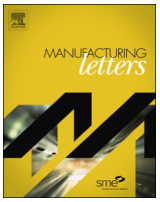
Available online at www.sciencedirect.com

Manufacturing Letters

00 (2026) 000–000



54th SME North American Manufacturing Research Conference (NAMRC 54, 2026)

A Multimodal Manufacturing Safety Chatbot: Knowledge Base Design, Benchmark Development, and Evaluation of Five RAG Approaches

Ryan Singh[a], Austin Hamilton[b], Amanda White[a], Michael Wise[c,d], Ibrahim Yousif[c], Arthur Carvalho[a], Jay Shan[a], Reza Abrishambaf[c], Mohammad Mayyas[c], Lora A. Cavuoto[d], Fadel M. Megahed[a,*]

[a]*Farmer School of Business, Miami University, 800 E. High Street, Oxford, OH 45056, USA*
[b]*Department of Computer Science and Software Engineering, Miami University, 105 Tallawanda Road, Oxford, OH 45056, USA*
[c]*Department of Engineering Technology, Miami University, 1601 University Blvd., Hamilton, OH 45011, USA*
[d]*Department of Mechanical and Manufacturing Engineering, Miami University, 650 E. High Street, Oxford, OH 45056, USA*
[e]*Department of Industrial and Systems Engineering, University at Buffalo, 407 Bell Hall, Buffalo, NY, 14260*

Abstract

Ensuring worker safety in modern manufacturing environments remains a critical challenge as Industry 5.0 reorients smart factories around human–AI collaboration. Following a design science research methodology, we start by identifying three essential requirements for next-generation safety training systems: 1) high accuracy, 2) low latency, and 3) affordability. Subsequently, we introduce an artifact that satisfies all such requirements, namely a multimodal chatbot powered by large language models designed to enhance worker safety and training in Industry 5.0 environments. Using modern retrieval-augmented generation techniques, we are able to ground the chatbot on curated regulatory and technical documentation. To test our solution, a domain-specific benchmark of expert-validated question–answer pairs was developed across three representative machines: a Bridgeport manual mill, a Haas TL-1 CNC lathe, and a Universal Robots UR5e collaborative robot. Using a full-factorial experimental design, we evaluated XXX RAG configurations on automated and human-assessed metrics of correctness, latency, and cost. Moreover, our results show that retrieval strategy and model configuration significantly affect system performance, with the most achieving accuracy and latency metics here, while costing $ per query on average. This final deployed chatbot supports text, image, and voice interactions, providing accessible, explainable, and real-time safety guidance. Overall, this work contributes an open-source framework, benchmark, and evaluation methodology to advance AI-driven safety training and human-centric manufacturing.

Keywords: Artificial intelligence; BM25 ranking; Graph-based retrieval; Human-machine interaction; Keyword search; Large language models; Semantic search

1. Introduction

Manufacturing remains the backbone of the United States economy, generating over $2.9 trillion in annual GDP in 2024, or roughly 10% of total U.S. economic output, according to the Bureau of Economic Analysis [1, Page 22]. The sector employs over 12.7 million workers [2] and acts as a critical engine for innovation, productivity, and national competitiveness, driving around 54% of the research and development done by American businesses with 10 or more domestic employees in 2021 [3]. These figures underscore not only the economic vitality of the sector but also the strategic importance of ensuring the safety and productivity of its workforce.

Within the past 15 years, manufacturing workplaces have experienced a shift, with the advent of Industry 4.0 and the integration of the internet of things (IoT) and cyber-physical systems [4]. Smart manufacturing platforms with the ability to integrate real-time data have enabled monitoring and adaptation of manufacturing processes and systems. However, the empha-

* Corresponding author. Tel.: +1-513-529-4185.
E-mail address: fmegahed@miamioh.edu (Fadel M. Megahed).

sis of these systems was often at the equipment level. In the last five years, a new industrial revolution (Industry 5.0) has refocused the manufacturing environment around the humans in the system [4, 5]. This has been centered on human-robot collaboration and the integration of artificial intelligence (AI).

Despite its economic importance, the manufacturing industry continues to face persistent safety challenges. According to the U.S. Bureau of Labor Statistics [6], the manufacturing sector reported over 320,000 workplace injuries, making it one of the most hazardous occupational domains. Common incidents include cuts, crush injuries, and caught-in/between accidents. These injuries not only endanger workers' lives but also impose substantial costs on employers and the national economy. An estimated $ 7.5 billion is spent in the manufacturing sector each year on the direct costs of medical payments and lost wages [7]. Contributing factors to workplace safety incidents include systems and worker-level factors.

Along with the traditional safety concerns, the evolving manufacturing environment has created new challenges for manufacturing operators, as they need to know about and interact with new systems. For example, variations in equipment make it difficult for operators to remember use procedures across a range of systems. When this is accompanied by the time pressure imposed by high production demands, safety incidents may be more likely to occur.

Prior studies have presented current applications of AI in manufacturing to support operators in performing their tasks, including the presentation of maintenance guidelines, process monitoring, predicting machine downtime, and worker training [8]. Modernizing worker training is a critical area for advancing worker performance and mitigating safety risk. Traditional training has often involved classroom or online lecture-based content presented upon job entry and/or at fixed intervals (e.g., annually). This approach is static and fails to reflect the complexity of modern manufacturing environments or worker experiences. In this static setup, when an operator needs information about a piece of equipment, they may not be able to easily access information in a user manual or online. There is thus a pressing need for adaptive, context-aware training systems capable of providing personalized, accurate, and readily accessible safety information in real time.

