# Development and Demonstration of a Local RAG-Based QA System Using an Open-Source LLM for the SMR Level 1 PSA Report

Byun, Hyeonho [a], Kim, Hyeonmin [*a]
[a]*111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon 34057, Republic of Korea*
[*]*Corresponding author: hyeonmin@kaeri.re.kr*

***Keywords :*** PSA, RAG, LLM

## 1. Introduction

Probabilistic safety assessment (PSA) performed for nuclear power plants is typically composed of event-sequence-based Event Trees (ETs), system failure-logic-based Fault Trees (FTs), and quantitative results summarized in tables, together with numerous modeling assumptions and analysis inputs. PSA results are used in a wide range of safety-related activities, including risk-informed decision-making, identification of safety improvement items, and impact assessments for design and operational changes; therefore, traceability and verifiability of both the analysis process and results are essential. However, PSA deliverables are voluminous and structurally complex, requiring considerable expert time to search for and cross-check key information such as success criteria for specific event sequences, definitions of basic events, importance measures, and top-level risk metrics.

Recently, large language model (LLM)-based question-answering (QA) tools have shown the potential to automate document search and summarization through natural language interfaces. Nevertheless, in the nuclear domain, design and safety assessment materials often contain confidential or sensitive information, which limits the applicability of workflows that rely on uploading original documents to commercial, externally hosted LLM services from an information protection and security perspective. In addition, the hallucination phenomenon poses a critical concern for PSA tasks that are directly related to safety, since incorrect information may affect expert judgement and decision-making. Accordingly, practical use of generative AI in nuclear PSA requires: (i) an architecture operable in a local environment, (ii) evidence-grounded answer generation with explicit source attribution, and (iii) safety-oriented behaviors under retrieval failure or insufficient evidence (e.g., refusal to answer and/or explicit uncertainty reporting).

Retrieval Augmented Generation (RAG) is a pipeline in which relevant document chunks are retrieved first and then provided as conditions for answer generation, making it advantageous for source attribution and for reducing hallucinations. RAG is also suitable for safety assessment environments where documents are continuously updated, because the knowledge base can be updated without re-training (i.e., fine-tuning) the underlying pre-trained model. However, PSA reports contain not only narrative text but also semi-structured and non-textual elements such as ET/FT diagrams, tables. When applying an open-source LLM that supports only text embeddings, critical information embedded in these structural elements can be missed during retrieval. Therefore, it is essential to convert PSA structural components (e.g., ET branching logic and quantitative values in tables) into searchable text representations, and to design chunking and embedding strategies that preserve query relevance for reliable question answering.

In this study, we propose a PSA-oriented QA framework that combines a locally deployable open-source pre-trained LLM (gpt-oss:20b) with RAG to enable evidence-grounded QA under confidentiality constraints. We develop a preprocessing module to convert ETs and tables in PSA reports into text, and we implement chunking and indexing methods designed to preserve query relevance for vector database ingestion. As a demonstration, we construct a database for the SMR Level 1 PSA report and perform a set of representative queries, producing answers accompanied by explicit references to source locations (document sections and/or converted units). Through this work, we aim to show that the proposed approach can improve the efficiency of PSA document exploration and verification while maintaining reliability from a safety perspective via evidence-based responses.

## 2. Methods

Figure 1 shows the overall architecture of the proposed PSA-oriented QA framework.

