layers of protection (Lee et al., 2026).

Incorporating retrieval-augmented generation leads to noticeable changes in safeguard composition. RAG + GPT-4o demonstrates the closest alignment with the reference worksheet in ISD/passive safeguards, increasing their proportion to 31.3% compared to 21.1% for GPT-4o. It also maintains a balanced contribution from active (25.0%) and procedural (44.7%) measures compared to other models. This shift indicates that retrieval grounding encourages the model to incorporate more design- and equipment-based protections consistent with expert practices. RAG + GPT-4o-mini exhibits a dominant reliance on active safeguards (60.0%), with reduced procedural content (20.0%). This suggests that retrieval support enables smaller models to prioritize instrumentation and automated controls over administrative measures. Similarly, RAG + LLaMA 3.2 reduces procedural dominance compared to its baseline counterpart, however, procedural safeguards (55.6%) remain prevalent, reflecting the model's continued tendency toward narrative and procedural reasoning.

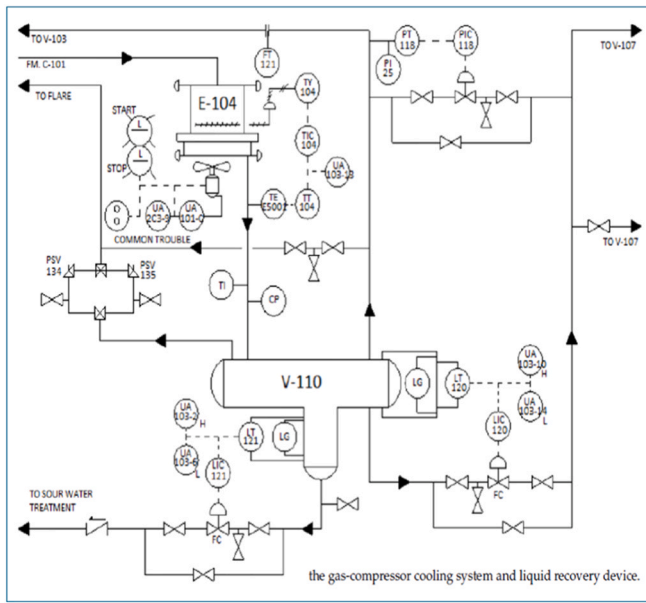
Overall, RAG-augmented models consistently reduce over-reliance on procedural safeguards and promote greater inclusion of inherent and engineering-based protections. This indicates that grounding generation in historical HAZOP records and technical documentation enhances the models' ability to emulate established safety design principles. Among all evaluated models, RAG + GPT-4o achieves a more

balanced distribution and closer correspondence to the reference worksheet compared to other models, demonstrating an improved capability to integrate passive, active, and procedural measures in a manner consistent with professional HAZOP practice. These findings confirm that retrieval augmentation not only improves semantic accuracy but also contributes to practically applicable safety recommendations by encouraging models to prioritize robust engineering controls over human-dependent interventions.

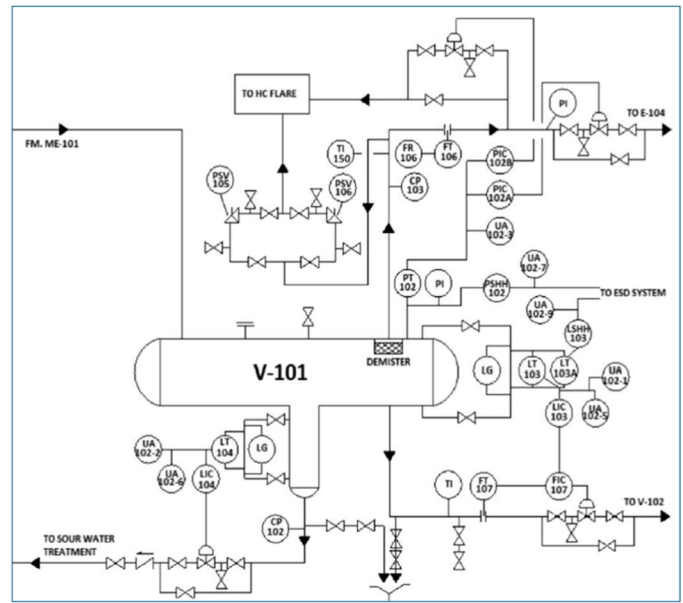
Since the RAG + GPT-4o model achieved the best performance among evaluated models, it was further applied to two additional P&ID nodes to validate the proposed method across different process configurations, as illustrated in Fig. 15. To provide a more comprehensive evaluation, an ablation study was also conducted on these nodes. Specifically, role-based prompting was compared against role-based CoT prompting to evaluate the impact of structured reasoning, while the baseline BERT model (bert-base-uncased) was compared with the HAZOP-Adapted BERT model to assess the contribution of domain-specific semantic representations. Accordingly, six models were evaluated on each node, with each model executed 30 times under three temperature settings (0.1, 0.2, and 0.4). In total, this resulted in 12 model configurations and 360 generated HAZOP worksheets. The same configurations applied to the reference node (Node #1) in this study were applied to these additional nodes. The generated worksheets were evaluated using BERT-based semantic similarity metrics, along with the valid scenario ratio to assess the correctness and completeness of the generated scenarios.

As shown in Fig. 15, the P&ID node #2 represents a gas-compressor cooling and liquid recovery system, while the P&ID node #3 illustrates a low-pressure separator. These nodes are part of a crude oil processing plant that aims to separate the crude oil into oil, gas, and water. The results of the BERT-based similarity metrics and valid scenario ratio are presented in Table 10 for Node #2 and Table 11 for Node #3. In addition, the PCA with K-Means clustering results, along with the optimal number of clusters determined using the Elbow method, are provided in Appendix D for all models across both nodes.

Table 10 and Table 11 demonstrate that integrating RAG, GPT-4o, HAZOP-Adapted BERT, and role-based CoT prompting significantly enhances the ability of LLMs to generate more accurate and reliable HAZOP scenarios. The findings clearly highlight the impact of retrieval grounding. Across both nodes, RAG-based models consistently achieve higher precision, recall, and F1-scores, as well as substantial



Node #2



Node #3

Fig. 15. Node#2 shows a gas-compressor cooling system (Marhavilas et al., 2021), Node#3 shows a low pressure separator (Marhavilas et al., 2020).

Table 10
BERT-based similarity metrics and valid scenario ratio for Node #2.

Model	BERT_P	BERT_R	BERT_F1	Scenario ratio
GPT-4o + Role-based Prompt	0.8428 ± 0.0134	0.8133 ± 0.0068	0.8275 ± 0.0081	25.72%
GPT-4o + Role-based CoT Prompt	0.8607 ± 0.0158	0.8387 ± 0.0075	0.8492 ± 0.0097	26.02%
RAG + GPT-4o + base BERT + Role-based Prompt	0.9095 ± 0.0159	0.8723 ± 0.0123	0.8901 ± 0.0105	34.07%
RAG + GPT-4o + base BERT + Role-based CoT Prompt	0.9206 ± 0.0015	0.9017 ± 0.0043	0.9108 ± 0.0029	38.62%
RAG + GPT-4o + HAZOP-Adapted BERT + Role-based Prompt	0.9245 ± 0.0014	0.9068 ± 0.0041	0.9156 ± 0.0028	39.34%
RAG + GPT-4o + HAZOP-Adapted BERT + Role-based CoT Prompt	0.9492 ± 0.0163	0.9283 ± 0.0076	0.9382 ± 0.0101	43.77%

improvements in the valid scenario ratio. For example, in Node #2, the baseline GPT-4o with role-based prompting achieves an F1-score of 82.75% and a valid scenario ratio of 25.72%, whereas the RAG-based model with HAZOP-Adapted BERT and CoT increases the F1-score to 93.82% and the valid scenario ratio to 43.77%. Similarly, for Node #3, the baseline GPT-4o with role-based prompting achieves an F1-score of 86.04% and a scenario ratio of 30.06%, while the RAG-based configuration improves these values to 93.31% and 48.2%, respectively. These results indicate that incorporating RAG with HAZOP-Adapted BERT and role-based CoT prompting enhances the valid scenario ratio by approximately 18%.

