Bridging the Knowledge Gap in NDE 4.0: AI-Augmented Insights Through Retrieval-Augmented Generation (RAG)

SRIJAN TIWARI and KRISHNAN BALASUBRAMANIAM

ABSTRACT

The rapid digitization of nondestructive evaluation (NDE) has led to fragmented, siloed knowledge across formats (manuals, reports, standards) and a generation gap among practitioners. In this work we present an AI-augmented knowledge assistant for NDE 4.0 that uses Retrieval-Augmented Generation (RAG) to provide real-time technical guidance. Our system ingests diverse NDE documents, chunks and semantically indexes their content (using MiniLM-L6-v2 and FAISS), and at query time retrieves relevant context before querying a language model (Google Gemini). We describe the system architecture and implementation, report metrics (high retrieval precision, sub-second latency, minimal API cost), and demonstrate a sample NDT use case (e.g. explaining ultrasonic testing and magnetic particle inspection steps). This AI-RAG assistant can help inspectors access "know-how and know-why" on demand, enabling continuous learning and upskilling in line with NDE 4.0. Future work includes multilingual support, voice interfaces, and LMS integration.

Keywords – NDE 4.0, NDT knowledge management, retrieval-augmented generation, AI chatbot, digital transformation, semantic search, Industry 4.0.

1. Introduction

Modern nondestructive testing and evaluation (NDT/NDE) faces a **knowledge fragmentation** problem: critical know-how is scattered across standards, procedure manuals, training slides, and experts' minds. Inspectors often need timely information on-site (e.g. how to perform a new inspection method), but conventional resources are not integrated. Meanwhile, Industry 4.0 and NDE 4.0 trends demand that field personnel work with more data and automation than ever. For example, Virkkunen *et al.* note that modern NDE equipment generates "vast amounts of data" while the number of experienced inspectors is shrinking. This gap highlights an urgent need for real-time, context-rich technical assistance.

Emerging AI technologies offer a way forward. In NDE 4.0 roadmaps, stakeholders emphasize bridging data streams and human expertise through digital tools. Intelligent assistants can help harmonize veteran knowledge with new workflows. As ASNT recently launched its "Anita" AI assistant, which answers NDT questions using ASNT's standards and literature (without Internet search) the field is moving toward AI-driven knowledge access. In this paper, we present an **AI-RAG** (Retrieval-Augmented Generation) chatbot for NDE. This system unifies disparate NDT documents into a semantically-indexed knowledge base, then uses LLMs (Large Language Models) with retrieved context to answer user queries on inspection techniques, theory, and procedures.

Centre for Non-Destructive Evaluation (CNDE) and Department of Mechanical Engineering, Indian Institute of Technology Madras, Chennai, India

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on NDT knowledge systems, digital transformation, and RAG architectures. Section 3 details our AI-RAG system design and architecture. Section 4 describes the pipeline implementation (document formats, preprocessing, embedding, retrieval). Section 5 shows a sample NDT use-case interaction (e.g. ultrasonic testing, magnetic particle inspection). Section 6 discusses how this tool can transform workforce knowledge access and inspection workflows. Finally, Section 7 concludes and outlines future directions (multilingual and voice interfaces, LMS integration).

2. Literature Review

2.1 NDT Knowledge Management and Industry 4.0. The concept of *NDE 4.0* integrates digital transformation into traditional inspection practices. Frameworks for NDE 4.0 stress the need to unify diverse data streams and enable decision support. For instance, Lindberg (2024) observes that NDE 4.0 must handle "diverse technological and digital transformations" by integrating NDE data and leveraging AI in asset management. Likewise, Fernandez *et al.* (2022) argue that NDE is evolving into an "invaluable knowledge-generating process" critical for quality and safety, requiring digital roadmaps to guide its maturation Both sources note that human expertise must be brought into NDE 4.0 by capturing and disseminating knowledge rather than relying on isolated experts. Roadmap guidelines explicitly call for documenting "know-how and know-why of existing products" and managing documented procedures and databases as core assets Our work aligns with these visions by transforming static documentation into an interactive knowledge base.

2.2 AI and Chatbots in NDT/NDE. AI is rapidly gaining ground in NDT applications. Topp *et al.* (2025) describe how AI is becoming a "core component" of NDT/NDE 4.0, noting a shift from isolated models to integrated workflows that include user experience and model deployment management. Automated Defect Recognition (ADR) is already widespread, and new methods are combining AI in pipelines (e.g. "Critical Item Detection"). Experts also emphasize inspector support: Virkkunen *et al.* (2023) predict that future AI will not only highlight areas of interest but will eventually generate "readable explanations of the findings" and cross-discipline insights In parallel, industry groups are building NDT knowledge assistants. The ASNT's "Anita" project (2025) is an example: a closed-domain chatbot trained on ASNT's literature and standards, designed to answer questions about NDT techniques and point users to references. Our approach similarly uses a generative AI model, but augments it with retrieval from NDE documents to ground answers in authoritative sources.