Following a design science research (DSR) methodology [9], our work begins by identifying three essential requirements that any next-generation safety training solution must satisfy: high accuracy, low cost, and low latency. To meet these requirements, we propose an open-source Retrieval-Augmented Generation (RAG)–based chatbot, grounded in regulatory documents, e.g., Occupational Safety and Health Administration (OSHA) standards and equipment manuals from relevant industrial machinery. A recent study in construction safety showed that chatbot-based training led to better hazard awareness, particularly for those participants who had less onsite experience [10]. One identified benefit was the customizability of the training materials based on the specific worker needs. By combining large language models (LLMs) with curated, domain-specific sources, our RAG system delivers precise, contextually relevant safety guidance while maintaining affordability and responsiveness suitable for real-world manufacturing settings.

A secondary contribution of this work is the development of a domain-specific benchmark to systematically evaluate the chatbot's safety reasoning capabilities. Focusing on three different equipment, namely Bridgeport Manual Mill, TL-1 CNC Machine, and the Universal Collaborative Robot (Cobot), we constructed datasets of relevant question–answer pairs, vetted by external researchers specializing in industrial safety and human factors and/or industry professionals. This benchmark allows for reproducible, quantitative assessment of the chatbot's accuracy and latency, establishing a foundation for future research in safety-oriented conversational AI systems.

In summary, our work contributes to the growing literature on AI for occupational safety and training by demonstrating how RAG systems can transform static safety instruction into dynamic, interactive, and data-driven learning experiences. By integrating regulatory grounding, multimodal reasoning, and open benchmarking, this research illustrates a viable pathway for using AI to enhance worker safety, compliance, and preparedness in the era of Industry 5.0.

Besides this introductory section, the rest of this paper is organized as follows. Section 2 reviews the relevant literature on leveraging modern AI technologies in manufacturing settings. Section 3 describes our DSR methodology, including the three design requirements guiding solution (artifact) development. Section 4 details the artifact design and development process. Section 5 demonstrates the proposed solution. Section 6 presents the artifact evaluation procedures, including the automated and human assessments used to compare different RAG configurations. Section 7 discusses the implications of our findings for Industry 5.0 safety training and human–AI collaboration. Finally, Section 8 concludes the paper and outlines avenues for future research.

2. Literature Review

Existing research in manufacturing has explored a range of AI-driven approaches to enhance worker safety, hazard awareness, and on-site decision support. In particular, three interrelated themes emerge in our literature review: 1) the use of AI, including LLMs and RAG techniques, for training and decision making, 2) multimodal interfaces (e.g., vision-and-language) for richer human–machine interaction, and 3) digital twin environments for simulation and operator support. This section reviews representative works in each area.

We start by noting that conversational AI frameworks have been successfully applied to real-time industrial data to assist operators in understanding machine status. For example, Jeon et al. [11] note that a lack of digital literacy among shop-floor workers often hinders the use of conventional monitoring interfaces. Their system, ChatCNC, mitigates this by allowing operators to query machine conditions, such as whether the machine is overheating, in plain language, instead of navigating complex software and user interfaces.

The above study illustrates a trend toward AI-mediated instruction and monitoring in manufacturing, aimed at proactively reducing human error and improving safety outcomes on the shop floor. More specifically, natural language techniques, such as LLMs, are attractive for their general reasoning ability and conversational interface, which align well with Industry 5.0's emphasis on human-centric smart manufacturing Along these lines, Wang et al. [12] proposed a collaborative robot that was endowed with vision–language capabilities that could visually map its environment and understand spoken or written instructions via an LLM-based parser. In another study, Lou et al. (2025) [13] proposed an LLM-enabled cognitive agent for manufacturing that plans and makes decisions for tasks like assembly and disassembly. By leveraging the LLM's expansive knowledge and reasoning capabilities, their agent exhibited a form of self-awareness about the manufacturing context, outperforming standard automation in achieving task goals. These examples show LLMs adapted to the manufacturing domain can serve as powerful conversational intermediaries, translating between human intentions and complex technical processes.

A key challenge in applying LLMs to manufacturing is the gap between the models' general training and the specialized, real-time knowledge needed on the factory floor. Retrieval-Augmented Generation (RAG) has emerged as an effective technique to bridge this gap. In a RAG architecture, the LLM is coupled with a retrieval mechanism that fetches relevant domain information (e.g., documents, sensor data, or knowledge base entries), which the LLM then uses to ground its responses. Close to our work, Fan et al. [14] introducing MaViLa, a vision–language model that can provide guidance in manufacturing settings by interpreting live visuals as well as leveraging domain knowledge via RAG. While MaViLa centers on technologies for broad manufacturing tasks such as process understanding and skill acquisition, our work targets manufacturing safety as its primary domain, leveraging RAG to ground chatbot responses in a curated safety knowledge base.

Another significant thread in manufacturing is the development of multimodal interfaces that combine vision, language, and other data modalities to more fully mirror how humans perceive and communicate. Many shop-floor scenarios are inherently visual (e.g., identifying a hazard, checking if protective equipment is worn, etc.). As such, an AI assistant that can both see and talk has clear advantages over text-only systems. The industrial metaverse vision put forth by Li et al. [15] paints a picture of highly multimodal manufacturing environments where workers and AI interact through virtual/augmented reality interfaces. In such settings, digital avatars, spatial audio, and mixed-reality visuals would all complement traditional textual data, creating a rich interface for addressing safety and production issues. The convergence of vision and language in these works underscores that a truly effective manufacturing chatbot should be multimodal, capable of both processing images or sensor readings and generating language to fully support safety monitoring and operator assistance. As such, multimodality is at the heart of the solution we develop in our work.

Finally, the use of digital twins and virtual environments has become prominent as a means to integrate the aforementioned technologies (LLMs, RAG, and multimodal interfaces) in a realistic, safety-oriented setting. A digital twin is a virtual replica of a physical system, continuously updated with real-world data, which can be used for simulations, monitoring, and what-if analyses. In manufacturing, digital twins have been adopted to facilitate proactive safety management and training without disrupting actual operations. Gautam et al. [16] present a compelling example by linking an LLM-based agent system with a factory's digital twin. In their implementation, multiple specialized LLM agents (for machine expertise, data visualization, fault diagnosis, etc.) ingest streaming IIoT data from machines on the shop floor and reflect the factory's state inside a virtual model. The digital twin not only mirrors machine readings and production metrics in real time, but also provides an interactive 3D interface where an operator can query an avatar representing the AI about the system's status or troubleshooting steps. By consulting the twin, operators can gain insights without physically inspecting hazardous equipment, and they receive immediate, expert guidance.