Comparing role-based prompting and role-based CoT prompting demonstrates the benefit of structured reasoning. For example, in Node #2, the F1-score increases from 82.75% to 84.92% when CoT is applied to GPT-4o, and from 91.56% to 93.82% for the RAG + HAZOP-Adapted BERT configuration, with the valid scenario ratio improving from 39.34% to 43.77%. Similarly, for Node #3, CoT improves the F1-score from 86.04% to 86.15% for GPT-4o and from 93.04% to 93.31% for

Table 11
BERT-based similarity metrics and valid scenario ratio for Node #3.

Model	BERT_P	BERT_R	BERT_F1	Scenario ratio
GPT-4o + Role-based Prompt	0.8627 ± 0.0093	0.8499 ± 0.0110	0.8604 ± 0.0077	30.06%
GPT-4o + Role-based CoT Prompt	0.8752 ± 0.0119	0.8495 ± 0.0065	0.8615 ± 0.0081	31.28%
RAG + GPT-4o + base BERT + Role-based Prompt	0.9370 ± 0.0094	0.8933 ± 0.0116	0.9139 ± 0.0078	39.87%
RAG + GPT-4o + base BERT + Role-based CoT Prompt	0.9252 ± 0.0109	0.9007 ± 0.0052	0.9121 ± 0.0066	42.54%
RAG + GPT-4o + HAZOP-Adapted BERT + Role-based Prompt	0.9427 ± 0.0093	0.9199 ± 0.0110	0.9304 ± 0.0077	45.2%
RAG + GPT-4o + HAZOP-Adapted BERT + Role-based CoT Prompt	0.9424 ± 0.0181	0.9253 ± 0.0128	0.9331 ± 0.0128	48.2%

the RAG + HAZOP-Adapted BERT configuration, while the scenario ratio increases from 45.2% to 48.2%. Although the improvements are smaller than those achieved by RAG, they are consistent across both nodes, indicating that guiding the model through step-by-step reasoning enhances the completeness and coherence of the generated HAZOP scenarios.

In addition, the comparison between the baseline BERT and the HAZOP-Adapted BERT highlights the importance of domain-specific semantic representation. Across both nodes, replacing the base BERT with the HAZOP-Adapted BERT leads to noticeable improvements in all evaluation metrics. For example, in Node #2, under the RAG + CoT configuration, the F1-score increases from 91.08% to 93.82%, while the valid scenario ratio improves from 38.62% to 43.77%. For Node #3, the F1-score increases from 91.21% to 93.31%, and the scenario ratio increases from 42.54% to 48.2%. These improvements indicate that the domain-adapted model more effectively captures hazard-related terminology and process-specific relationships, thereby improving retrieval quality and enabling the LLM to generate more accurate and relevant

HAZOP scenarios.

The experimental results across the three P&ID nodes used in this study demonstrate that the proposed domain-augmented method consistently outperforms baseline LLMs across multiple evaluation dimensions. In terms of semantic similarity assessment, RAG-based models incorporating HAZOP-Adapted BERT exhibit stronger alignment with expert-prepared HAZOP worksheets, achieving F1-Score exceeding 92%, compared to approximately 82–87% for baseline models. Scenario analysis further indicates improved valid scenario ratios and more stable scenario generation, indicating closer correspondence with expert reasoning patterns. RAG-based models improved the valid scenario ratio by approximately 10–18%, with the RAG + GPT-4o + HAZOP-Adapted BERT + role-based CoT prompting achieving the highest ratio of 41.53% for the reference node (Node #1), 43.77% for Node #2, and 48.2% for Node #3. The ablation study further reveals that HAZOP-Adapted BERT contributes more to performance improvement (approximately 5–6%) than CoT prompting (approximately 1–4%), highlighting the importance of domain-specific semantic representation in enhancing retrieval quality and downstream generation performance.

The deviation analysis of the reference node further confirms this advantage, as the proposed method generates a higher number of correct deviations and achieves substantially higher valid deviation ratios than baseline models. RAG + GPT-4o achieved the highest valid deviation ratio (81.25%), compared to 58.38% for RAG + LLaMA 3.2. Among baseline models, GPT-4o achieved the best performance (74.13%), while LLaMA 3.2 exhibited the lowest ratio (44.13%). In addition, the safeguard and recommendation distribution analysis reveals that retrieval grounding promotes a more balanced integration of inherently safer, active, and procedural controls, thereby reducing the over-reliance on administrative measures observed in baseline LLMs. Collectively, these results highlight the effectiveness of combining domain-adapted embeddings, dense retrieval, and controlled prompting to enhance the reliability and practical relevance of automatically generated HAZOP worksheets.

It should also be noted that the selected P&ID nodes focus on scenarios associated with a subset of deviation types. For example, in the reference node (Node #1), eight deviations were considered, including more flow, less flow, no flow, reverse flow, high pressure, low level, less level, and loss of containment. While these deviations cover major operational hazards, other relevant deviation types may exist but were not represented in this case. Notably, the proposed method generated additional deviations, such as low and high temperature, low pressure, inadequate isolation, and corrosion, which were not documented in the real worksheets. This observation suggests that the proposed method is capable of identifying plausible hazards beyond the predefined scope, although further expert validation is required to assess their practical relevance.

Despite these promising results, several limitations should be considered for future investigation. First, the proposed method was applied and evaluated using three representative P&ID node, which limits the generalizability of the findings across different process configurations and industrial contexts. Future studies should validate the method on a broader range of nodes and plant sections to establish its robustness. Second, the quality and composition of the training corpus significantly influence model performance. In this study, historical BSEE incident reports were used as a major data source; however, these reports lack explicit deviation annotations, which may constrain the model's ability to fully capture deviation-specific reasoning patterns. Third, the dataset employed for domain-adaptive pretraining of the BERT-base-uncased model represents approximately 0.5% of the original pretraining corpus (3.3 billion words), which may limit the extent to which domain-specific linguistic and semantic knowledge can be internalized. Fourth, the valid scenario ratio remains relatively low (41.53–48.2%) compared with expert-prepared scenarios, indicating that the proposed method should be regarded as a decision-support system that complements, rather than replaces, expert judgment. This

limitation is primarily attributed to the completeness of the knowledge base. Most of the HAZOP data extracted from the literature consists of partial worksheets that include only selected scenarios, while omitting others, resulting in incomplete coverage of HAZOP records. Only 11 actual HAZOP reports were incorporated, which further restricts the diversity and comprehensiveness of the knowledge base. In practical deployment, the knowledge base should include a substantially larger collection of complete HAZOP studies to improve scenario coverage and reliability. Fifth, the top-K retrieval parameter was set to 10 without systematic optimization. Varying this value may affect the performance of RAG-based models in generating HAZOP worksheets. Therefore, a sensitivity analysis should be conducted to evaluate the impact of the K value on model performance. These limitations highlight the need for richer, more comprehensive HAZOP-specific datasets and broader empirical validation, which are expected to further improve scenario validity and enhance the overall reliability and practical applicability of the proposed method.

6. Conclusion

This study proposed an integrated, domain-specific method that combines domain-adaptive language modeling and retrieval-augmented generation (RAG) to support automated HAZOP worksheet generation. The method incorporates multimodal LLMs (GPT-4o and GPT-4o-mini) for generating P&ID node descriptions, HAZOP-Adapted BERT for improving domain-specific semantic representation, and FAISS-based retrieval for knowledge grounding. Unlike prior studies that rely solely on standalone LLMs—often resulting in nonfactual outputs and hallucinations, the proposed method enhances the reasoning capability of LLMs (GPT-4o, GPT-4o-mini, and LLaMA 3.2) by grounding generation in contextual knowledge derived from historical HAZOP worksheets and incident investigation reports. This knowledge-grounded method improves both the accuracy and reliability of the generated HAZOP worksheets.