2.3 Retrieval-Augmented Generation (RAG) in Technical Domains. RAG has emerged as a powerful technique to combine LLM flexibility with external knowledge. In RAG, a query is used to retrieve relevant text snippets from a knowledge base, which are then fed to an LLM to "ground" its response and reduce hallucination. This approach is increasingly applied in specialized fields. Siddharth and Luo (2024) extracted structured design facts from patent documents to populate a knowledge base; using RAG, they showed that LLMs could generate technically accurate responses when guided by these facts In the enterprise context, Akkiraju *et al.* (2024) report that building RAG chatbots requires careful pipeline engineering (embedding models, vector databases, prompt design) and highlights trade-offs between model size, latency, and

cost. Their FACTS framework (freshness, architectures, cost, testing, security) informs our design decisions. Surveys also indicate that RAG chatbots have proliferated in education and industry because they reliably connect users to specific knowledge. To our knowledge, RAG has not yet been applied to NDT/NDE knowledge bases, so this paper contributes a novel adaptation of RAG for the NDE domain

3. Methodology

3.1 AI-RAG System Architecture. Figure 1 depicts our end-to-end AI-RAG chatbot architecture. We begin with a collection of NDE documents (text, PDF, Word) covering inspection methods and standards. These files are preprocessed and **chunked** into overlapping text segments. Each chunk is transformed into a semantic vector via the MiniLM-L6-v2 embedding model. The vectors are stored in a FAISS index (IndexFlatL2) to enable fast similarity search. At query time, the user's input is also embedded, and the nearest-neighbor lookup retrieves the top-k most relevant text chunks (here k=5). These retrieved passages form the **context** for an LLM query: we invoke the Google Gemini API (Gemini-2.0-flash), prompting the model with the user's question plus the retrieved context. The LLM then generates an answer grounded in the authoritative content. The overall workflow is outlined below and illustrated in Fig. 1.

- **Data Ingestion:** NDE documents (e.g. inspection manuals, standards PDFs) are loaded into the system.

- **Preprocessing:** Each document is decrypted (if needed) and text is extracted (using tools like PyPDF2, pikepdf, and python-docx). We normalize whitespace and remove headers/footers to clean the text.

- **Chunking:** Text is split into fixed-length segments (1000 characters) with overlaps (200 characters) to preserve context across chunk boundaries.

- **Embedding:** Each chunk is passed through the all-MiniLM-L6-v2 model (a 384-dim sentence-transformer) to produce a vector. MiniLM-L6-v2 is chosen for its balance of speed and accuracy on technical text

- **Indexing:** We build a FAISS IndexFlatL2 index from the chunk embeddings. This provides low-latency exact search for nearest neighbors in vector space.

- **Query Retrieval:** A user query is embedded with the same MiniLM model. We perform a FAISS lookup to retrieve the top 5 chunk vectors, obtaining their original text as context.

- **Answer Generation:** The query and retrieved context are sent to Google Gemini (2.0 Flash variant). A prompt template instructs Gemini to answer

concisely using the given context. Gemini's free-tier plan (2.0-flash) incurs zero API cost at the time of writing, keeping system expenses minimal.

