INL/RPT-26-90824
Revision 0
DATE: 02/27/26

Evaluation of AI-Enabled Digital Documented Safety Analysis: A Case Study

February 2026

Generative AI for Nuclear Safety Basis Document Generation within NRIC's Digital Ecosystem

Sonali Sinha Roy, Harleen Kaur Sandhu, Nancy Gomez Dominguez, Kevin O'Rear, Drew Rizk, Paul Schuck, Peter Suyderhoud, Jacob Rymer, and Matthew Lund
Idaho National Laboratory

INL is a. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC.

NRIC National Reactor Innovation Center

DISCLAIMER

REVISION LOG

Revision No.	Date	Affected Pages	Description
0	02/27/26	All	New Issue

ABSTRACT, SUMMARY, AND FOREWORD

The traditional approach for developing safety basis and licensing documentation in the nuclear industry relies heavily on static, document-based processes that are time-consuming and lack integration across engineering disciplines. Leveraging digital engineering and generative artificial intelligence (GenAI) offers a promising alternative by enabling automation, improving data traceability, and strengthening alignment between evolving design information and safety documentation.

This report presents a case study evaluating AI-assisted generation of a Documented Safety Analysis (DSA) chapter within the National Reactor Innovation Center's (NRIC's) digital ecosystem. The project successfully demonstrated integration of a structured generative AI document development tool with the NRIC digital thread, enabling retrieval and synthesis of authoritative engineering data from a single source of truth. The evaluation was conducted within the constraints of the U.S. Government Azure cloud environment, where model availability and software maturity limited technical performance. Despite these constraints, subject matter expert (SME) review concluded that the AI-generated chapter provided an acceptable technical first draft and is expected to reduce document development effort by approximately 50 percent compared to traditional manual authoring.

While opportunities remain to improve technical precision, traceability, and formatting consistency through model upgrades and embedding enhancements, the results validate both the feasibility and the strategic value of AI-enabled, digital-thread-aware safety documentation. This effort establishes a foundation for continued maturation, systematic validation, and structured early adoption of AI-assisted workflows to accelerate advanced nuclear technology deployment.

ACKNOWLEDGEMENTS

This document was sponsored by the National Reactor Innovation Center (NRIC). NRIC is a national program funded by U.S. Department of Energy's Office of Nuclear Energy and is dedicated to the demonstration and deployment of advanced nuclear energy. Neither the U.S. Government nor any agency thereof makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe on privately owned rights. References herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise do not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

CONTENTS

FIGURES

TABLES

ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
ASME	American Society of Mechanical Engineers
BPV	Boiler and Pressure Vessel
DSA	Documented Safety Analysis
DOE	Department of Energy
DOME	Demonstration of Microreactor Experiments
ECAR	Engineering Calculations and Analysis Report
ECCS	Emergency Core Cooling Systems
FAA	Federal Aviation Administration
FDA	Food and Drug Administration
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
HTGR	High-Temperature Gas Reactor
INL	Idaho National Laboratory
LLM	Large Language Model
MCP	Model Context Protocol
NEI	Nuclear Energy Institute
NRC	Nuclear Regulatory Commission
NRIC	National Reactor Innovation Center
OCR	Optical Character Recognition
PDSA	Preliminary Documented Safety Analysis
RAG	Retrieval-Augmented Generation
RPS	Reactor Protection System
SME	Subject Matter Expert
U.S.	United States

Evaluation of AI-Enabled Digital Documented Safety Analysis

1. INTRODUCTION

1.1 Background and Motivation

Safety basis documentation development and review under U.S. Department of Energy (DOE) authorization have emerged as critical constraint throttling deployment of advanced nuclear reactors, with traditional processes demanding extraordinary resource investment that delays the delivery of these technologies. Traditional Documented Safety Analysis (DSA) processes rely on static documents with limited traceability [U.S. DOE]. The volume and complexity of information within nuclear licensing applications or authorization reviews demands innovative approaches to document generation and data management.