A total of 180 HAZOP worksheets were generated and benchmarked against a reference HAZOP worksheet extracted from the literature. In addition, an ablation study was conducted on additional two P&ID nodes to evaluate the impact of HAZOP-Adapted BERT and CoT prompting on model performance. The RAG-based models were evaluated and compared with baseline LLMs using two main criteria: (i) model-based assessment, based on semantic similarity (BERT-based metrics), and (ii) HAZOP-based assessment, including scenario generation capacity, deviation analysis, and safeguard and recommendation distribution.

The findings indicate that the proposed method consistently outperforms baseline LLMs across multiple evaluation dimensions. In terms of semantic similarity assessment, RAG-based models incorporating HAZOP-Adapted BERT achieved F1-Score exceeding 92%, compared with approximately 82–87% for baseline models. For scenario analysis, RAG-based models incorporating HAZOP-Adapted BERT improved the valid scenario ratio by approximately 10–18%. In addition, the ablation study shows that HAZOP-Adapted BERT contributes more to performance improvement (approximately 5–6%) than CoT prompting (approximately 1–4%). For deviation analysis of the reference node, RAG + GPT-4o achieved the highest ratio (81.25%), compared to 58.38% for RAG + LLaMA 3.2. In addition, safeguard and recommendation distribution analysis revealed that RAG-based models increased emphasis on passive and active protections, reflecting established industry best practices. Collectively, these results demonstrate the potential of domain augmented LLMs for deployment in process industries to support the generation of more reliable HAZOP worksheets.

Despite these promising results, limitations remain. First, the proposed method was evaluated using three P&ID nodes, which limits the generalizability across different process configurations. Second, the quality and completeness of the training corpus directly affect model performance. In particular, the incident investigation reports used in the

knowledge base do not explicitly include HAZOP deviations. Third, the domain-adaptive pretraining corpus represents only a small fraction (approximately 0.5%) of the original BERT pretraining data, which limit the extent of domain specialization that can be achieved. Fourth, the valid scenario ratio remains relatively low compared with expert-prepared scenarios, indicating that the proposed method should be regarded as a decision-support system that complements, rather than replaces, expert judgment.

Future research will focus on extending the proposed method to multiple P&ID nodes and complete process units to improve generalizability. Expanding the corpus with more HAZOP worksheets from past industrial projects and linking them explicitly to corresponding P&ID nodes is expected to further enhance model performance. In addition, incorporating larger-scale domain-adaptive pretraining is expected to strengthen knowledge representations and retrieval accuracy. A sensitivity analysis of the K value used in content retrieval will be conducted to evaluate how the number of retrieved items affects the ability of LLMs to generate accurate HAZOP worksheets. These directions will contribute toward the development of practical and scalable AI-assisted systems for automated HAZOP report generation.

Appendix A

Prompt to generate P&ID node descriptions.

P&ID and HAZOP Context (GENERAL VIEW):

Piping and Instrumentation Diagrams (P&IDs) represent process equipment, pipelines, valves, and instrumentation for process plants. In HAZOP studies, a P&ID node is a logical section of the P&ID selected for detailed analysis by the HAZOP team to identify potential hazards and operability malfunctions. Each node typically includes one or more major equipment items (e.g., pumps, heat exchangers, vessels) together with the connecting lines, valves, and instruments that define the normal flow path.

ROLE:

You are a process engineer, participated in many HAZOP sessions. Your task is to explain the description of the P&ID node to the HAZOP team for the detailed analysis. Base your answer only on the information visible in the P&ID node image. If something is unclear or not shown, state that it is uncertain instead of guessing.

QUESTION:

Please describe the P&ID node shown in the image, including:

- Design intent and normal operating conditions.
- Inputs and outputs streams to and from the node.
- Process flow and connectivity through the node.

Appendix B

B.1. Prompt to generate HAZOP worksheets.

ROLE:

You are a senior HAZOP facilitator conducting a HAZOP study in accordance with IEC 61882.

INTERNAL REASONING:

Reason step-by-step internally, but do NOT reveal chain-of-thought. Provide only the final structured HAZOP worksheet.

GOAL/TASK:

Use the provided evidence (retrieved node descriptions and their HAZOP scenarios + incident narratives and their causes, consequences, and recommendations) plus the given node image and description to generate HAZOP worksheet.

Analyze the provided node description and P&ID node image and the evidence to identify design intent elements (flow, pressure, temperature, level, containment, composition, isolation, Corrosion), then systematically apply HAZOP guide words (NO/NOT, MORE, LESS, AS WELL AS, PART OF, REVERSE, OTHER THAN) to generate deviations.

You MUST evaluate EACH deviation. If a deviation is not credible for this node, state "Not credible for this node" with a brief justification.

Stop when all credible deviations for this node are examined. Do NOT target or assume any fixed number of deviations.

For each credible deviation:

- Identify realistic causes.
- Identify realistic consequences.
- Identify safeguards ONLY if they are visible in the P&ID or explicitly stated.
- Propose required actions to address hazards, operability issues, or uncertainty.

CRITICAL INSTRUCTIONS:

- Generate ONE complete HAZOP worksheet for THIS RUN ONLY.
- Do NOT merge with previous runs.
- Do not invent equipment, tags, alarms, or safeguards not supported by the node description/image/evidence. If you are unsure, write "Not evident

CRediT authorship contribution statement

Ehab Elhosary: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Osama Moselhi:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of Competing Interest

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

Acknowledgment

The authors extend their appreciation to Mitacs and Hatch Ltd. for funding this research work through project no. IT37249.

from node/evidence".

- Each row = one deviation scenario.
- Each deviation may have multiple causes, consequences, safeguards, and recommendations.
- If evidence is insufficient, say "INSUFFICIENT_EVIDENCE" for that field.
- Use BSEE only to support realistic accident sequences and safeguards, not to force irrelevant items.
- If information is missing or unclear, write "Uncertain" and convert it into an action.

B.2. The real HAZOP worksheet of the reference node (Node #1) used in this study.