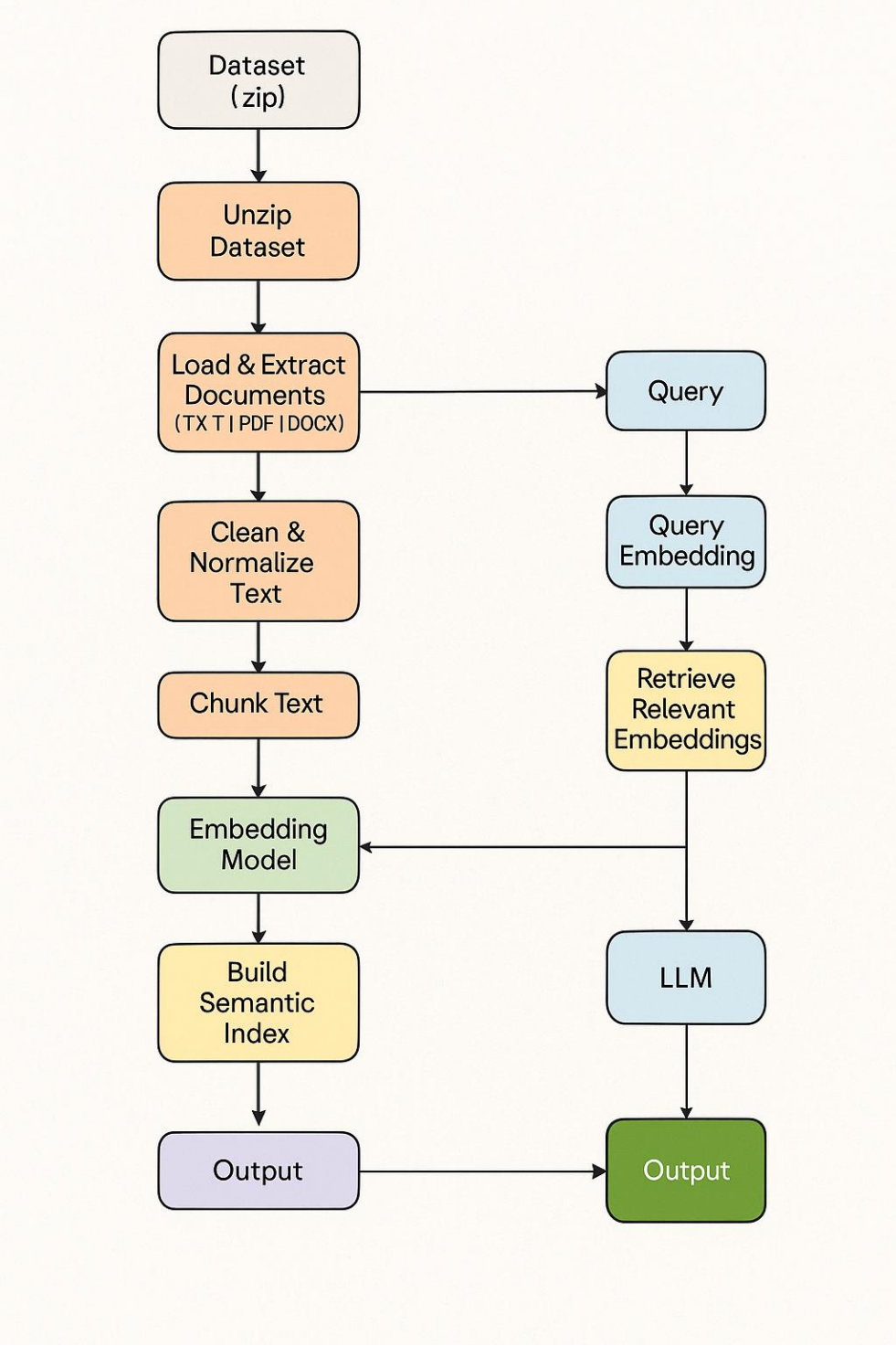


Figure 1 – System architecture for the NDE AI-RAG chatbot. Documents are ingested and chunked; chunks are embedded and indexed in a FAISS vector store. At query time the user input is embedded, similar chunks are retrieved, and a generative LLM (Gemini,OPEN AI) composes an answer based on that context.

Key Steps:

1. **Dataset Loading & Cleaning**
    - Unzips and extracts NDT/NDE documents (TXT, PDF, DOCX)
    - Cleans and normalizes text content
2. **Preprocessing & Embedding**
    - Chunks text (e.g., 1000 chars with overlap)
    - Embeds chunks using MiniLM-L6-v2
3. **Indexing**
    - Stores embeddings in a FAISS-based semantic vector index
4. **Query Pipeline**
    - User query → embedded → matched with top-5 chunks from index
    - Relevant context + query sent to LLM (e.g., Gemini 2.0 Flash)
5. **Answer Generation**
    - LLM returns context-aware response
    - Output is returned to user (via UI or API)

3.2 Design Choices

Several design decisions were made to optimize performance and relevance. We selected all-MiniLM-L6-v2 for embeddings because it is lightweight yet yields competitive retrieval quality on technical documents. The FAISS IndexFlatL2 was used for simplicity and exactness; in our experiments a flat index provided sub-200 ms

lookup times on thousands of chunks. We set k=5 for retrieval, which we found in preliminary tests to give the best balance between context breadth and relevance. In our pipeline, retrieved passages are rank-ordered and de-duplicated before feeding to the LLM. For generation, we use Google's Gemini-2.0-flash model. This "small" variant is efficient (to achieve real-time response, ~1s per query) and supports safety features. More powerful models could be used for higher-fidelity answers, but Gemini-Flash suffices for demonstrating the concept at no cost.