Recognizing these challenges as fundamental barriers to the commercial deployment of advanced reactors, the National Reactor Innovation Center (NRIC) has led the development of a transformative digital engineering ecosystem that fundamentally reimagines how safety analysis and licensing documentation are created and maintained [Lund et al. 2024]. NRIC's strategy employs a disciplined systems engineering approach to establish a digital thread—a single source of truth that maintains traceability and consistency across all design and safety information. The NRIC-developed digital ecosystem integrates safety analysis and licensing documentation into a dynamic, data-driven framework that evolves with the reactor design, replacing static documentation with an interconnected digital infrastructure. NRIC's broader vision for this digital ecosystem includes future development of artificial intelligence (AI)-enabled graphical user interfaces (GUI) that would enable DOE reviewers to directly access the digital thread and query information to validate consistency across engineering documents and analyses, potentially streamlining the authorization review process. By leveraging resources from across the Idaho National Laboratory (INL), including expertise from the Nuclear Safety and Assurance Department, Autonomous Engineering within the Scientific Computing and AI directorate, and the Materials and Fuels Complex (MFC) Reactor Engineering department, NRIC is well positioned to deliver this innovative capability to accelerate advanced nuclear technology deployment.

The integration of digital engineering and AI technologies has demonstrated substantial benefits in reducing project risk and improving efficiency across industries. For instance, digital twin technologies have been shown to reduce schedule delays by approximately 20% [Ritter and Rhodes 2023], and GE Vernova reported $1.6 billion in cost savings through digital engineering initiatives [Swartzendruber et al. 2025]. AI-powered document processing tools complement these advances, outperforming traditional optical character recognition (OCR) by understanding context and relationships within text and enabling intelligent automation of complex documentation tasks [Patel 2025]. Within the nuclear sector specifically, Mandelli et al. [2023] demonstrated that natural language processing (NLP) can extract structured data from unstructured equipment reliability reports to support nuclear safety analysis.

As the next phase in advancing its digital ecosystem capability, NRIC is evaluating AI-enabled document generation to further accelerate the delivery of advanced nuclear technologies. While NRIC's digital ecosystem is operational with a manually maintained digital thread, ongoing efforts continue to evolve the system through increased automation. Current DOE authorization requirements necessitate generating physical safety basis documents. This paper proposes a methodology for evaluating both the accuracy and efficiency of AI-powered document generation through targeted case study. The case study will assess the feasibility of a generative AI tool to create documentation sourced from the NRIC digital ecosystem, comparing AI-generated safety basis documents against what INL nuclear safety analysts would consider for human-generated documents, both in technical quality and resources. Leveraging INL's strategic partnership with Microsoft, NRIC utilized Microsoft's *Generative AI for Permitting Solution Accelerator* for the initial digital tool evaluation, while maintaining a technology-agnostic approach that will ultimately utilize the best available AI tool compatible with the NRIC digital ecosystem.

The potential benefits of AI-assisted documentation for nuclear safety applications within NRIC's digital framework include enhanced efficiency, improved adaptability, and reduced cost. NRIC's digital thread creates unique advantages for AI integration: AI tools can query the ecosystem to analyze information for consistency, identify all locations where specific design elements are referenced, and provide comprehensive information. However, given the safety critical nature of nuclear applications' rigorous credibility assessment is essential before deployment of AI tools. The current study addresses this need by initiating a systematic evaluation that validates the AI-generated safety basis documentation while comparing the effort required for AI-assisted generation against traditional human-led approaches. This work establishes a foundation for future integration of AI capabilities that leverage NRIC's single-source-of-truth architecture to support DOE authorization processes for advanced nuclear reactor projects.

This study represents the first successful demonstration of AI-enabled safety basis document generation directly linked to the NRIC digital thread. The effort validated that structured generative AI tools can access authoritative project data within INL's secure government cloud environment and produce a technically reviewable draft chapter. Although constrained by model limitations and pre-release software maturity, the outcome provides early evidence—supported by SME assessment—that AI-assisted workflows can reduce DSA development effort by approximately 50 percent while maintaining acceptable technical quality for refinement.

1.2 Scope and Objectives

The objectives of the current effort are:

- Evaluate Retrieval-Augmented Generation (RAG) systems credibility for producing nuclear safety documentation
- Identify key variables affecting output quality
- Determine actionable insights for potential adoption, validation, and risk mitigation.