No.	GW	Deviation	Possible Causes	Possible Consequences	Existing Safeguards	Recommendations
1	More	More water level in V-106	P-102 Shutdown	Poor desalinization in ME-101	LV-102 will close, Water will overflow to drain, High Level Alarm LT-102	Install a Backup P-102
2	More	More water level in V-106	Pipeline Blockage to P-102	Poor desalinization in ME-101	LV-102 will close, Water will overflow to drain, High Level Alarm LT-102	None
3	More	More water level in V-106	LV-102 Failure	Water waste	Water will overflow to drain, High Level Alarm LT-102	None
4	More	More water level in V-106	LT-102 Failure	Water waste	Water will overflow to drain	Install a LSH alarm in V-106
5	Less	Less Water Level in V-106	LV-102 Failure	P-102 pumping air. Possible explosive environment in ME-101, Poor desalinization in ME-101	LSLL-119/126 will shutdown the transformer, Low Oil Level Alarm, Low Level Alarm LT-102, LV-102 by-pass line, Install a LSLL in V-106 to shutdown P-102	Install a LSLL in V-106 to shutdown P-102
6	Less	Less Water Level in V-106	LT-102 Failure	P-102 pumping air. Possible explosive environment in ME-101, Poor desalinization in ME-101	LSLL-119/126 will shutdown the transformer, Low Oil Level Alarm	Install a LSLL in V-106 to shutdown P-102, Schedule a regular LT-202 test.
7	Less	Less Water Level in V-106	No potable water supply to the LV-102	P-102 pumping air. Possible explosive environment in ME-101, Poor desalinization in ME-101	LSLL-119/126 will shutdown the transformer, Low Oil Level Alarm, Low Level Alarm LT-102	None
8	Less	Less Water Level in V-106	Blockage in the pipeline to V-106	P-102 pumping air. Possible explosive environment in ME-101, Poor desalinization in ME-101	LSLL-119/126 will shutdown the transformer, Low Oil Level Alarm, Low Level Alarm LT-102	None
9	More	More Pressure in the discharge of P-102	Blockage in the line to crude oil pipeline.	Damage to P-102, Pipeline rupture	PSV-107 will open	None
10	More	More Pressure in the discharge of P-102	Increased Crude oil pressure	Damage to P-102, Pipeline rupture	PSV-107 will open, PSV-103/104 will open, High Pressure Alarm PT-101	None
11	More	More flow from LV-102 to V-106	LV-102 Failure	Water Waste	High Level Alarm LT-102, Water overflow to drain	None
12	More	More flow from LV-102 to V-106	LT-102 Failure	Water Waste	Water overflow to drain	None
13	Less	Less flow from LV-102 to V-106	LV-102 Failure	P-102 pumping air. Possible explosive environment in ME-101, Poor desalinization in ME-101	Low Level Alarm LT-102, LSLL-119/126 will shutdown the transformer, Low Oil Level Alarm, LV-102 by-pass line	Install a LSLL in V-106 to shutdown P-102
14	Less	Less flow from LV-102 to V-106	LT-102 Failure	P-102 pumping air. Possible explosive environment in ME-101, Poor desalinization in ME-101	LSLL-119/126 will shutdown the transformer, Low Oil Level Alarm	Install a LSLL in V-106 to shutdown P-102
15	Reverse	Reverse Flow from P-102 to V-106	Pump shutdown	Sour crude oil to V-106, Oil spill in the unit, H ₂ S release, Fire hazard, Explosion hazard	Check valve at P-102 discharge, Check valve at injection point	None
16	No	No flow from P-102	P-102 shutdown	Poor desalinization in ME-101	Check valve at P-102 discharge, Check valve at injection point	Install a backup pump
17	No	No containment in the pipelines after P-102	Pipeline failure due to corrosion	Loss of pressure in the water pipelines, Crude oil spill, H ₂ S Release, Explosion hazard, Fire Hazard	Check valve at injection point, H ₂ S detectors in the area.	Perform regular inspections on the pipelines

Appendix C

Sample of test sentences used to evaluate HAZOP-Adapted BERT versus baseline BERT in predicting masked tokens.

No.	Sentence_with_[MASK]	Real word	Model	Rank	Predicted token	Probability
S1	High pressure deviation may damage process [MASK].	equipment	Adapted BERT	1	equipment	0.435
S1	High pressure deviation may damage process [MASK].	equipment	bert-base-uncased	6	equipment	0.033
S2	Overpressure may result in equipment [MASK] and hydrocarbon release.	failure	Adapted BERT	1	failure	0.660
S2	Overpressure may result in equipment [MASK] and hydrocarbon release.	failure	bert-base-uncased	1	failure	0.411

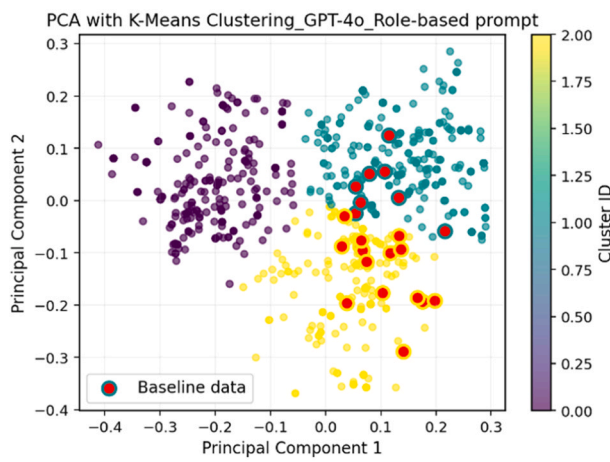
(continued on next page)

(continued)

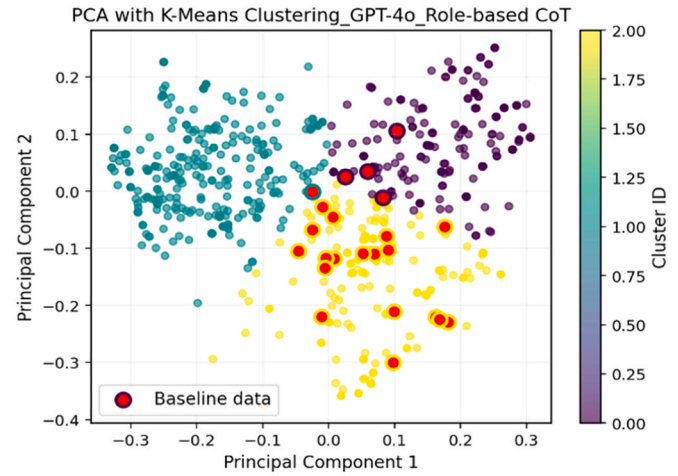
No.	Sentence_with_[MASK]	Real word	Model	Rank	Predicted token	Probability
S3	A high-level alarm should [MASK] the emergency shutdown.	initiate	HAZOP-Adapted BERT	1	initiate	0.436
S3	A high-level alarm should [MASK] the emergency shutdown.	initiate	bert-base-uncased	5	initiate	0.056
S4	The operator must follow the lockout/tagout [MASK].	procedure	HAZOP-Adapted BERT	1	procedure	0.534
S4	The operator must follow the lockout/tagout [MASK].	procedure	bert-base-uncased	1	procedure	0.106
S5	The rupture disk provides secondary pressure [MASK].	protection	HAZOP-Adapted BERT	2	protection	0.206
S5	The rupture disk provides secondary pressure [MASK].	protection	bert-base-uncased	2	protection	0.067
S6	A high level alarm will [MASK] the emergency shutdown system.	activate	HAZOP-Adapted BERT	1	activate	0.310
S6	A high level alarm will [MASK] the emergency shutdown system.	activate	bert-base-uncased	1	activate	0.262
S7	Low liquid level can cause pump [MASK] and overheating.	damage	HAZOP-Adapted BERT	3	damage	0.045
S7	Low liquid level can cause pump [MASK] and overheating.	damage	bert-base-uncased	4	damage	0.016
S8	The flow transmitter monitors the process [MASK].	flow	HAZOP-Adapted BERT	1	flow	0.431
S8	The flow transmitter monitors the process [MASK].	flow	bert-base-uncased	1	flow	0.036
S9	The emergency shutdown system isolates the affected [MASK].	equipment	HAZOP-Adapted BERT	1	equipment	0.150
S9	The emergency shutdown system isolates the affected [MASK].	equipment	bert-base-uncased		—	
S10	The interlock prevents unsafe process [MASK].	conditions	HAZOP-Adapted BERT	3	conditions	0.099
S10	The interlock prevents unsafe process [MASK].	conditions	bert-base-uncased		—	0.435

Appendix D

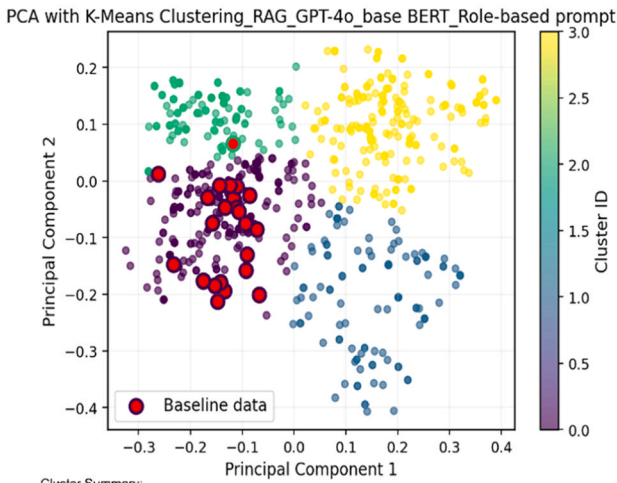
Results of PCA with K-Means clustering, along with the optimal number of clusters determined using the elbow method, for Node #2 and Node #3.