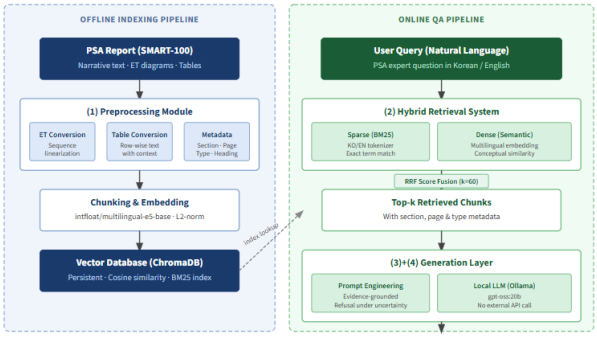


**Figure 1. Overall architecture of the proposed PSA-oriented RAG-based QA framework**

The system operates entirely in a local environment to maintain confidentiality of proprietary PSA documents. The pipeline consists of four main compartments: (1) preprocessing module for structural element conversion, (2) hybrid retrieval system combining sparse and dense methods, (3) locally-hosted open-source LLM (gpt-oss:20b) for answer generation, and (4) prompt engineering layer enforcing evidence-grounded responses with explicit source attribution. The gpt-oss:20b model (OpenAI, 2025), an open-weight MoE model with 21B total parameters, available via Ollama. All document embeddings and vector storage are managed through ChromaDB persistent client, enabling continuous updates without model retraining.

### *2.1 Preprocessing Module for PSA Structural Elements*

PSA reports contain not only narrative text but also semi-structured components such as ETs, FTs, and quantitative tables. Since the selected open-source LLM supports only text embeddings, critical information embedded in theses structural elements must be converted into searchable text representations while preserving their logical relationships and quantitative values.

**Event Tree Conversion:** ET diagrams are parsed to extract (1) initiating event definitions, (2) safety function headers and success criteria, (3) branching logic (success/failure paths), and (4) end-state identifiers. Each branch sequence is represented as a linearized text string following the format: "Initiating Event → Safety Function 1 [Success/Failure] → Safety Function 2 [Success/Failure] → ... → End State [ID]." This enables retrieval of complete event sequences and their associated success criteria through natural language queries.

**Table Conversion:** Quantitative tables (e.g., importance measures, frequency results, basic event probabilities) are converted into row-wise text entries. Each row is formatted as: "Parameter: [value], Context: [table caption and column headers], Source: [section number]." This preserves the numerical data and its semantic context for embedding.

**Metadata Enrichment:** All converted text chunks are augmented with structured metadata including section number, page number, document type (narrative/ET/table), and parent headings, enabling precise source attribution during answer generation.

### *2.2 Hybrid Retrieval Architecture*

To address the challenge of accurately retrieving relevant information from PSA documents containing mixed Korean-English technical terminology, we implement a hybrid retrieval system that combines lexical matching and semantic similarity. The BM25 component employs a custom tokenizer using regular expressions to separately extract Korean Hangul tokens and English alphabetic tokens. This ensures that exact matches for technical terms are properly captured regardless of language. The semantic component uses the ***intfloat/multilingual-e5***-base sentence transformer model, which applies a "query:" prefix during encoding to distinguish query representations from document representations. Embeddings are L2-normalized and stored in ChromaDB with cosine similarity as the distance metric.

For a given query $q$, both retrieval methods independently rank documents. The BM25 ranker produces rank $r_{\mathrm{BM25}}(d)$ for document $d$, while the dense retriever produces rank $r_{\mathrm{dense}}(d)$. The final relevance score is computed as:

$$\mathrm{RRF(d,q)} = \frac{w_{BM25}}{k + r_{BM25}} + \frac{w_{dense}}{k + r_{dense}} \qquad (1)$$

where $w_{BM25} = w_{dense} = 0.5$ (equal weighting), and $k = 60$ (standard RRF constant, (RRF: Reciprocal Rank Fusion)). The top-$k$ documents are selected after retrieving and scoring 2$k$ candidates from each method. This hybrid approach balances precision for specific term queries (via BM25) and recall for conceptual queries (via dense retrieval).

### *2.3 Evidence-Grounded Answer Generation*

The retrieved document chunks serve as the sole context for answer generation. The prompt template explicitly constrains the LLM to ground all statements in the provided context, ensuring that answers are derived exclusively from retrieved chunks. The model is instructed to refuse answering if the evidence is insufficient, thereby avoiding hallucination-based responses. All responses must include explicit citations of source locations, such as section numbers and document types (e.g., “Section 3.2.1 ET conversion”, “Table 4-5”), to enable traceability and verification by human experts.