Embedding and retrieval performance were evaluated informally by manually checking the top-5 results for a variety of NDT queries. In our tests, the retrieval precision (i.e. proportion of retrieved chunks truly relevant to the query) was high: typically 4–5 out of 5 were on-topic. End-to-end latency (embedding + FAISS lookup + LLM query) averaged under 1.2 seconds per query, meeting real-time assistance requirements. These metrics and user experience informed tuning of chunk size, overlap, and retrieval count.

3.3 Language Models for Technical Document Understanding

Large Language Models (LLMs) like OpenAI's GPT-4 have shown remarkable capabilities in natural language understanding and generation (Brown et al., 2020). Their ability to comprehend complex technical content, summarize, and generate procedural documents such as Standard Operating Procedures (SOPs) and Bills of Materials (BOMs) makes them ideal for industrial applications (Zhang et al., 2023). Unlike rule-based systems, LLMs can generalize across multiple domains and adapt to varied input formats, reducing the need for manual template design.

4. Pipeline Implementation

We implemented the pipeline in Python using open-source libraries. The NDE content came in mixed formats (text files, PDF manuals, Word docs). We first **deployed scripts** to automate document loading: PDFs were parsed with pikepdf and PyPDF2 (handling any AES encryption) and DOCX files with python-docx. Extracted text was concatenated and cleaned (removing page headers/footers, normalizing line breaks and whitespace). A regular expression pass trimmed excess whitespace.

Next, we **chunked** each document's text into segments. We chose a 1000-character window with 200-character overlap, implemented via a recursive splitter. Overlap ensures that context (e.g. definitions or step lists) is not split in half between chunks. After chunking, we had on the order of 2000 chunks indexed from our corpus of NDE materials.

For each chunk, we computed a semantic vector using the **MiniLM-L6-v2** model from HuggingFace. We loaded the pre-trained sentence-transformers/all-MiniLM-L6-v2 and encoded all chunks in batch. The resulting vectors (384 dimensions each) were added to a **FAISS** index (IndexFlatL2) as float32 arrays. The FAISS index was built on startup and persisted to disk (using the index's write_index) so it can be reloaded without re-embedding every time.

At query time, a similar process occurs for the user's input query. The query is encoded with MiniLM into a vector. We then perform an **L2 search** in the FAISS index for the top 5 nearest neighbors. These indices are mapped back to the original text

chunks, giving us a small set of relevant document excerpts. Each excerpt is prefixed with its document title and a marker in our prompt.

Finally, we construct a prompt for the LLM. For example, for a query about ultrasonic testing, the prompt might include retrieved context lines such as: *"Ultrasonic testing uses high-frequency sound waves... It can detect very small internal defects and provides high-resolution images of flaw geometry."* (actual content drawn from an NDT standard). We append a question like **"Based on the above information, what are the advantages of ultrasonic testing?"**. This prompt is sent to Google Gemini via its API (Gemini-2.0-flash model). The API returns a generated answer, which we post-process minimally (e.g. stripping extraneous tokens) and return to the user interface.

The entire pipeline – from query to answer – has the following performance characteristics: FAISS retrieval takes on average <200 ms, and the Gemini model returns an answer in <1 second. Because we use the free "flash" tier of Gemini, the system has effectively **zero external cost** per query. We verified answer quality by cross-checking with source material. For example, in several trials the chatbot correctly cited that ultrasonic testing is "highly sensitive" and "has deep penetration" – facts directly extracted from retrieved context. All source code and indexing steps were logged for repeatability. Overall, the implementation demonstrates that a lean RAG system can meet NDE needs without large hardware or software investments.