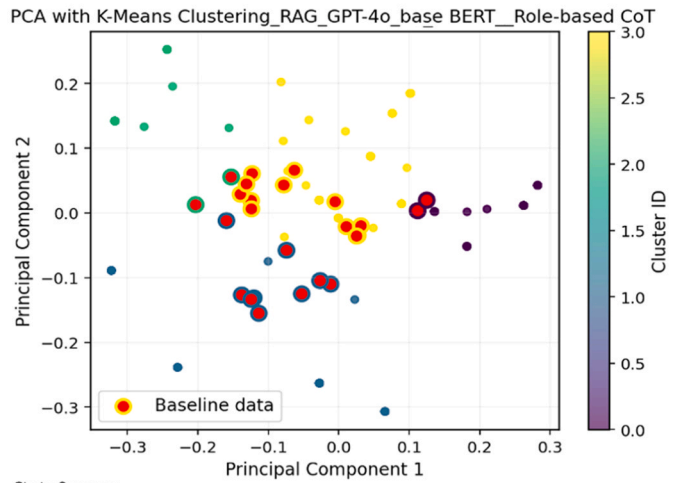
Cluster Summary:
 Cluster 0: 236 elements (Real=0, Gen=236)
 Cluster 1: 289 elements (Real=8, Gen=281)
 Cluster 2: 195 elements (Real=16, Gen=179)



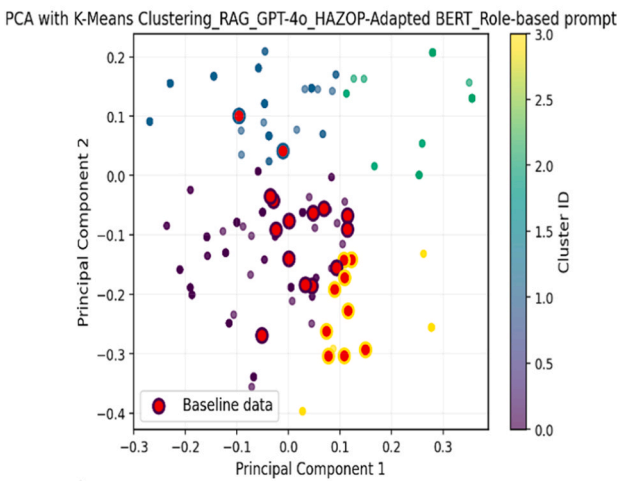
Cluster Summary:
 Cluster 0: 229 elements (Real=4, Gen=225)
 Cluster 1: 302 elements (Real=1, Gen=301)
 Cluster 2: 204 elements (Real=19, Gen=185)



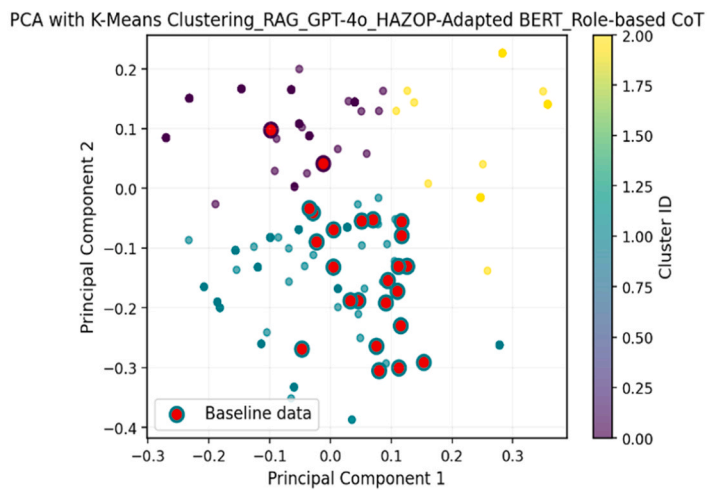
Cluster Summary:
 Cluster 0: 253 elements (Real=23, Gen=230)
 Cluster 1: 113 elements (Real=0, Gen=113)
 Cluster 2: 133 elements (Real=1, Gen=132)
 Cluster 3: 200 elements (Real=0, Gen=200)



Cluster Summary:
 Cluster 0: 144 elements (Real=2, Gen=142)
 Cluster 1: 149 elements (Real=9, Gen=140)
 Cluster 2: 76 elements (Real=2, Gen=74)
 Cluster 3: 236 elements (Real=12, Gen=224)



Cluster Summary:
 Cluster 0: 262 elements (Real=13, Gen=249)
 Cluster 1: 243 elements (Real=2, Gen=241)
 Cluster 2: 105 elements (Real=0, Gen=105)
 Cluster 3: 47 elements (Real=9, Gen=38)



Cluster Summary:
 Cluster 0: 230 elements (Real=2, Gen=228)
 Cluster 1: 275 elements (Real=22, Gen=253)
 Cluster 2: 97 elements (Real=0, Gen=97)