Collectively, the above studies demonstrate the building blocks of a modern manufacturing safety assistant. They show that LLMs can interface naturally with human operators, that RAG can inject crucial real-time or domain-specific knowledge into responses, that multimodal understanding (vision + language) is key for interpreting factory situations. This observation motivates us to combine all these elements specifically for manufacturing safety training and hazard prevention. In particular, we design a retrieval-augmented multimodal chatbot that unifies these strands, i.e., an AI agent that can draw on an extensive safety knowledge base, perceive the manufacturing environment, and engage in dialog in order to deliver effective safety guidance and risk analysis in real time.

3. Research Methodology

We address the challenge of modernizing hazard-recognition training in manufacturing by adopting a design science research methodology (DSR) [9]. This methodological approach is well-suited for the development and evaluation of information technology solutions (artifacts) that aim to solve complex, real-world problems. Following DSR ensures that our work systematically integrates scientific rigor with practical relevance, producing outcomes that are both theoretically grounded and operationally viable within industrial settings.

In particular, our research process follows the six-stage model outlined by Peffers et al. [9], which we elaborate on throughout the paper: 1) problem identification and motivation, 2) definition of solution objectives, 3) artifact design and development, 4) demonstration, 5) evaluation, and 6) communication. We have already identified and motivated the problem in the preceding section, and the last stage of the DSR methodology, i.e., communication, is represented by this paper. That said, we now turn to the second stage, namely defining the solution objectives, which are derived "*from the problem definition and knowledge of what is possible and feasible*" [9]. To operationalize this stage, we established a set of *design requirements*

that any proposed artifact must satisfy to address the underlying hazard-recognition and safety-training challenges.

Our first design requirement concerns system *accuracy*, which represents the non-negotiable foundation of any effective safety-training system. In industrial environments, inaccurate or misleading instructions can lead to improper procedures, equipment misuse, and, in the worst cases, fatal accidents. A training artifact that compromises factual precision risks undermining the very purpose of hazard education. Therefore, every element of the system, whether textual guidance, visual cues, or procedural demonstrations, must convey information that is technically correct, operationally safe, and fully aligned with regulatory and manufacturer standards. The following requirement definition captures the above discussion.

> **Design Requirement #1 (Accuracy)**: Training artifacts must deliver information that is technically correct and operationally safe.

Our second design requirement concerns system *latency*, reflecting the need for timely and responsive feedback within dynamic manufacturing environments. In industrial contexts, hazards can emerge and evolve in seconds, and delayed or sluggish responses can render safety guidance ineffective or even dangerous. A training system that fails to deliver instructions, alerts, or assessments in real time risks losing its relevance to the unfolding situation on the shop floor. Therefore, any proposed solution must minimize latency in both data processing and user feedback, ensuring that workers receive immediate, context-appropriate information as they interact with machinery or simulated environments. Low latency is essential not only for maintaining user engagement and situational awareness but also for reinforcing procedural learning under realistic temporal conditions. The following definition summarizes our second design requirement.

> **Design Requirement #2 (Latency)**: Artifacts must provide timely and responsive feedback, ensuring that safety instructions are rapidly provided under changing conditions in manufacturing environments.

Our third design requirement concerns the *cost* of deployment and operation, emphasizing the importance of economic accessibility for broad adoption across the manufacturing sector. Safety innovation cannot achieve its intended societal impact if the associated technologies are prohibitively expensive to implement, maintain, or scale. Many small and medium-sized manufacturers, which represent the majority of industrial employers, operate with limited budgets for workforce development and technology infrastructure. Therefore, training solutions must minimize recurring expenses such as software licensing, hardware requirements, and data-processing costs, without compromising instructional quality or safety integrity.

> **Design Requirement #3 (Cost)**: Artifacts must minimize deployment and operational costs to ensure affordability and, consequently, adoption across manufacturers of varying sizes.

In the following sections, we design, demonstrate, and evaluate an artifact that satisfies the above design requirements. We do so by proposing the six-phase framework in Figure 1 for building manufacturing-safety chatbots grounded in expert-curated operational and safety documents. These phases align perfectly with our DSR methodology. For example, Phases 1 and 2 are part of the third DSR stage called "artifact design and development." In particular, in Phase 1, we curate and preprocess safety materials from OSHA, the National Institute for Occupational Safety and Health (NIOSH), and Original Equipment Manufacturer (OEM) manuals to build a reliable, structured corpus. In Phase 2, we construct multiple knowledge bases that integrate keyword, semantic, and graph indices to support flexible information retrieval for question answering (Q&A). In Phase 3, we transition to the fourth design science stage, namely artifact demonstration. In particular, we demonstrate how a grounded, multimodal chatbot can be developed based on the knowledge base and indexing from the previous phase. Phases 4 to 6 are linked to the fifth stage of our DSR methodology, namely artifact evaluation. Phase 4 develops a benchmark Q&A dataset, with expert-validated questions and reference answers. We evaluate multiple retrieval-augmented generation (RAG) configurations using automated metrics in Phase 5 to identify a reasonable subset of configurations for human evaluation. In Phase 6, we employ blind human evaluation to select the "best" configuration. The evaluations in Phases 4 to 6 enable us to implement a single configuration and deploy a multimodal chatbot that integrates text, visual, and speech interactions, ensuring that its responses remain grounded in our curated source documents.