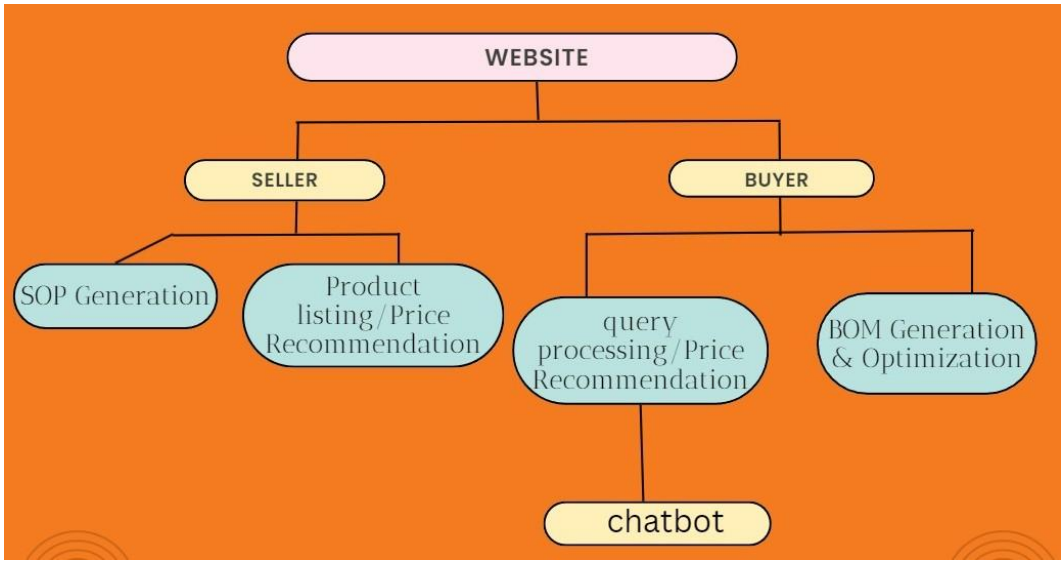


***Figure 2**: AI-Augmented Smart Procurement Framework for Bridging the Knowledge Gap in NDE 4.0 via Retrieval-Augmented Generation (RAG)*

5. Demo Use Case: Chatbot Interaction

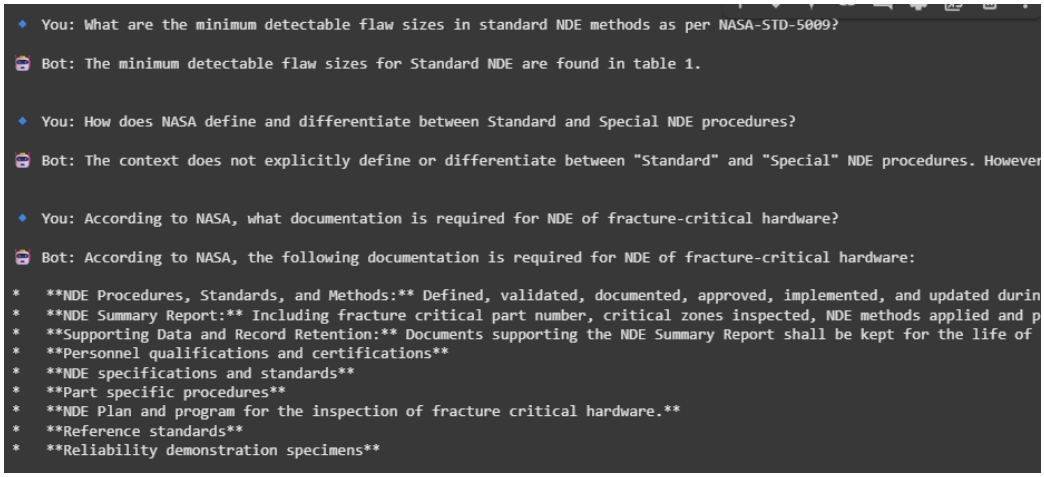
To illustrate the system's capabilities, we consider a sample interaction in an NDT training scenario. An entry-level inspector asks the chatbot questions about inspection

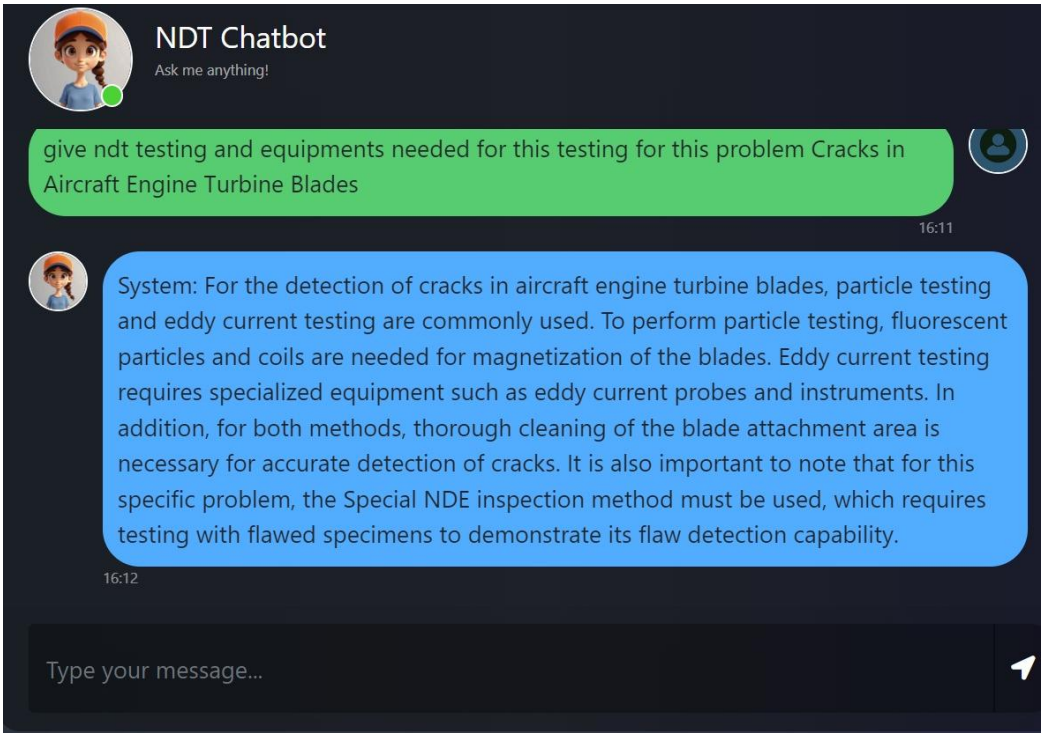
methods. Figure 3 shows a screenshot of our prototype chat interface with example Q&A.