Fig. 16. PCA with K-Means clustering for Node #2

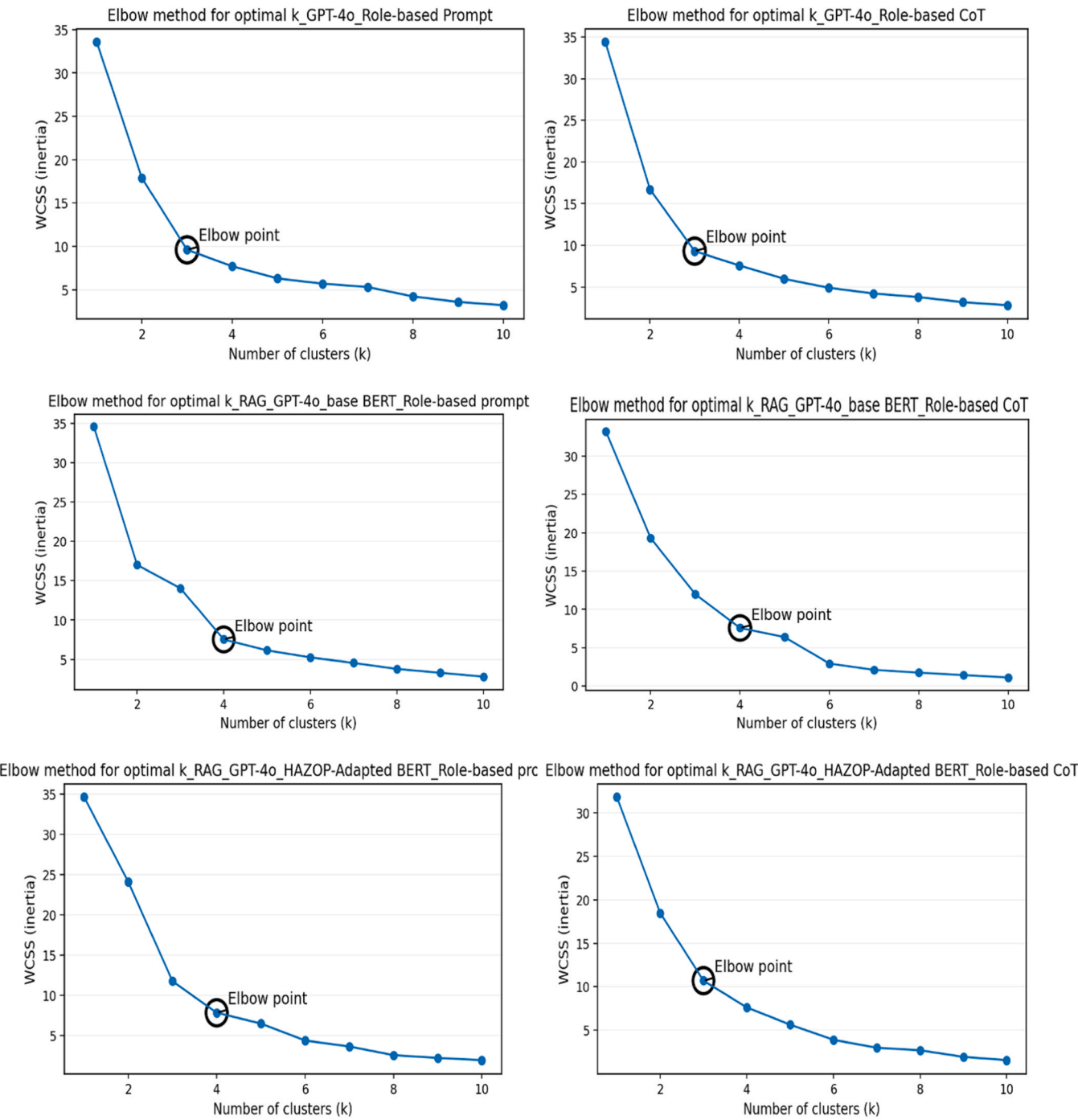


Fig. 17. Optimal number of clusters for Node #2 determined by Elbow method

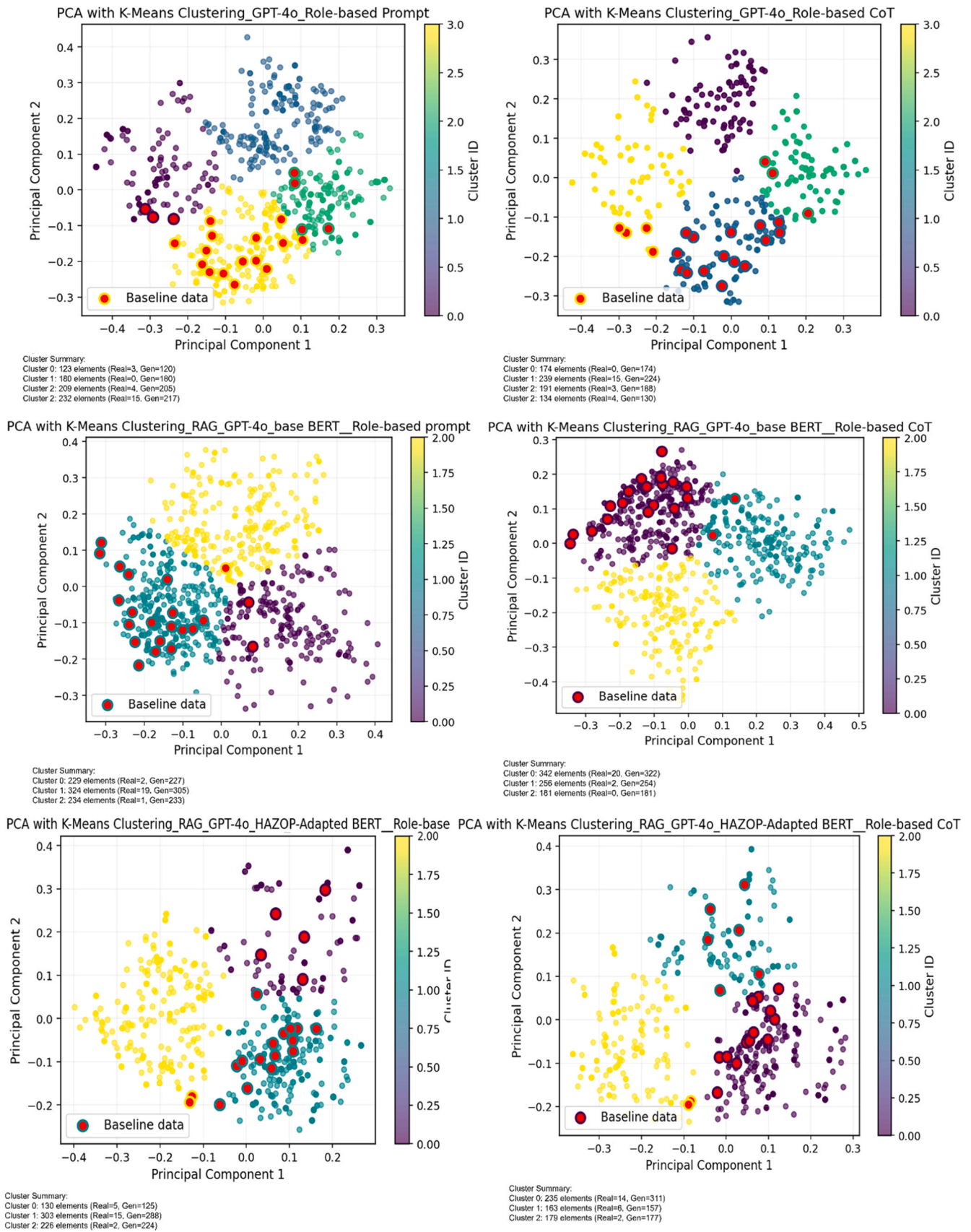


Fig. 18. PCA with K-Means clustering for Node #3

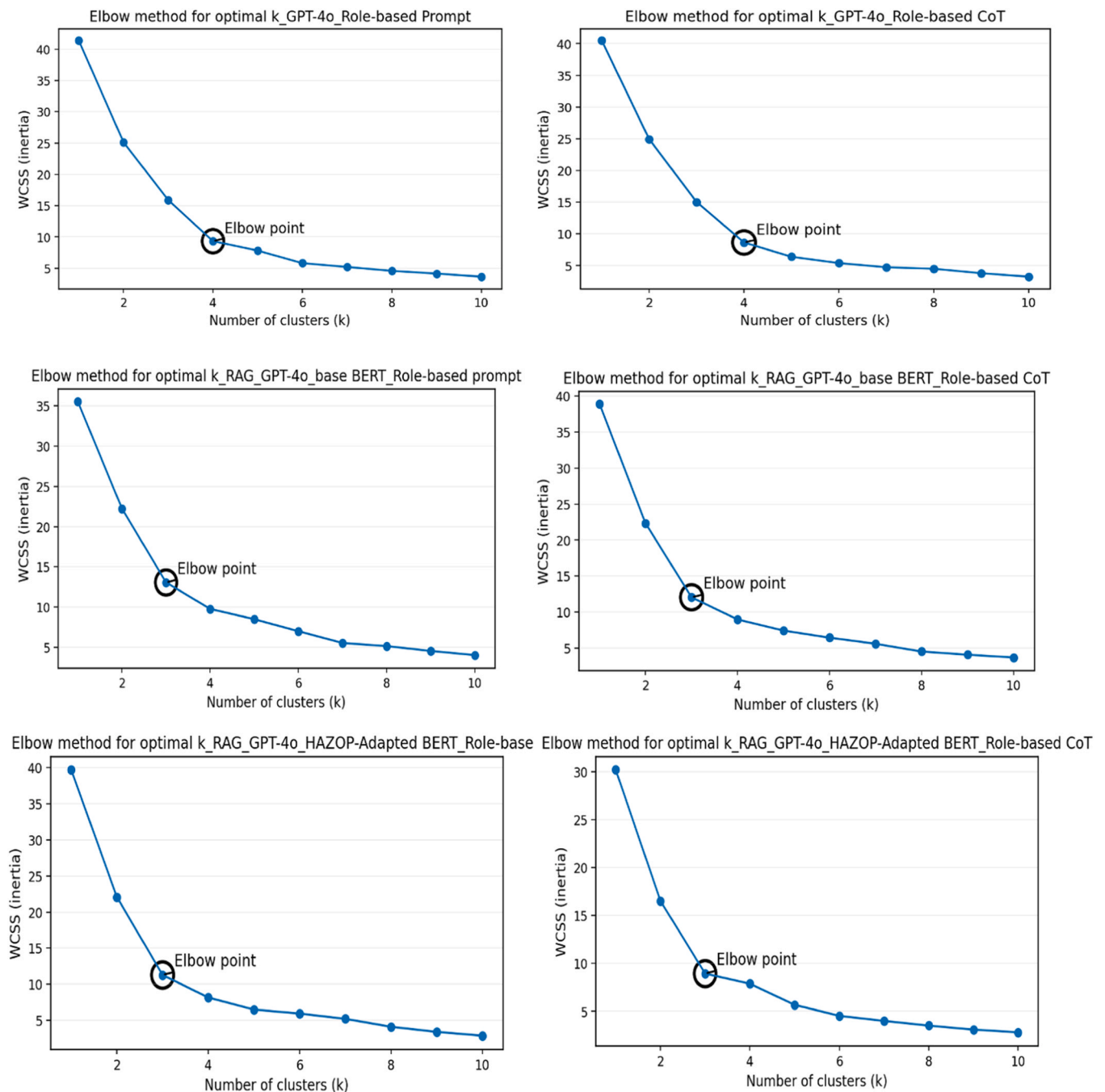


Fig. 19. Optimal number of clusters for Node #3 determined by Elbow method

Data availability

Any data, models, or code generated or used during this study can be available by the corresponding author upon reasonable request.

References

- Almousa, A.A., Hoseyni, S.M., Cordiner, J., 2025. Incorporating resilience into HAZOP to enhance process safety in industrial facilities. *J. Loss Prev. Process Ind.* 96, 105630. <https://doi.org/10.1016/J.JLP.2025.105630>.
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P., 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *Assoc. Comput. Linguist.* 1, 675–718. <https://doi.org/10.18653/V1/2023.IJCNLP-MAIN.45>.
- Belaroussi, R., 2025. Subjective assessment of a built environment by ChatGpt, gemini and grok: comparison with architecture, engineering and construction expert perception. *Big Data Cogn. Comput.* 9 (4), 100. <https://doi.org/10.3390/bdcc9040100>.
- Bholowalia, P., Kumar, A., 2014. EBK-means: a clustering technique based on elbow method and K-Means in WSN. *Int. J. Comput. Appl.* 105 (9), 17–24. <https://doi.org/10.5120/18405-9674>.
- BSEE, 2025. Offshore Incid. Investig. (<https://www.bsee.gov/what-we-do/incident-investigations/offshore-incident-investigations>).
- Bui, N., Nguyen, G., Nguyen, N., Vo, B., Vo, L., Huynh, T., Tang, A., Tran, V.N., Huynh, T., Nguyen, H.Q., Dinh, M., 2025. Fine-tuning large language models for improved health communication in low-resource languages. *Comput. Methods Prog. Biomed.* 263, 108655. <https://doi.org/10.1016/J.CMPB.2025.108655>.