The current report is limited to the AI-augmented generation of one chapter of a safety basis document based on a single test case, as described in Section 1.3. Section 3 provides the team's insight and guidance for wider adoption of AI-enabled safety basis document generation in the nuclear industry.

1.3 Description of Case Study

To evaluate the viability of using Generative AI tools to assist with writing nuclear safety documents, the team proposes a structured case study aimed at generating the Preliminary Documented Safety Analysis (PDSA) for NRIC's Demonstration of Microreactor Experiments (DOME) test facility. The current effort focuses on generating a single chapter from the Nuclear Energy Institute's NEI-21-07 format: Chapter 1, General Plant and Site Description and Overview of the Safety Case.

DOME, developed and constructed by NRIC at INL, is repurposed from the containment structure of the Experimental Breeder Reactor-II (EBR-II) and is designed to host experimental reactors up to 20 megawatts thermal, which will enable industry testing of advanced reactor concepts. For this case study, DOME is assumed to host a hypothetical, generic high-temperature gas reactor (HTGR) designed by NRIC.

A PDSA is the preliminary safety basis for DOE facilities associated with the as-designed configuration. It describes the facility, its safety structures, systems, and components, provides hazard and accident analyses, and details the safety management plan. A PDSA is followed by a more comprehensive DSA prior to turning over the as-constructed facility to operations.

The case study includes three evaluation components: a comparison to other AI models, a parametric evaluation, and subject matter expert (SME) review. Each of these components utilizes Microsoft's Generative AI for Permitting Solution Accelerator (described in Section 1.4) to generate sections of a PDSA using DOME and HTGR information. The components evaluate the efficacy of this generative AI tool for accelerating the development of a PDSA and the performance of the specific software deployment.

1.4 Description of Document Generation Tool

This case study explores using Microsoft's *Generative AI for Permitting Solution Accelerator* to fill the generative AI role. This software is a pre-release build capable of being developed into a bespoke solution to fit an organization's needs. In its original format, the tool can be used as a platform for defining and generating documents of any type. The software allows users to define a document format, known as a process, link associated reference libraries, and chat with AI assistants about a specific process or generated result. Figure 1 shows the home page of the deployed Microsoft *Generative AI for Permitting Solution Accelerator*.

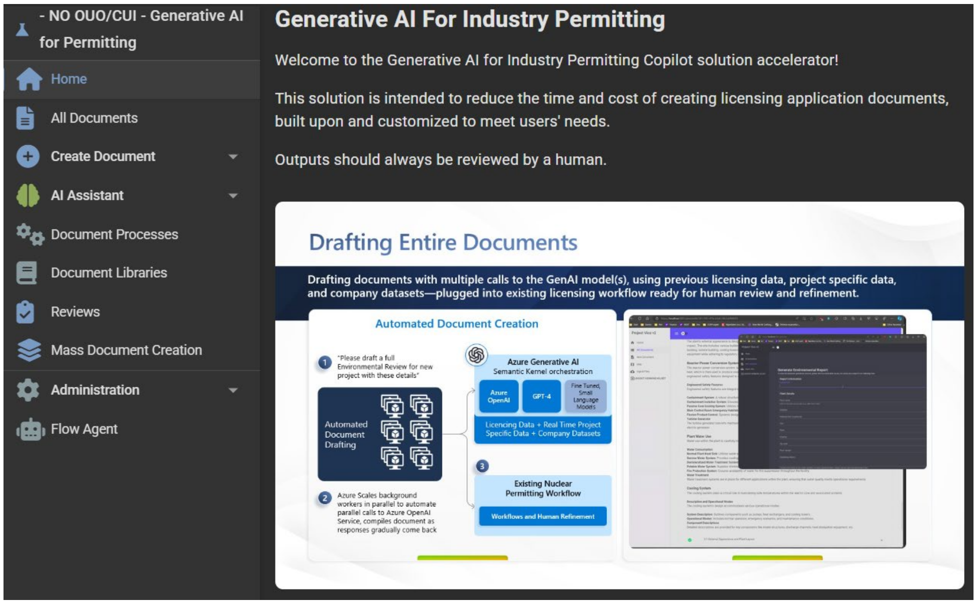

Figure 1. INL Deployment of the Solution Accelerator.