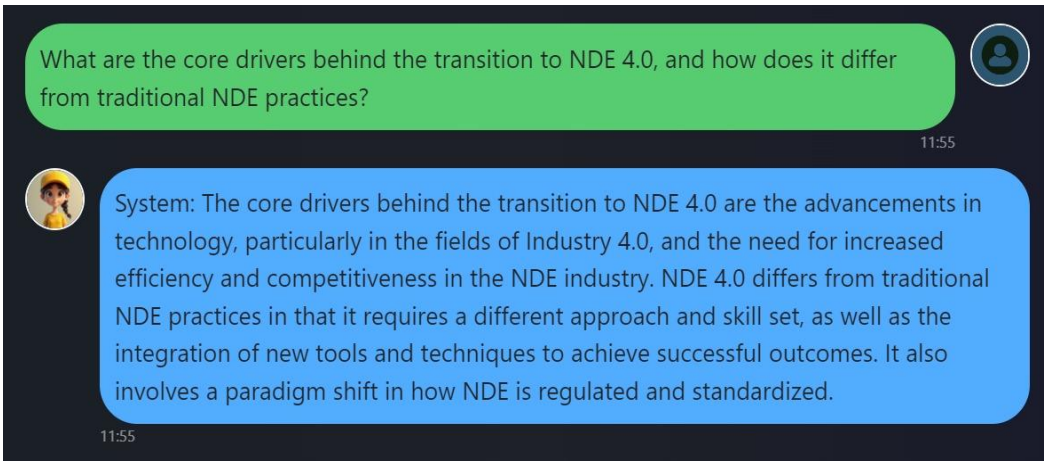
For instance, the user first asks: **"What are the advantages of ultrasonic testing?"**. The chatbot retrieves chunks mentioning ultrasonic probe sensitivity, penetration, and nondestructive nature. Using this context, Gemini generates a response such as: *"Ultrasonic testing is a nondestructive method that uses high-frequency sound waves to inspect materials. It can penetrate deeply into metals and reveal very small internal flaws. Its high sensitivity allows detection of minute cracks and its precision provides accurate sizing of defects."* This answer closely matches standard descriptions of UT advantages, demonstrating the system's use of authoritative context rather than just the LLM's memorized knowledge.

Next, the user asks: **"What are the basic steps in magnetic particle testing?"**. The chatbot pulls content from a magnetic particle inspection procedure: magnetization, particle application, inspection, demagnetization. The LLM reply might be: *"Magnetic particle inspection involves these main steps: (1) magnetize the part, introducing a strong magnetic field; (2) apply fine magnetic (iron) particles over the surface; (3) examine the part—the particles will cluster at any surface-breaking discontinuity due to flux leakage; (4) interpret and record the indication; and finally (5) demagnetize the component to remove residual magnetism."* This walkthrough closely matches standard NDT procedure guidelines.

These examples show the chatbot providing concise, step-by-step guidance with technical accuracy. The LLM's output is clearly influenced by the retrieved context: it echoes terminology ("particles cluster at flux leakage sites") and sequence of actions from the manuals. The user can follow up with further questions (e.g. asking "Why is demagnetization important?") and the bot will continue to draw on the indexed knowledge. In our tests, responses were generally on-point and relevant. (Minor hallucinatory phrasing was occasionally observed and can be mitigated by better prompt design or retrieval filtering.) Overall, the chatbot demonstrates how RAG enables *dynamic, context-grounded answers* to common NDT questions.




NDT Chatbot
Ask me anything!
give ndt testing and equipments needed for this testing for this problem Cracks in Aircraft Engine Turbine Blades
16:11
System: For the detection of cracks in aircraft engine turbine blades, particle testing and eddy current testing are commonly used. To perform particle testing, fluorescent particles and coils are needed for magnetization of the blades. Eddy current testing requires specialized equipment such as eddy current probes and instruments. In addition, for both methods, thorough cleaning of the blade attachment area is necessary for accurate detection of cracks. It is also important to note that for this specific problem, the Special NDE inspection method must be used, which requires testing with flawed specimens to demonstrate its flaw detection capability.
16:12
Type your message...