- CCPS, 1998. Guidelines for Design Solutions for Process Equipment Failures, 1st ed.). American Institute of Chemical Engineers. <https://doi.org/10.1002/9780470935286>.
- CER, 2025. Rep. Compliance Enforc. (<https://www.cer-rec.gc.ca/en/safety-environment/industry-performance/reports-compliance-enforcement/index.html>).
- Chen, B., Zhang, Z., Langrené, N., Zhu, S., 2025. Unleashing the potential of prompt engineering for large language models. *Patterns* 6 (6), 101260. <https://doi.org/10.1016/j.patter.2025.101260>.
- CSB. (2025). *Investigations*. (https://www.csb.gov/investigations/?F_InvestigationId=3587&Type=undefined).
- Danish, A.A.K., Chatterjee, S., 2026. Bridging the safety-specific language model gap: Domain-adaptive pretraining of transformer-based models across several industrial sectors for occupational safety applications. *Expert Syst. Appl.* 299, 130068. <https://doi.org/10.1016/j.eswa.2025.130068>.
- Daramola, O., Stålhane, T., Omoronyia, I., Sindre, G., 2013. Using ontologies and machine learning for hazard identification and safety analysis. *Managing Requirements Knowledge*. Springer, pp. 117–141.
- Devlin, J., Chang, M.-W., Lee, K., Google, K.T., Language, A.I., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proc. 2019 Conf. North* 4171–4186. <https://doi.org/10.18653/v1/N19-1423>.
- Ekrampiooya, A., Boroushaki, M., Rashtchian, D., 2023. Application of natural language processing and machine learning in prediction of deviations in the HAZOP study worksheet: A comparison of classifiers. *Process Saf. Environ. Prot.* 176, 65–73. <https://doi.org/10.1016/j.psep.2023.06.004>.
- Ekrampiooya, A., Boroushaki, M., Rashtchian, D., 2024. Predicting possible recommendations related to causes and consequences in the HAZOP study worksheet using natural language processing and machine learning: BERT, clustering, and classification. *J. Loss Prev. Process Ind.* 89. <https://doi.org/10.1016/j.jlp.2024.105310>.
- Elhosary, E., Moselhi, O., 2024. Automation for HAZOP study: A state-of-the-art review and future research directions. *J. Inf. Technol. Constr.* 29, 750–777. <https://doi.org/10.36680/j.itcon.2024.033>.
- Elhosary, E., Moselhi, O., 2025a. Evaluating Natural Language Processing Algorithms for Improved Hazard and Operability Analysis. *Geod. AI* 4, 100026. <https://doi.org/10.1016/j.geoi.2025.100026>.
- Elhosary, E., Moselhi, O., 2025c. Utilization of Artificial Intelligence in HAZOP Studies and Reports. *42nd Int. Symp. Autom. Robot. Constr.* 617–624. <https://doi.org/10.22260/ISARC2025/0081>.
- Elhosary, E., Moselhi, O., 2025b. Intelligent countermeasures analysis in oil and gas projects utilizing topic modeling. *J. Loss Prev. Process Ind.* 98, 105751. <https://doi.org/10.1016/j.jlp.2025.105751>.
- Elhosary, E., Moselhi, O., Bucur, C., 2024, November 5. Evaluation of AI-assisted HAZOP Software Tools. *Hazards 34 Process Saf. Conf. Symp. Ser.* 171. (<https://www.icheme.org/knowledge-networks/communities/special-interest-groups/safety-and-loss-prevention/resources/hazards-conference-archive/hazards-34/>).
- Feng, X., Dai, Y., Ji, X., Zhou, L., Dang, Y., 2021. Application of natural language processing in HAZOP reports. *Process Saf. Environ. Prot.* 155, 41–48. <https://doi.org/10.1016/j.psep.2021.09.001>.
- Gao, P., Li, W., 2022. Integration of HAZOP and Bayesian network in city gas explosion emergency response processes. *Emerg. Manag. Sci. Technol.* 2 (1), 1–9. <https://doi.org/10.48130/emst-2022-0019>.
- Hosseini, S.H., Pourzangbar, A., 2026. How well do DeepSeek, ChatGPT, and Gemini respond to water science questions? *Environ. Model. & Softw.* 196, 106772. <https://doi.org/10.1016/j.envsoft.2025.106772>.
- IEC. (2016). Hazard and Operability Studies (HAZOP Studies)—Application Guide (IEC 61882:2016); (Vol. 2).
- Joubert, F., Steyn, E., Pretorius, L., 2021. Using the HAZOP Method to Conduct a Risk Assessment on the Dismantling of Large Industrial Machines and Associated Structures: Case Study. *J. Constr. Eng. Manag.* 147 (1). [https://doi.org/10.1061/\(asce\)co.1943-7862.0001942](https://doi.org/10.1061/(asce)co.1943-7862.0001942).
- Kojima, T., Shane Gu, S., Reid Google Research, M., Matsuo, Y., Iwasawa, Y., 2022. Large Language Models are Zero-Shot Reasoners. *Adv. Neural Inf. Process. Syst.* 35, 22199–22213.
- Kranz, P., Schirmer, F., Kaupp, T., Daun, M., 2024. Generative AI Copilot to Support Safety Analyses of Human-Robot Collaborations: Hazard Operability Analysis and GPT-4. *IEEE Softw.* 41 (6), 65–72. <https://doi.org/10.1109/MS.2024.3414445>.
- Lahitani, A.R., Permasari, A.E., Setiawan, N.A., 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. *CITSM 2016 Proc.* 2016 4th Int. Conf. Cyber IT Serv. Manag.. <https://doi.org/10.1109/CITSM.2016.7577578>.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *8th Int. Conf. Learn. Represent.* 2020. (<http://arxiv.org/abs/1909.11942>) (ICLR).
- Lee, J., Park, S., Oh, S., Ma, B., 2026. Can large language models automate the HAZOP process without human intervention? *Saf. Sci.* 194, 107039. <https://doi.org/10.1016/j.ssci.2025.107039>.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* 33, 9459–9474. (<https://github.com/huggingface/transformers/blob/master/>).
- Li, Z., Zhao, J., 2025. HAZOPT: A HAZOP analysis completeness tool based on knowledge graph reasoning. *Process Saf. Environ. Prot.* 197, 107025. <https://doi.org/10.1016/j.psep.2025.107025>.
- Liu, G., Boyd, M., Yu, M., Halim, S.Z., Quddus, N., 2021. Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques. *Process Saf. Environ. Prot.* 152, 37–46. <https://doi.org/10.1016/j.psep.2021.05.036>.
- Liu, H., Yin, H., Luo, Z., Wang, X., 2025. Integrating chemistry knowledge in large language models via prompt engineering. *Synth. Syst. Biotechnol.* 10 (1), 23–38. <https://doi.org/10.1016/j.synbio.2024.07.004>.
- Marhavilas, P.K., Filippidis, M., Koulinas, G.K., Koulouriotis, D.E., 2019. The integration of HAZOP study with risk-matrix and the analytical-hierarchy process for identifying critical control-points and prioritizing risks in industry – A case study. *J. Loss Prev. Process Ind.* 62, 103981. <https://doi.org/10.1016/j.jlp.2019.103981>.
- Marhavilas, P.K., Filippidis, M., Koulinas, G.K., Koulouriotis, D.E., 2020. An expanded HAZOP-study with fuzzy-AHP (XPA-HAZOP technique): Application in a sour crude-oil processing plant. *Saf. Sci.* 124, 104590. <https://doi.org/10.1016/j.ssci.2019.104590>.
- Marhavilas, P.K., Filippidis, M., Koulinas, G.K., Koulouriotis, D.E., 2021. Safety Considerations by Synergy of HAZOP/DMRA with Safety Color Maps—Applications on: A Crude-Oil Processing Industry/a Gas Transportation System. *Processes* 9 (8), 1299. <https://doi.org/10.3390/pr9081299>.
- Mokhtarname, R., Urbas, L., Safavi, A.A., Salimi, F., Zerafat, M.M., Harasi, N., 2024. An artificial neural network approach to enrich HAZOP analysis of complex processes. *J. Loss Prev. Process Ind.* 91, 105382. <https://doi.org/10.1016/j.jlp.2024.105382>.
- Mukharrom, D.A., Susilo, H., Putra, Z.A., 2025. On the Use of Artificial Intelligent with Large Language Model as a Tool for Evaluating Quality of Hazard and Operability Study. *9th Int. Conf. Instrum. Control Autom. (ICA)* 2025, 337–342. <https://doi.org/10.1109/ICA65945.2025.11252369>.
- Park, S.W.J., Pasman, H.J., 2018. Is HAZOP a Reliable Tool? What Improvements are Possible? *J. Korean Inst. Gas.* 22 (2), 1–20. <https://doi.org/10.7842/KIGAS.2018.22.2.1>.
- Peepkerom, M., Kouwenhoven, T., Brown, D., Jordanous, A., 2024. Is. Temp. Creat. Parameter Large Lang. Models. (<https://arxiv.org/pdf/2405.00492>).
- Peng, S., Kalliamvakou, E., Cihon, P., Demirer, M., 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. *Softw. Eng.* <https://doi.org/10.48550/arXiv.2302.06590>.
- Rossing, N.L., Lind, M., Jensen, N., Jørgensen, S.B., 2010. A functional HAZOP methodology. *Comput. Chem. Eng.* 34 (2), 244–253. <https://doi.org/10.1016/j.compchemeng.2009.06.028>.
- Salus Technical (2023). SALUS HAZOP AI. Accessed September, 2024. (<https://salus-technical.com/>).
- Single, J., Schmidt, J., Denekke, J., 2019. Computer-aided hazop studies: Knowledge representation and algorithmic hazard identification. *WIT Trans. Built Environ.* 189, 55–66. <https://doi.org/10.2495/SAFE190061>.
- Tikayat Ray, A., Bhat, A.P., White, R.T., Nguyen, V.M., Pinon Fischer, O.J., Mavris, D.N., 2023. Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS). *Aerospace* 2023 10 (9), 10. <https://doi.org/10.3390/AEROSPACE10090770>.
- TSB, 2025. Pipeline transportation safety investigations and reports. (<https://www.tsb.gc.ca/eng/rappports-reports/pipeline/index.html>).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, L., 2017. Attention Is All You Need. *Comput. Lang.*
- Wagstaff, K., Cardie, C., Rogers, S., Icm, S.S., & 2001, undefined. (2001). Constrained k-means clustering with background knowledge. *Cse.Msu.EduK Wagstaff, C Cardie, S Rogers, S Schrödlcm1, 2001cse.Msu.Edu, 577–584.* (<https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf>).
- Wang, F., Gu, W., 2022. Intelligent HAZOP analysis method based on data mining. *J. Loss Prev. Process Ind.* 80. <https://doi.org/10.1016/j.jlp.2022.104911>.
- Wang, F., Gu, W., Bai, Y., Bian, J., 2023. A method for assisting the accident consequence prediction and cause investigation in petrochemical industries based on natural language processing technology. *J. Loss Prev. Process Ind.* 83. <https://doi.org/10.1016/j.jlp.2023.105028>.
- Wang, Y., Luo, H., Fang, W., 2025. An integrated approach for automatic safety inspection in construction: Domain knowledge with multimodal large language model. *Adv. Eng. Inform.* 65, 103246. <https://doi.org/10.1016/j.aei.2025.103246>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi Quoc, E.H., Le, V., Zhou, D., 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv. Neural Inf. Process. Syst.* 35, 24824–24837.
- Yang, H., Li, C.Y., Wang, Q., 2026. Domain-augmented large language models for automated root cause classification of offshore process incidents. *J. Loss Prev. Process Ind.* 99, 105845. <https://doi.org/10.1016/J.JLP.2025.105845>.
- Zhang, J., Zhang, W., Wei, Q., Zhong, H., Miao, Y., 2026. GRAG-ProSafe QAS: Graph retrieval-augmented generation for process production safety intelligent question and answer system. *Expert Syst. Appl.* 298, 129947. <https://doi.org/10.1016/J.ESWA.2025.129947>.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019. BERTScore: Evaluating Text Generation with BERT. *8th Int. Conf. Learn. Represent. ICLR.* (<https://arxiv.org/pdf/1904.09675>), 2020.
- Zhao, J., Cui, L., Zhao, L., Qiu, T., Chen, B., 2009. Learning HAZOP expert system by case-based reasoning and ontology. *Comput. Chem. Eng.* 33 (1), 371–378. <https://doi.org/10.1016/j.compchemeng.2008.10.006>.
- Zheng, C., Chen, G., Chen, H., Xu, Q., Zhao, Y., Zhao, Y., Yang, Y., 2025. Chemical process safety domain knowledge graph-enhanced LLM for efficient emergency response decision support. *Can. J. Chem. Eng.* 103 (10), 4814–4843. <https://doi.org/10.1002/CJCE.25700;WGROUP:STRING:PUBLICATION>.