To encourage accurate use of domain-specific terminology, the prompt assigns the LLM a role as "SMR PSA expert," which helps maintain consistency with nuclear engineering and PSA vocabulary conventions. Additionally, the system is designed to produce explicit uncertainty statements rather than generating plausible but incorrect answers when retrieved chunks do not contain relevant information. This refusal behavior under uncertainty is critical for safety-related applications where incorrect information may affect expert judgment and decision-making.

### *2.4 Evaluation Methodology: QA Performance Assessment Using SMR Level 1 PSA Report*

To evaluate the QA performance of the proposed RAG system, a vector database was constructed exclusively from the SMR-100 Level 1 PSA report (full internal-event analysis). Restricting the knowledge base to a single, well-characterized document focuses the evaluation on retrieval precision and answer quality rather than cross-document disambiguation. Seven

representative queries (Q1–Q7) were selected to cover three categories of information needs commonly encountered in PSA review practice, as summarized in Table 1.

Category I (Simple Fact Retrieval and Quantitative Queries, Q1–Q4) targets single numerical results or structured criteria directly accessible from tables and body text: total CDF (Q1), PSIS success criteria (Q2), total number of analyzed initiating events (IEs, Q3), and the LOOP CDF contribution (Q4). These queries assess the system's ability to locate key quantitative results and success criteria—the most elementary yet frequently needed tasks in PSA review. Category II (Comparative Queries with Multi-hop Reasoning, Q5–Q6) requires integrating information across multiple document entries: a multi-attribute comparison of the SLOCA1 and SLOCA2 initiating events (Q5), and identification of the top three CDF contributors with quantified percentages (Q6). These queries test whether the system can aggregate and rank data across multiple retrieved chunks. Category III (System Logic and Accident Progression Queries, Q7) requests a detailed comparison of the accident progression between SLOCA1 and SGTR (Q7), requiring synthesis of multi-attribute information scattered across different sections of the document—the most demanding query type from a reasoning standpoint.

Each response was evaluated on two dimensions: content accuracy and source accuracy. Content accuracy assesses whether the generated answer correctly and completely reflects the information in the SMR PSA report. Source accuracy examines whether the cited section numbers, table identifiers, and page references correspond to actual document locations. Evaluation was performed independently by two LLM-based evaluators (GPT-4-turbo and Claude-opus-4.6) and supplemented by manual verification against the original report. Each dimension was scored on a three-level scale: Accurate (○), Partial Error (△), or Inaccurate (✕).

## 3. Results and Discussion

Table 1 summarizes the evaluation results for all seven queries. Each response was independently evaluated by GPT-4-turbo, Claude-opus-4.6, and manual document verification; the combined judgment determined the final rating for each dimension.

**Table 1. QA evaluation results for seven queries**

| Q# | Category | Content Accuracy | Source Accuracy |
|---|---|---|---|
| Q1 | I | ○ | ○ |
| Q2 | I | ○ | △ |
| Q3 | I | ○ | ○ |
| Q4 | I | ○ | ○ |
| Q5 | II | ○ | ○ |
| Q6 | II | ○ | ○ |
| Q7 | III | △ | △ |

○ Accurate, △ Partial Error, ✕ Inaccurate

For Category I (Q1–Q4), the system demonstrated reliable performance on direct numerical retrieval. Q1 (Total CDF = $6.20 \times 10^{-8}$/yr) and Q4 (LOOP contribution = 0.74%, $4.60 \times 10^{-10}$/yr) returned exact values with correct citations to Table 7-1 (p. 405). Q2 (PSIS success criteria) was content-accurate—correctly identifying the 2/4 train criterion for PSIS and PSIS-R—but exhibited a source error: the reported page number (p. 14) corresponded to the table-of-contents entry rather than the actual body location of Section 5.3.1.2.1 (p. 230), suggesting that page-number metadata in the vector store may reflect chapter listings rather than true content locations. Q3 shows accurate context based on exact source from PSA report.