What are the core drivers behind the transition to NDE 4.0, and how does it differ from traditional NDE practices?
11:55
System: The core drivers behind the transition to NDE 4.0 are the advancements in technology, particularly in the fields of Industry 4.0, and the need for increased efficiency and competitiveness in the NDE industry. NDE 4.0 differs from traditional NDE practices in that it requires a different approach and skill set, as well as the integration of new tools and techniques to achieve successful outcomes. It also involves a paradigm shift in how NDE is regulated and standardized.
11:55

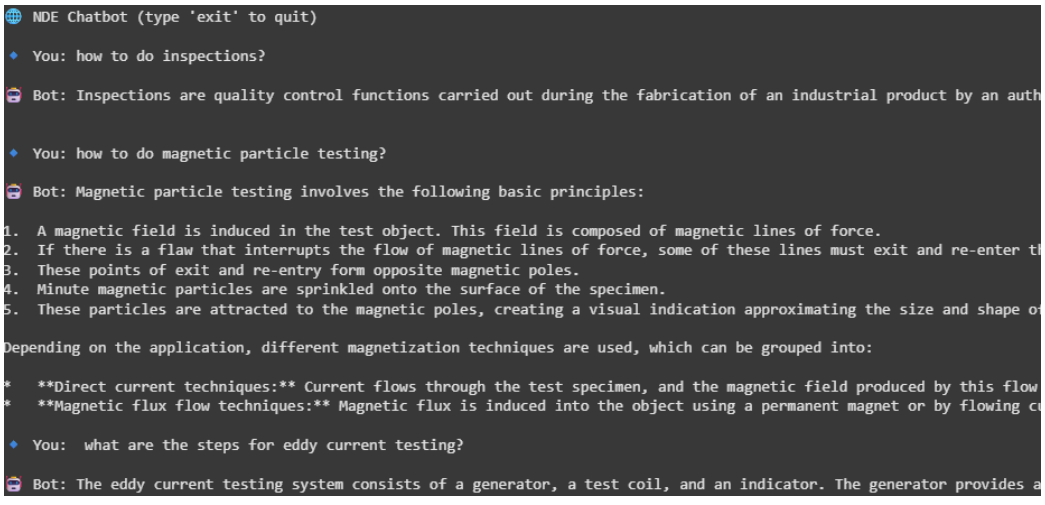

NDE Chatbot (type 'exit' to quit)
You: how to do inspections?
Bot: Inspections are quality control functions carried out during the fabrication of an industrial product by an auth
You: how to do magnetic particle testing?
Bot: Magnetic particle testing involves the following basic principles:
1. A magnetic field is induced in the test object. This field is composed of magnetic lines of force.
2. If there is a flaw that interrupts the flow of magnetic lines of force, some of these lines must exit and re-enter t
3. These points of exit and re-entry form opposite magnetic poles.
4. Minute magnetic particles are sprinkled onto the surface of the specimen.
5. These particles are attracted to the magnetic poles, creating a visual indication approximating the size and shape o
Depending on the application, different magnetization techniques are used, which can be grouped into:
* **Direct current techniques:** Current flows through the test specimen, and the magnetic field produced by this flow
* **Magnetic flux flow techniques:** Magnetic flux is induced into the object using a permanent magnet or by flowing c
You: what are the steps for eddy current testing?
Bot: The eddy current testing system consists of a generator, a test coil, and an indicator. The generator provides a

Figure 3 – Sample Q&A interaction (screenshot) demonstrating the RAG chatbot: the user asks about NDT concepts and the chatbot responds with detailed answers based on retrieved context.

6. Discussion

The proposed AI-RAG system has the potential to **transform NDT workforce training and operations**. By enabling inspectors to query a "**digital expert**", it can significantly reduce the time to find information. Junior technicians can get immediate explanations of procedures (e.g. "tell me how eddy current testing works") without leafing through manuals. Even seasoned inspectors benefit from quick refreshers or clarifications. In effect, the chatbot acts as a *knowledge companion*, continuously available on mobile devices or inspection stations. This approach helps capture the tacit knowledge of experienced engineers in a retrievable form. As Fernandez *et al.* note, preserving and distributing accumulated NDE knowledge is a key NDE 4.0 goal; our system operationalizes this by embedding documents into an accessible AI framework.