The tool provides users with the ability to generate a specifically formatted document of their choice, called a document process. Multiple processes can be defined within the tool, allowing for any number of individual documents and topics to be addressed through a single deployment. The tool also allows creation of various reference libraries for uploading information that can be used by the large language model (LLM) to generate content or inform document process structure. These document libraries may be linked to any of the document processes created for reuse of resources across multiple efforts from a single document upload. These libraries may be hosted on automatically generated blob storage within the deployment location (in this instance, the INL Azure Gov tenant).

The software utilizes an agentic approach to content generation with orchestrator, research, content generation, and review agents. The number of each of these agents can be scaled depending on the available token limits for the user's Azure subscription.

The permitting tool comes with commercially available large language models (LLMs) Generative Pre-trained Transformer (gpt)-4o for text generation and reasoning and text-embedding-ada-002 for embeddings. Models may be interchangeable from available options on the Azure OpenAI Service/Microsoft Foundry; however, this evaluation did not deviate from the aforementioned models.

The permitting tool was deployed within INL's Azure Government tenant with the assistance of Microsoft and used with NRIC's DOME and HTGR information. This deployment is specifically hosted on the *usgovvirginia* tenant which currently (as of February 2026) only has approval to host deployments for the **gpt-4 series** generative AI (GenAI) models and **text embedding 2 and 3 series** models. Azure government has set a retirement date of March 31, 2026, for the gpt-4o model used in this evaluation. A more modern replacement model of gpt-5.1 will be required in future deployments.

1.4.1 Tool Parameters

The following parameters/settings in the Microsoft Solution Accelerator are available for users to modify:

1. AI model configuration:
   a. RAG model (only **gpt-4o** permitted for INL's deployment)
   b. Validation model (only **text-embedding-ada-002** permitted for INL's deployment)
   c. Completion service: Agentic or Generic
2. Chunking mode: Simple or Semantic
3. Chunk size: Number of tokens
4. Overlap: Number of tokens
5. Embedding model
6. Embedding dimensions: 256, 512, 1024, 1536, 3072
7. Document retrieval settings
   - Preceding chunks before matching chunk
   - Following chunks after matching chunk
   - Maximum source documents to retrieve
   - Maximum chunks per source document
   - Minimum match relevance (%).

1.4.2 User Prompts

In addition to settings listed above, each document process created in the tool allows users to configure the specific prompts and outline associated with their process. There are two main types of prompts for a given document process: system generated and user generated. System generated prompts are automatically developed upon creation of a document process. These prompts apply to all sections of a document process, making them important for tuning the expected content within the entire document. Individual section prompts are user defined and only apply to the specific section in which the user writes them. The most relevant prompt types are described below.

1.4.2.1 AGENT ORCHESTRATOR AND AGENTIC PROMPTS

Agent orchestrator and agentic prompts are system generated upon creation of a document process. They provide the overall instructions for the use of the underlying agents, which include content generation, research, review, chat, and orchestration. The agent prompts decompose tasks in a document process and assign them to a specific LLM agent. The orchestrator agent is responsible for coordinating the efforts between these agents.

1.4.2.2 SECTION GENERATION PROMPTS

Similarly to the agentic prompts, section generation prompts are generated upon document process creation. Section generation is accomplished through a progressive multi-pass approach. To accomplish this, the content generation agent is given three section generation prompts: section summary, section generation main, and section generation multi-pass continuation. In combination, these prompts instruct the content generation agent on how to develop each individual section in the outline.

1.4.2.3 INDIVIDUAL SECTION PROMPTS

Individual section prompts are optional prompts created manually by the user to generate individual sections. With these prompts, the user has maximum flexibility and can vary the level of detail, use a variety of prompting techniques, and decide how to specify the reference files relevant to the section being generated. Section prompts are where the user defines the output for the document process.

1.4.3 Document Generation Workflow

The team used NEI-21-07 to provide the document outline and content structure for the PDSA within the tool. A document process (i.e., generic outline) was created to match the outline provided in NEI-21-07. The individual sections were populated using prompts created from NEI guidance and from expected content communicated from SMEs to the project team. Settings for the document process were tuned, and appropriate reference documents were linked. Generally, once the initial setup for the process is performed, the content may be generated repeatedly with minimal user input. Figure 2 shows the workflow for generating a document in the Microsoft Solution Accelerator.