4. Artifact Design and Development

In the third stage of our DSR methodology, we designed and developed a concrete artifact. While many digital solutions could theoretically address the challenge of delivering accessible, accurate, and adaptive safety training in manufacturing, few can simultaneously meet the requirements of accuracy, low latency, and affordability identified in Section 3. For example, generic e-learning platforms or static digital manuals often fall short because they lack contextual awareness, interactivity, and real-time responsiveness. In contrast, modern AI technologies, such as a multimodal chatbot capable of processing text, images, and speech, offers a uniquely flexible and human-centered medium for safety communication. In particular, by embedding RAG techniques within a multimodal interface, such a system can ground its responses in authoritative safety documents while allowing workers to query information naturally through the mode most convenient to their environment. This convergence of adaptive retrieval, multimodal reasoning, and conversational accessibility positions a RAG-based chatbot as a suitable artifact to operationalize the design requirements and realize the goals of Industry 5.0–aligned, human-centric safety training. We next describe the first two phases of our development framework (Figure 3), which focus on data curation and the construction of RAG methods.

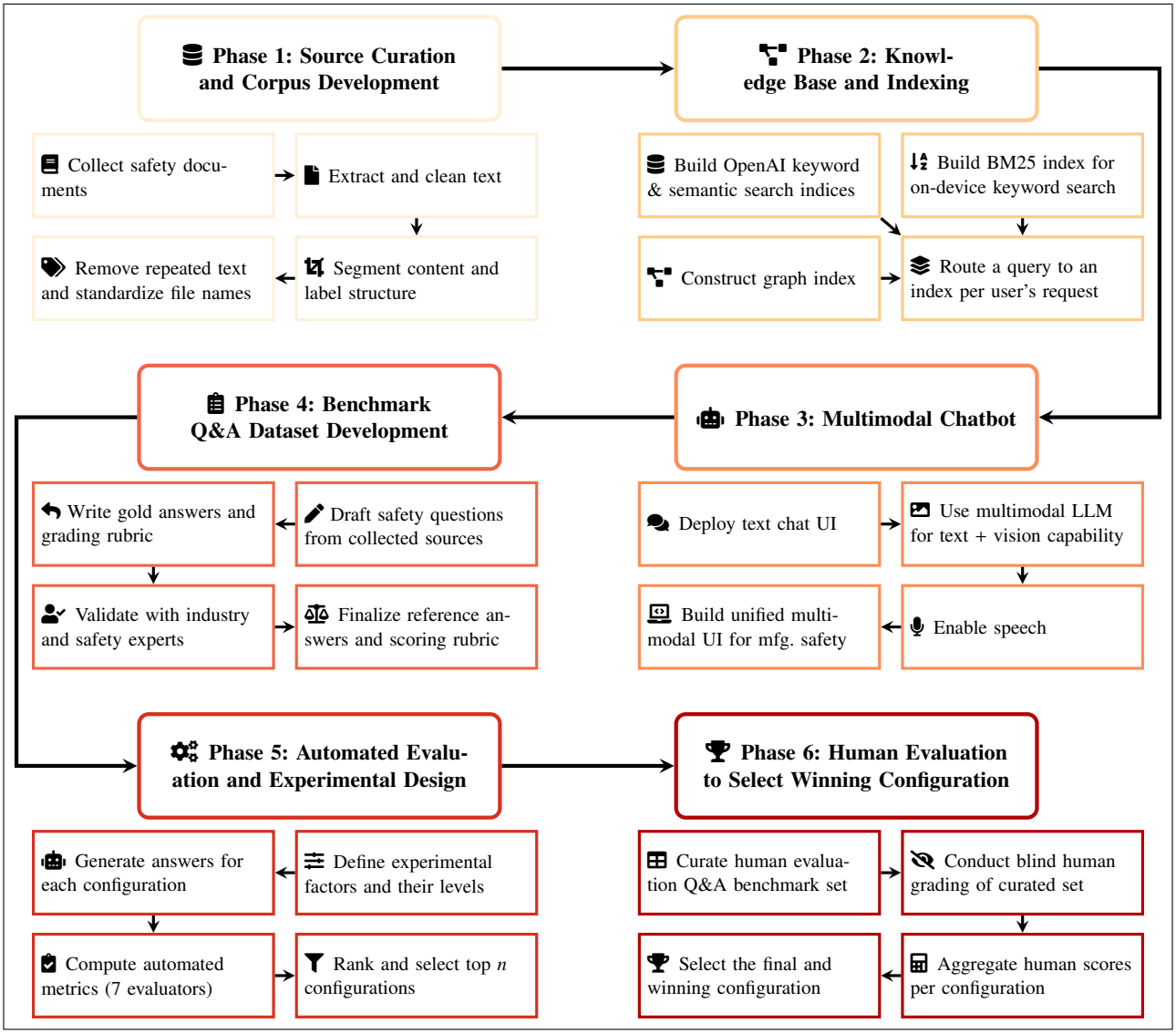


Figure 1. An overview of our proposed six-phase framework for developing, demonstrating, and evaluating a multimodal manufacturing-safety chatbot.

4.1. Phase 1: Source Curation and Corpus Development

Phase 1 marks the foundation of the chatbot development framework by focusing on the systematic curation and preparation of safety-critical source materials. This stage ensures that all subsequent modeling and retrieval processes are grounded in authoritative, high-quality data. It involves collecting and organizing diverse safety documents into a unified corpus suitable for RAG. This phase includes four steps: 1) collecting the source documents, 2) extracting and cleaning text, 3) segmenting the content into meaningful units, and 4) standardizing how each segment is saved.

4.1.1. Collect Safety Documents

Grounding LLM responses in authoritative documents reduces hallucinations and improves factual correctness when compared to generation from the LLM alone [17]. We curated four categories of documents to ensure that the chatbot's responses are grounded in authoritative safety expectations: 1) OSHA Laws and Regulations, i.e., CFR 1910 Subparts O and N, 2) OSHA technical guidance documents, 3) NIOSH safety alerts, and 4) OEM manuals for our target machines.