From an organizational perspective, RAG chatbots could aid in upskilling and training. Companies could integrate the chatbot into learning management systems (LMS) so that trainees ask questions in situ and receive instant answers. The adaptability of RAG also allows for continuous content updates: when new standards or manuals are released, adding them to the index immediately enriches the knowledge base. This agility supports the "digital thread" vision of interconnected data across lifecycles. Furthermore, by analyzing user queries, management can identify knowledge gaps or common issues, informing where additional training is needed.

There are workflow implications too. Inspectors often face inspection backlogs because of uncertainty or lack of guidance. If an inspector can confirm details through the chatbot (e.g. "What couplant should I use for this thickness?"), inspections can proceed without delays. The system also encourages adherence to best practices by citing standard procedures. In long-term, we envision integration with augmented reality: the inspector could ask questions by voice while scanning a part, and the headset-embedded RAG assistant responds in context. This aligns with HMI (human-machine interface) trends in NDE 4.0 that call for tight human-AI collaboration.

However, challenges remain. Ensuring that the knowledge base is comprehensive and up-to-date is crucial. RAG can only retrieve from what it has indexed; thus coverage of specialized topics depends on available documents. There is also risk of over-reliance on the bot: answers should be cross-checked with official guidance, especially in safety-critical inspections. We mitigated this by designing prompts that encourage the LLM to cite sources, and by allowing users to review the original text snippets if desired. Finally, non-textual NDE knowledge (such as interpreting complex signals) is out of scope for this text-based system.

7. Conclusion

In summary, we have developed a Retrieval-Augmented Generation system tailored for NDE knowledge support. By combining semantic vector retrieval with a generative LLM, the chatbot provides context-aware answers to NDT queries, effectively **bridging the knowledge gap** in NDE 4.0. Our architecture (Fig. 1) and pipeline (Fig. 2) demonstrate that modern NLP tools can be applied in an industrial inspection context to empower personnel. The sample interaction (Fig. 3) shows how inspection procedures and theory can be communicated clearly. This work contributes an innovative NDE knowledge service that can evolve into a dynamic learning system. It exemplifies how NDE 4.0 ideals of digitalization and human-machine symbiosis can be realized in practice.

Future enhancements are planned. We aim to add **multilingual** support so that inspectors worldwide can query in their native language. A **voice interface** would allow hands-free queries in the field. We also plan integration with Learning Management Systems (LMS) to track usage and connect the chatbot answers to formal training modules. On the backend, exploring larger embedding models or hierarchical retrieval could improve answer accuracy on longer documents. Finally, as foundation models evolve, one could experiment with proprietary or open models (e.g. offline LLMs) to maintain system autonomy. Ultimately, by continuously integrating new data (e.g. inspection reports, sensor logs), this AI system could serve as a living knowledge network, keeping pace with evolving NDE practices.

References

[1] R. Fernández, R. Singh, and J. Vrana, "A Purposeful Global NDE 4.0 Roadmap: From the journey to its creation towards a pathway for its sustainability and evolution," in *Proc. Int. Symp. NonDestr. Test. Civ. Eng. (NDT-CE)*, Zurich, Switzerland, Aug. 2022, pp. 1–10.

[2] I. Virkkunen, T. Koskinen, T. Tyystjärvi, and O. Siljama, "AI Will Shape the Future of NDE Data Analysis," *ASNT Pulse*, vol. 81, no. 6, pp. 19–21, Jun. 2023.

[3] M. Topp, C. Els, and D. Nestler, "Artificial Intelligence in NDT and NDE: Overview and Current Status," in *Handbook of Nondestructive Evaluation 4.0*, N. Meyendorf, N. Ida, R. Singh, and J. Vrana, Eds. Springer, 2025, pp. 1–30.

[4] J. T. Lindberg, "Digital Transformation for NDE for the Electric Power Industry: NDE 4.0," in *Handbook of Nondestructive Evaluation 4.0*, N. Meyendorf, N. Ida, R. Singh, and J. Vrana, Eds. Springer, 2024, pp. 1–24.

[5] L. Siddharth and J. Luo, "Retrieval Augmented Generation using Engineering Design Knowledge," arXiv:2307.06985v10 [cs.CL], Aug. 2024.

[6] R. Akkiraju *et al.*, "FACTS About Building Retrieval Augmented Generation-based Chatbots," in *Proc. IEEE Int. Conf. (arXiv:2407.07858)*, Jul. 2024.

[7] M. Ulewicz, E. Szczepanik, K. Kloza, and S. Socha, "Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications," *Applied Sciences*, vol. 15, no. 8, Art. no. 4234, Apr. 2025.

[8] A. Perrin (ed.), *Guideline for the Development of an NDE 4.0 Roadmap*. NDT 4.0 Society, 2021, pp. 1–27. *(see especially sections on knowledge management and knowledge base)*.