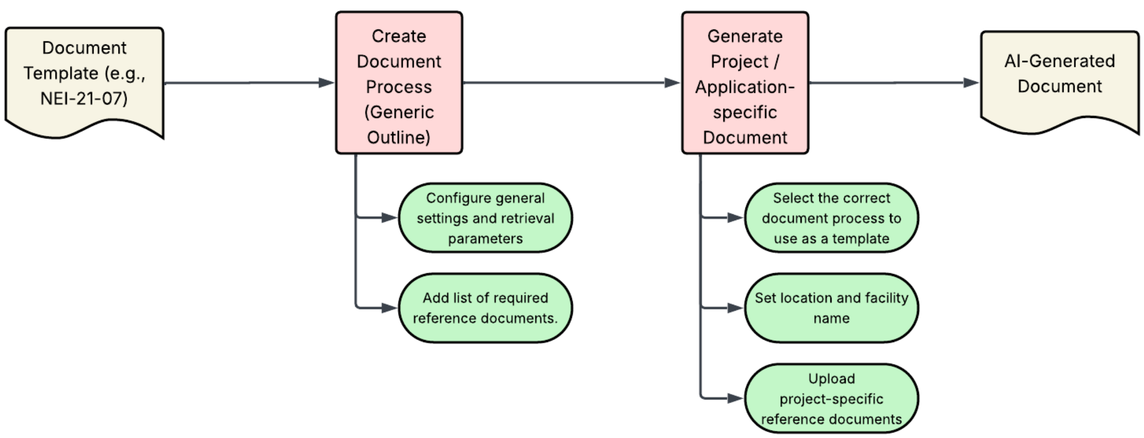

Figure 2. Document Generation Workflow in the Microsoft Solution Accelerator.

In this phase of the project, the source documents were uploaded through the Microsoft Solution Accelerator's user interface to an object storage location within INL's Microsoft Azure GovCloud tenant. The tool vectorizes the documents to create a knowledge base for RAG, which is then utilized to generate sections of the PDSA in response to user prompts. The prompts help the tool identify the desired content for the section and which source document to use for generating the response.

1.5 Integration with NRIC Digital Ecosystem

The NRIC ecosystem consists of interconnected data stores and software packages to enable enhanced collection and communication of project data from a single source of truth, automation of routine tasks (including document creation), change management and decision making, risk mitigation, and reduced development time [Lund et al. 2024]. The ecosystem focuses on enabling links between the original information developed on a project, known as the source of truth, and any software or documents that consume that information thus improving the flow of data, safety analyses, engineering requirements, and computer modeling [Darrington et al. 2020]. NRIC's ecosystem comprises open-source data adapters that move engineering data from disparate tool suites in and out of a centralized data hub: DeepLynx Nexus. Figure 3 shows a high-level overview of the NRIC digital ecosystem.