4.1.2. Extract and Clean Text

For each document, we first determined whether it was machine-readable (i.e., if the text can be searched or selected in a standard PDF reader). For machine-readable PDFs, we used the `PyMuPDF` library in Python, which returns page-level text while preserving paragraph structure. If the document stored text as embedded images, we used the Mistral OCR (optical character recognition) API [18] to produce a text layer with page boundaries and paragraph breaks. After OCR, the output was treated the same as machine-readable text, ensuring a consistent representation across documents regardless of their original format.

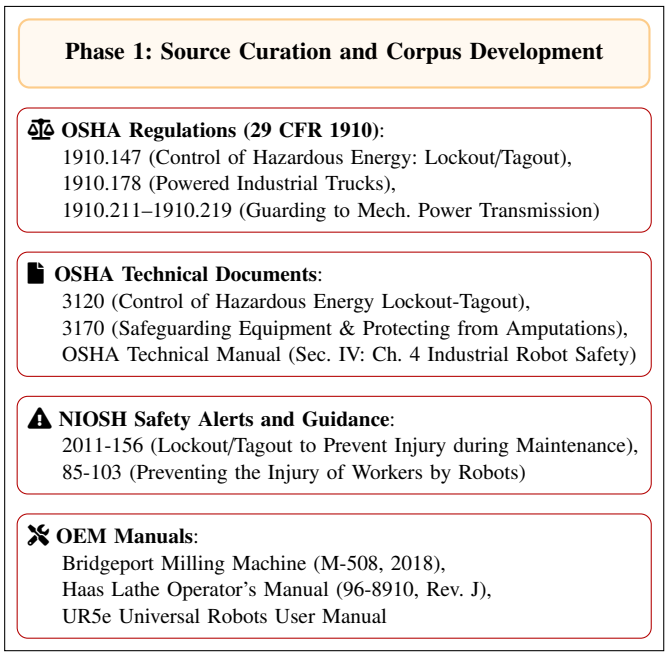


Figure 2. Regulatory, safety, and OEM documents forming our safety corpus.

4.1.3. *Segment content.*

For longer documents such as OSHA standards and OEM manuals, we used the table of contents to identify the start of chapters and subchapters, and split the document into those same sections. This segmentation keeps each unit of text aligned with meaningful instructional topics. Shorter narrative documents, such as certain NIOSH safety alerts, typically do not include a usable table of contents, so we applied a length-based rule: if the document exceeded 3,000 words, we divided it into smaller sections. This threshold balances retrieval performance in retrieval-augmented generation pipelines by preventing segments from becoming either too large or too fragmented.

4.1.4. *Remove repeated text and standardize file names.*

Once segmented, each section was saved as its own PDF, and any repeated header, footer, or boilerplate content was removed when present. We standardized filenames to include the source type (e.g., OSHA, NIOSH, OEM), the document identifier, the section title, and the original page range. This naming convention ensures that each file is traceable back to its authoritative source and simplifies manual review, debugging, and alignment in downstream retrieval steps.

4.2. *Phase 2: Knowledge Base and Indexing*

The goal of this phase is to construct the knowledge bases and retrieval mechanisms that the chatbot relies on when answering questions. Prior research has shown that different retrieval strategies can surface different types of relevant content [17]. Because manufacturing safety questions may be phrased in either precise terminology or broader hazard descriptions, we examined multiple retrieval methods in this phase rather than relying on a single indexing strategy. In later phases, our evaluations will determine which of these approaches performs best when answering the constructed benchmark questions developed in Phase 3.

4.2.1. *Build OpenAI keyword and semantic search indices.*

We uploaded the segmented PDF documents from Phase 1 into an OpenAI-managed vector store (i.e., a database that stores documents as numerical representations so they can be compared based on similarity). During upload, we defined a chunk size of 4,000 tokens (approximately 3,000 words; default is 800 tokens) so that our segmented documents are not further chunked by the API. The `OpenAI Retrieval API` supports two retrieval modes using the same stored data [19]. In keyword search mode, exact word matches between the query and the word matching between the query and stored content. In semantic search mode (search based on similarity of meaning rather than exact wording), the system rewrites or embeds the query to retrieve conceptually similar text. In our estimation, keyword searches are useful when users recall terminology directly from manuals, while semantic searches are useful when users describe their questions in everyday language.

4.2.2. *Build the BM25 index and retriever.*

We also constructed a `BM25 index`, using the BM25 retriever from the Python `LangChain` library, to support lexical search across the segmented corpus. `BM25` uses probabilistic term-weighting to rank documents and remains a strong baseline for retrieval since its introduction [20]. As a keyword-based retrieval system, it can also perform well on queries containing distinctive regulatory language, part labels, or procedural step names. Since we run `BM25` locally, it is free (no API costs) and has low latency.

4.2.3. *Construct the graph index.*

Graph-based retrieval methods allow the system to retrieve information not only by textual similarity but also by leveraging structural relationships across safety concepts, procedures, and component interactions. Following the `LangChain` library's `GraphRetriever` design [21], we implemented three graph RAG variants that differ in how they explore or constrain neighborhoods within the graph. First, `Graph-Eager` expands outward through connected nodes to gather broader contextual and procedural information, which is useful when a user describes a situation imprecisely. Second, `Graph-MMR` applies maximal marginal relevance to balance relevance with diversity, reducing redundant passages when multiple related segments exist in the same procedure or manual section. Third, we include a vanilla similarity search over *AstraDB*, which is simply the traditional $k-$nearest-neighbor chunk retrieval based on embedding similarity. This baseline does not traverse the graph and instead retrieves only the top k most similar text chunks, providing an anchor for measuring whether graph traversal adds value beyond standard vector search. Note that *AstraDB* serves as the online vector storage backend in all three approaches.