Category II queries (Q5–Q6) yielded fully accurate results for both dimensions. The SLOCA1 vs. SLOCA2 comparison (Q5) correctly identified the key distinguishing feature (PRHRS requirement for SLOCA1 only, per Section 4.2.2.1.1), system operation differences, and the shared occurrence frequency ($1.51 \times 10^{-4}$/yr, from NUREG/CR-6928 data), all with accurate citations. The top-CDF-contributor ranking (Q6) accurately computed the combined SLOCA1/2 contribution (37.54% + 37.53% ≈ 75.1%) and correctly identified GTRN (13.1%) and Excessive LOCA (6.9%) as the second and third contributors, demonstrating multi-step aggregation capability within retrieved table data.

The Category III query (Q7) revealed two distinct error types. First, a section reference error: the system cited Section 4.3.5 for SGTR, which corresponds to SGHR (Steam Generator Header Rupture) in the SMR PSA report, rather than the correct Section 4.3.3 (SGTR). This confusion of similar acronyms and adjacent section numbers constitutes a retrieval attribution error that could direct a reviewer to the wrong section. Second, a content inaccuracy: the system stated that 'SLOCA1 has no radioactive materials in the RCS,' whereas RCS always contains radioactive coolant; the correct distinction is that SLOCA releases are confined within the containment, whereas SGTR can bypass containment via the secondary system. This nuanced factual error is potentially safety-relevant and underscores the need for expert review of complex comparative responses.

Overall, six of seven queries achieved accurate or near-accurate results (one dimension rated △ at most), confirming the system's practical utility for PSA document exploration. The single identified failure mode is systematic and addressable: domain-specific disambiguation rules can reduce SGTR/SGHR-type confusions at retrieval time. Integrating such improvements, along with explicit uncertainty signaling when retrieved context contains potentially ambiguous information, is the primary direction for future work.

## 4. Conclusions

In this study, a locally deployable RAG-based QA framework for PSA document exploration was proposed and demonstrated using the SMR Level 1 PSA report.

The system integrates hybrid retrieval (BM25 and dense semantic search with intfloat/multilingual-e5-base), preprocessing of PSA structural elements (ETs and tables converted to searchable text), and an evidence-grounded generation pipeline with explicit source attribution. Evaluation on seven representative queries across three information categories showed that the system reliably handles direct numerical retrieval and comparative multi-attribute queries, achieving accurate results on six of seven queries. Single systematic failure modes were identified: Section/acronym disambiguation errors, where SGTR (Section 4.3.3) was confused with SGHR (Section 4.3.5) together with an inaccurate radioactive material description. Such failure is safety-relevant and addressable through enhanced table-type metadata and acronym disambiguation at retrieval time. The locally deployable architecture satisfies confidentiality requirements for proprietary PSA documents, and the evidence-grounded response design provides a foundation for reliable, traceable QA assistance in nuclear safety assessment.

## ACKNOWLEDGEMENT

This research was supported by National Research Council of Science & Technology (NST) grant by the Korea government (MNIST) (No. GTL24031-000).

## REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kueettler, M. Lewis, W.-t. Yih, T. Rocktaeschel, S. Riedel, and D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 9459–9474, 2020.

[2] S. E. Robertson and H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond,Foundations and Trends in Information Retrieval, Vol. 3, No. 4, pp. 333–389, 2009.

[3] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 758–759, 2009.

[4] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, Multilingual E5 Text Embeddings: A Technical Report,arXiv preprint, arXiv:2402.05672, 2024.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, and others, Llama 2: Open Foundation and Fine-Tuned Chat Models,arXiv preprint, arXiv:2307.09288, 2023.

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,arXiv preprint, arXiv:2311.05232, 2023.

[7] W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl,Fault Tree Handbook, NUREG-0492, U.S. Nuclear Regulatory Commission, Washington, DC, 1981.

[8] U.S. Nuclear Regulatory Commission,A Guide to the Performance of Probabilistic Risk Assessments for Nuclear Power Plants, NUREG/CR-2300, Washington, DC, 1983.

[9] U.S. Nuclear Regulatory Commission,Domestic Estimated Frequencies for Initiating Events at U.S. Nuclear Power Plants: Results of the 2010 Survey, NUREG/CR-6928, Washington, DC, 2012.