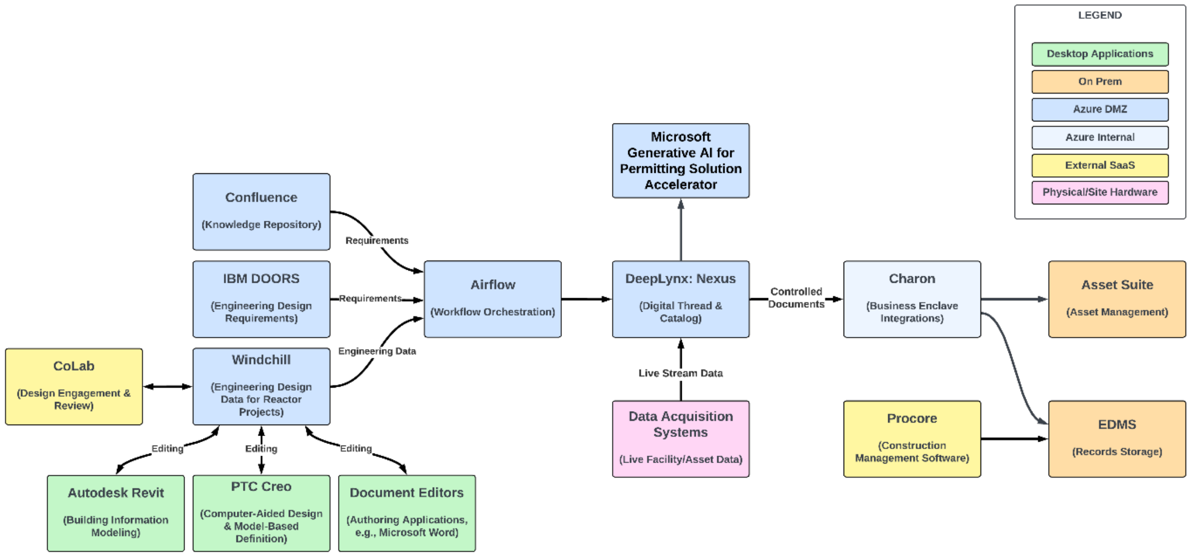

Figure 3. NRIC digital ecosystem.

The team's ongoing efforts focus on connecting Microsoft's tool to INL's enterprise databases to retrieve information from various types of dynamic data sources in addition to static documents. The progressive approach toward integrating this software into the NRIC digital ecosystem is outlined in Roy et al. (2025).

To test integration into NRIC's digital ecosystem [Lund et al. 2024], a Model Context Protocol (MCP) server was created, as shown in Figure 4, to provide an interface to the DeepLynx Nexus backend API, enabling interaction with Nexus data through natural language queries and agentic requests directly from the tool. The MCP server acts as an authenticated proxy, extracting the user's bearer token and forwarding it to the appropriate Nexus backend endpoint. The authentication middleware validates the forwarded token using OpenID Connect (OIDC) to verify the token signature with Microsoft's public keys, confirm the issuer is legitimate, and check the token's expiration. Upon successful validation, user identity claims (email, SSO ID, name) are extracted, and the user is either matched to an existing database record or provisioned as a new user. The MCP server does not generate its own tokens—it passes through the user's Entra credentials (from the Microsoft Solution Accelerator), allowing the Nexus backend to treat requests as authenticated application programming interface (API) calls made on behalf of real users. Figure 4 shows the flow of authentication claims via the Nexus MCP server.