4.2.4. *Retrieval Routing Across Knowledge Bases.*

We implemented a unified retrieval router for our six retrieval mechanisms: (a) OpenAI keyword search, (b) Ope-

nAI semantic search, (c) `BM25` retriever, (d) `Graph-Eager`, (e) `Graph-MMR`, and (f) vanilla similarity search. The router standardizes how results are formatted and passed to the language model, ensuring that downstream evaluation compares retrieval strategies on an equal footing. Operationally, this means each retrieval strategy can be swapped in or out without modifying the rest of the chatbot pipeline, enabling controlled experimentation in Phase 4 and simplifying deployment in Phase 6.

5. Artifact Demonstration

The fourth stage of our DSR methodology is to "*demonstrate the use of the artifact to solve one or more instances of the problem*" [9].

We deployed the winning RAG pipeline as a publicly accessible multimodal chatbot interface that supports three natural input modes: (a) text chat, (b) image-based queries, and (c) voice (speech-to-speech) interaction. The chat interface enables workers to ask safety-related questions using everyday language, while the system retrieves and synthesizes answers grounded in the authoritative documents curated earlier in the pipeline. The image input mode allows the user to upload photographs of equipment, controls, or workspace conditions; the chatbot interprets the visual context and retrieves relevant safety passages for that specific situation. In addition, a speech input mode enables hands-free operation of the chatbot.

The deployed system was implemented as a `Streamlit` application [22] that integrates OpenAI's "Responses" and "Realtime" APIs to support text, image, and speech interactions through a unified interface. Retrieval relies on the same corpus and vector store structure established in the earlier phases, ensuring traceability and grounding of responses. The system also logs session-level analytics and provides transparent citations for retrieved content, supporting explainability and trust. A screenshot of the deployed application is shown in Figure 3, and the full code implementation is provided in the project repository (see "Data and Code Section Availability" section).

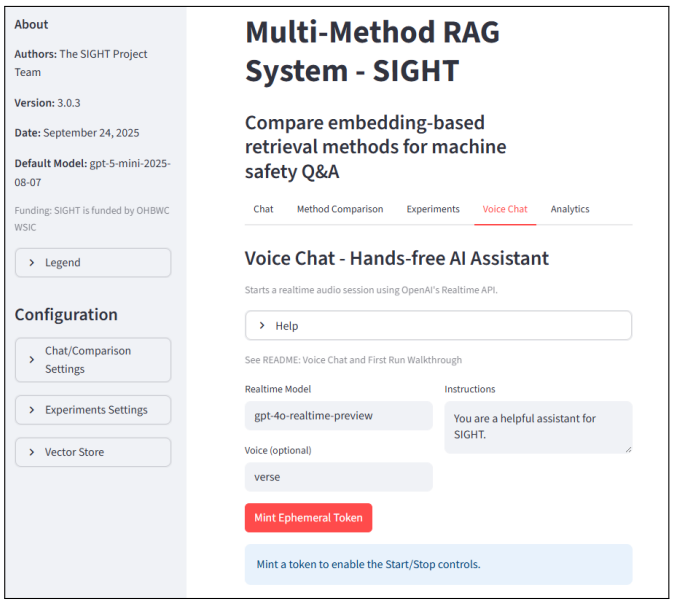


Figure 3. A screenshot of the voice module of our deployed chatbot. The chatbot is available at `https://sight.fsb.miamioh.edu/`.

6. Artifact Evaluation

Note: when evaluating artifact agasint accuracy mention that "instructional content must be validated against authoritative safety documentation and reviewed by domain experts prior to delivery to ensure factual precision and user trust, while avoiding improper actions and serious injury" , and consistent with official standards and manufacturer guidelines.

6.1. Phase 4: Benchmark Q&A Dataset Development

To evaluate the performance of proposed systems, we developed a comprehensive benchmark consisting of questions and gold-standard answers derived from the operational and safety manuals of the three target machines: the Bridgeport Manual Mill, the TL-1 CNC Lathe, and the Universal Collaborative Robot (Cobot). This benchmark was designed to assess the chatbot's ability to retrieve, reason about, and accurately convey safety-critical information across diverse levels of difficulty.

The process began with our research team conducting a detailed review of each machine's official manuals. From these materials, we systematically extracted and formulated questions that represent realistic scenarios faced by operators and trainees in manufacturing settings. Each question was paired with a corresponding answer grounded explicitly in the manuals or regulatory guidelines, ensuring factual correctness and traceability. For instance, an example of an easy question might be: "What is a pinch hazard in robot operation?" with the corresponding answer: "A pinch hazard occurs when body parts can be caught between moving robot parts or surfaces." Such questions can typically be answered from a single section of a manual. To better simulate realistic industrial information retrieval tasks, we also created more challenging questions that required integrating knowledge from multiple sections or documents. For example, a complex question might ask: "What are the external connection ports on the robot?" The answer ("Teach Pendant Port, SD Card slot, Ethernet, USB 2.0 and USB 3.0, Mini Displayport, 10A Mini Blade Fuse") requires synthesizing details scattered across different technical sections. This ensured the benchmark tested both direct retrieval and reasoning capabilities of the system.

After the initial Q&A pairs were derived, we implemented a validation phase to ensure content accuracy and industrial relevance. Each question and answer was reviewed either by external researchers with expertise in industrial safety or by experienced operators from our partner companies. Their feedback was used to refine the clarity, technical precision, and contextual relevance of the benchmark items. This iterative validation process guaranteed that the benchmark accurately represented the complexity of real-world safety knowledge and could be reliably used for evaluating AI system performance across different machines and hazard contexts.

In total, we created 60 questions for the UR5 Cobot, 51 for the Haas TL-1 lathe, and 42 for the Bridgeport manual mill. Due to space constraints, we focus on the UR5 Cobot, which exhibits greater operational complexity and a higher degree of human–machine interaction than the other two machines. The UR5 benchmark sufficiently tests the system's ability to retrieve and synthesize information across multiple text segments ("chunks") produced during corpus preprocessing. Including additional machines would provide no-to-little insight beyond what is demonstrated through the Cobot evaluation.

6.2. *Phase 5: Automated Evaluation and Experimental Design.*

We used a full-factorial experimental design to study how retrieval method, model specification, and prompting choices influenced the quality of answers generated by the AI. The purpose was to identify which configuration patterns consistently yielded correct responses across multiple correctness measures before advancing a small number of candidate configurations for human evaluation in Phase 5.

Table 1. Factors and levels in the Phase 4 full-factorial experimental design.

Factor	Purpose	Levels
Retrieval Approach	Determines how relevant safety text is selected	{`OpenAI Keyword`, `OpenAI Semantic`, `BM25`, `Graph Eager`, `Graph MMR`, `Vanilla`}
LLM Used	Controls model capacity and reasoning ability	{`gpt-5-mini-2025-08-07`, `gpt-5-nano-2025-08-07`}
Max Output Tokens	Constrains verbosity and potential reasoning detail for generated responses	{500, 1000, 2500, 5000}
Reasoning Effort	Sets structured reasoning depth for LLM	{minimal, low, medium, high}
Top–k Retrieval Depth	Number of chunks returned; for graph methods applied after traversal	{5, 10}
System Prompt	Evaluates stability under different answer formats	{A, B} *(Both are chain-of-thought prompts)*
Few-shot Sub-Prompt	Few-shot in-context learning to teach the LLM answering style	{A}
Total Pipelines = $6 \times 2 \times 4 \times 4 \times 2 \times 2 \times 1 =$ **3, 072**		

6.2.1. *Define experimental factors and levels.*

We vary seven factors, summarized in Table 1, resulting in 3, 072 retrieval-generation *pipelines*. We use a pipeline to refer to a specific combination of the seven factor levels used to generate an answer from the LLM/chatbot. Each pipeline is evaluated across all 50 benchmark questions, and the full experiment is replicated three times, yielding $3072 \times 50 \times 3 = 460,800$ generated responses per evaluation metric.

6.2.2. *Generate answers for each configuration.*

For each pipeline, we generate an answer to every benchmark question using that pipeline's specific settings. All outputs are stored together with their pipeline, question, and replicate identifiers to support consistent evaluation.

6.2.3. *Compute automated evaluation metrics.*

For each of the candidate pipelines, we computed a set of primary and secondary evaluation metrics using the 150 generated answers (50 benchmark questions $\times$ 3 replicates). The primary metrics corresponded to the three design requirements in Section 3. First, correctness was assessed using an LLM-as-judge that compares each generated answer to the gold reference answer and returns a binary decision. Second, latency was measured as the mean end-to-end response time for answer generation. Third, cost was recorded but not applied as a post-hoc filter, since cost constraints were built into the experiment design by restricting model selection to `gpt-5-mini-2025-08-07` and `gpt-5-nano-2025-08-07`. Note these LLMs are considered to be high performing and relatively low-cost. As of October 2025, the `gpt-5-mini-2025-08-07` was priced at \$0.250 / 1M input tokens and \$2/ 1M output tokens, and `gpt-5-nano-2025-08-07` was priced at \$0.050 / 1M input tokens and \$0.400 / 1M output tokens [23].

The secondary metrics provided additional characterization of answer quality. BLEU and ROUGE-L measured text similarity between generated and reference answers, while cosine similarity evaluated semantic alignment using TF–IDF vector representations [24–26]. We also computed three LLM-as-judge metrics that evaluate if the (a) retrieved context was relevant to the question, (b) final answer remained grounded in that context (rather than hallucinating), and (c) response was helpful and meaningfully addressed the user query. Each LLM-based judgment used standardized evaluation prompt templates provided in the LangChain "RAG Evaluation" tutorial [27]. The full set of metrics and their definitions are summarized in Table 2.

6.2.4. *Aggregate and rank pipeline performance.*

We first applied threshold-based filtering to enforce the design requirements from Section 3. For each pipeline, we computed the correctness proportion p based on the binary correctness outcomes from Step 4.3 across all 150 evaluations (50 questions $\times$ 3 replicates). Only pipelines with $p \geq 0.95$ were retained (i.e., an error rate of 5% or less). Among these accuracy-qualified pipelines, we then computed the mean end-to-end latency and retained only those with an average response time of 4 seconds or less. Cost was not applied as a post-hoc filter as the we already constrained model selection to low-cost configurations.

For the pipelines that meet both criteria, we constructed a composite performance ranking. For each accuracy metric in Table 2, we ranked pipelines within each question and replicate, then averaged these ranks to obtain a mean rank per metric. The composite score for each pipeline was defined as the average of its mean ranks across the seven accuracy metrics. The two pipelines with the lowest composite scores advanced to Phase 5 for blind expert evaluation.

Table 2. Summary of evaluation metrics used in our framework/experiment.

Type	Metric	Definition / Prompt Basis
Primary	Correctness vs. Reference	Binary decision from LLM-as-judge comparing the generated answer with its corresponding gold answer
Primary	Latency	Mean model response time
Primary	Cost	Token usage cost for `gpt-5-mini` and `gpt-5-nano`
Secondary	BLEU	n-gram precision with brevity penalty and smoothing (sentence-level BLEU)
Secondary	ROUGE-L	Longest common subsequence F-measure between generated and reference answers
Secondary	Cosine Similarity	Cosine similarity $\left(\frac{X \cdot Y}{\|X\| \|Y\|}\right)$ computed between TF–IDF vectors of generated (X) and reference (Y) answers to measure semantic alignment independent of length
Secondary	Retrieval Relevance	LLM-as-judge indicating if retrieved passages relate to the question
Secondary	Groundedness (Faithfulness)	LLM-as-judge check that answer does not introduce information outside retrieved evidence
Secondary	Helpfulness	LLM-as-judge check that the answer directly and meaningfully addresses the question

6.3. *Phase 6: Human Evaluation to Select Winning Configuration*

Arthur, please take a stab at this to set the details of how we will proceed.