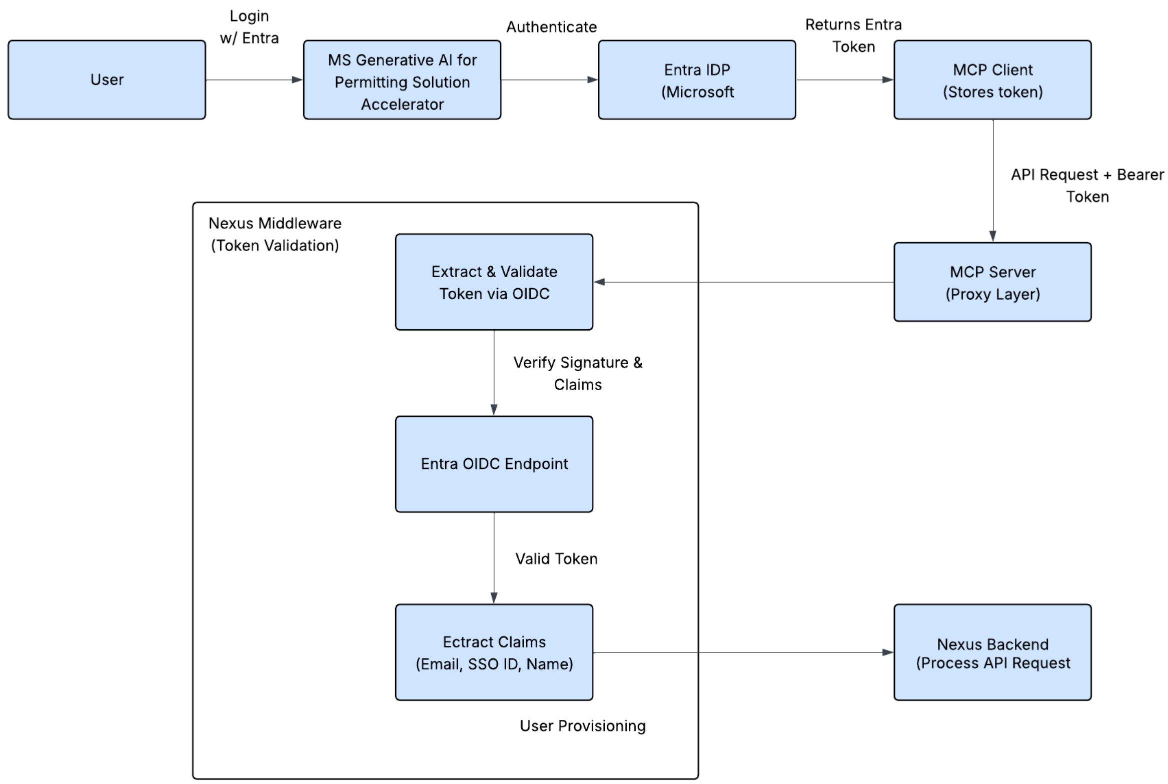

Figure 4. MCP flow via Entra Authentication.

This MCP server allows the Microsoft Solution Accelerator to access project records, which are housed within the INL DeepLynx Nexus server for use with document process generation. These records, taken from the various DOME and HTGR system design description requirement sets, provide detailed context for document generation. This MCP plugin is accompanied by changes to the research agent prompts within the document process to ensure that the research agent accessed DeepLynx data prior to determining that needed information was unavailable. Figure *5* shows the Microsoft Solution Accelerator UI accessing a selection of data from the NRIC DeepLynx Nexus project area.

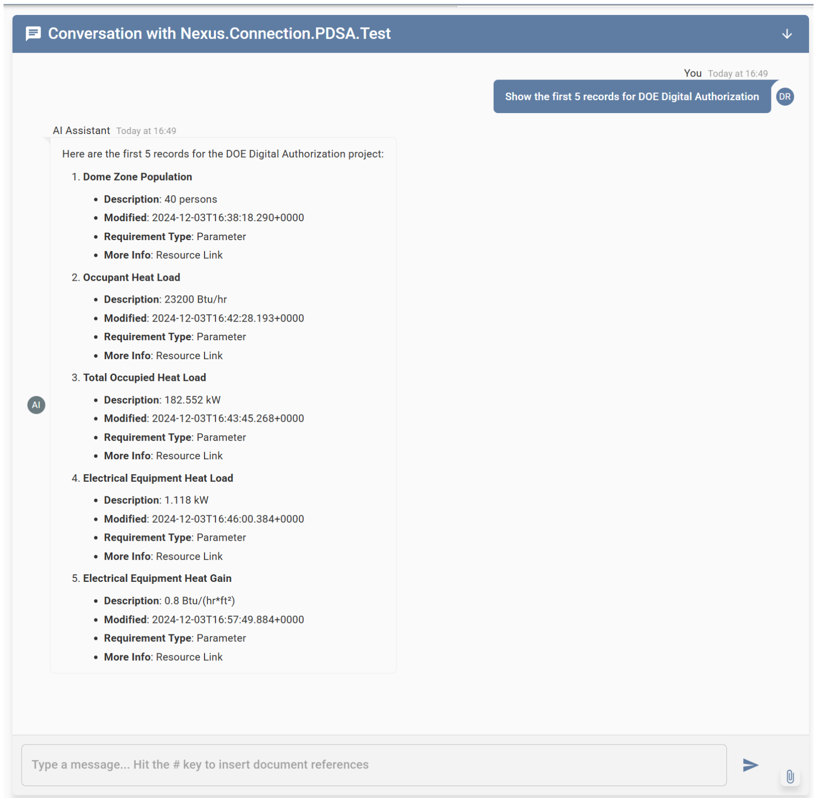

Figure 5. NRIC digital ecosystem integration.

2. DOCUMENT GENERATION AND EVALUATION

2.1 Document Generation

To facilitate the completion of the case study within the applicable time and budget constraints, the target document format and list of references were fixed in advance.