7. Discussion

ABC

8. Concluding Remarks

ABC

Data and Code Availability

All code and evaluation benchmarks used in this study are publicly available at Add repo here, including the Streamlit-based app code, preprocessing scripts, and experiment pipelines.

Acknowledgments

This work was funded by the Ohio Bureau of Workers' Compensation Worker Safety Innovation Center (WSIC) Grant `#WSIC26-250206-009` and by the *Raymond E. Glos Professorship* and the *Dinesh & Ila Paliwal Innovation Chair* at Miami University. The authors also acknowledge the *Farmer School of Business* for providing the computing infrastructure used to host the chatbot developed in this research. The authors are also thankful for Engineered Profiles, Yamaha Motor Company of North America, and MeetKai, Inc. for their feedback on our benchmark Q&A datasets.

References

[1] Bureau of Economic Analysis, Gross domestic product 2nd quarter 2025 (third estimate), gdp by industry, corporate profits (revised), and annual update, https://www.bea.gov/sites/default/files/2025-09/gdp2q25-3rd.pdf (2025).

[2] U. B. of Economic Analysis, Table 6.4d. full-time and part-time employees by industry, https://apps.bea.gov/iTable/?reqid=19&step=2&isuri=1&1921=survey, accessed: October 28, 2025 (2025).

[3] National Center for Science and Engineering Statistics, Research and development: U.S. trends and international comparisons, https://ncses.nsf.gov/pubs/nsb20246/u-s-business-r-d, accessed: October 28, 2025 (2024).

[4] X. Xu, Y. Lu, B. Vogel-Heuser, L. Wang, Industry 4.0 and industry 5.0—inception, conception and perception, Journal of Manufacturing Systems 61 (2021) 530–535.

[5] Y. Lu, H. Zheng, S. Chand, W. Xia, Z. Liu, X. Xu, L. Wang, Z. Qin, J. Bao, Outlook on human-centric manufacturing towards industry 5.0, Journal of Manufacturing Systems 62 (2022) 612–627.

[6] U. B. of Labor Statistics, Employer-reported workplace injuries and illnesses – 2023, https://www.bls.gov/news.release/osh.htm, accessed: October 28, 2025 (2024).

[7] Liberty Mutual Insurance, Workplace Safety Index 2025, Annual report, Liberty Mutual Insurance, Boston, MA, retrieved from https://business.libertymutual.com/workplace-safety-index/ (2025).

[8] K. S. Kiangala, Z. Wang, An experimental hybrid customized ai and generative ai chatbot human machine interface to improve a factory troubleshooting downtime in the context of industry 5.0, The International Journal of Advanced Manufacturing Technology 132 (5) (2024) 2715–2733.

[9] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, A design science research methodology for information systems research, Journal of Management Information Systems 24 (3) (2007) 45–77.

[10] X. Zhu, R. Y. M. Li, M. J. C. Crabbe, K. Sukpascharoen, Can a chatbot enhance hazard awareness in the construction industry?, Frontiers in Public Health 10 (2022) 993700.

[11] J. Jeon, Y. Sim, H. Lee, C. Han, D. Yun, E. Kim, S. L. Nagendra, M. B. Jun, Y. Kim, S. W. Lee, et al., ChatCNC: Conversational machine monitoring via large language model and real-time data retrieval augmented generation, Journal of Manufacturing Systems 79 (2025) 504–514.

[12] T. Wang, J. Fan, P. Zheng, An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing, Journal of Manufacturing Systems 75 (2024) 299–305.

[13] S. Lou, R. Tan, Y. Zhou, Z. Zhao, Y. Zhang, C. Lv, Large language model-enabled cognitive agent for self-aware manufacturing, Journal of Manufacturing Systems 82 (2025) 1213–1226.

[14] H. Fan, C. Liu, N. E. Janvisloo, S. Bian, J. Y. H. Fuh, W. F. Lu, B. Li, MaViLa: Unlocking new potentials in smart manufacturing through vision language models, Journal of Manufacturing Systems 80 (2025) 258–271.

[15] S. Li, H.-L. Xie, P. Zheng, L. Wang, Industrial metaverse: A proactive human-robot collaboration perspective, Journal of Manufacturing Systems 76 (2024) 314–319.

[16] A. Gautam, M. R. Aryal, S. Deshpande, S. Padalkar, M. Nikolaenko, M. Tang, S. Anand, Iiot-enabled digital twin for legacy and smart factory machines with llm integration, Journal of Manufacturing Systems 80 (2025) 511–523.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 9459–9474.

[18] Mistral AI Team, Mistral OCR, https://mistral.ai/news/mistral-ocr, accessed on 2025-10-26 (Mar. 2025).

[19] OpenAI, Retrieval API, https://platform.openai.com/docs/guides/retrieval, accessed on 2025-10-26 (2025).

[20] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends in Information Retrieval 3 (4) (2009) 333–389.

[21] LangChain, Graph RAG, https://python.langchain.com/docs/integrations/retrievers/graph_rag/, accessed on 2025-10-26 (2025).

[22] streamlit, streamlit v1.50.0, Python library. Available at https://pypi.org/project/streamlit/ (Sep. 2025).

[23] OpenAI, Pricing, https://openai.com/api/pricing/, accessed on 2025-10-27 (Aug. 2025).

[24] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[25] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[26] G. Chowdhury, Introduction to Modern Information Retrieval, Facet Publications, Facet, 2010.

[27] LangChain, Evaluate a RAG application, https://docs.langchain.com/langsmith/evaluate-rag-tutorial, accessed on 2025-10-27 (2025).