- Document template: NEI-21-07 Chapter 1
  - This is the new recommended template for future PDSA submissions to DOE.
- Reference corpus:
  - DOME system design descriptions (SDDs)
    - Containment building and isolation system
    - Utility systems
    - Containment ventilation and heat removal systems
  - Engineering Calculations and Analysis Reports (ECARs)
    - Hazards evaluation
    - System classification
    - Design and analysis
    - Failure modes and effects analysis
  - Generic HTGR concept of operations

- Generic HTGR Reactor Core, Cooling System, and Instrumentation and Control system design descriptions
- INL standardized safety analysis report
- NEI-21-07 and NEI-18-04 guidelines.

The evaluation team used the NEI-21-07 format to create document outlines in the Microsoft Generative AI for Permitting Solution Accelerator. This format guided the prompts used to generate content for each section of the DOME PDSA. Data sources and corresponding reference documents for each section were identified through consultations with the nuclear safety analysts and design engineers. A single chapter (Chapter 1: General Plant and Site Description and Overview of the Safety Case) was selected for generation and evaluation. Additionally, Section 1.1.4.4 was selected for detailed analysis such as comparison with other AI tools and impact assessment of retrieval parameters.

2.2 Evaluation of the Tool

GenAI is inherently probabilistic, making it difficult to evaluate AI-generated content. Additionally, GenAI comes with known risks like hallucinations and misinterpretation of context. These challenges make AI adoption in safety-critical domains like nuclear energy extremely difficult. Tools like Microsoft's *Generative AI for Permitting Solution Accelerator* have the potential to optimize licensing and regulatory workflows, which directly translates to time and cost savings. However, a thorough evaluation of the reliability and trustworthiness of these tools will be required before they can be adopted for real-world applications. "AI for nuclear safety" is a nascent and evolving market. Our goal with this study is to better understand the nuances, identify the advantages and shortcomings, and initiate a path toward well-rounded validation frameworks. To that end, a three-step evaluation was conducted, as illustrated in Figure *6*.

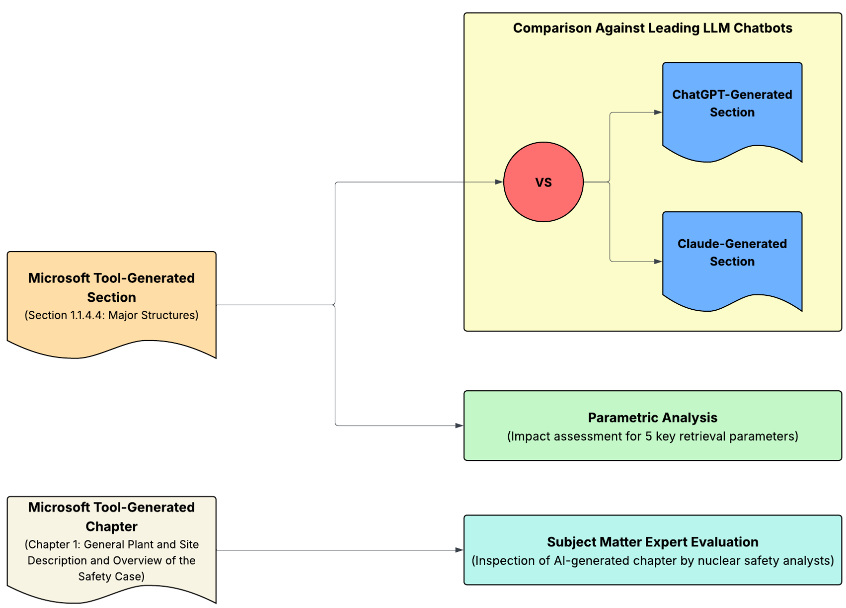

Figure 6. Evaluation of the AI tool.

2.2.1 Comparison Against Leading LLM Chatbots

To evaluate the credibility and utility of the Microsoft AI Generative AI for Permitting Solution Accelerator, it was compared against two traditional LLM chatbots, OpenAI's ChatGPT and Anthropic's Claude. The section of the targeted chapter (Section 1.1.4.4 "Major Structures" of the NEI-21-07 format) was generated using the same prompt format (shown in Figure 7) using the Microsoft Solution Accelerator, ChatGPT, and Claude. The Microsoft Solution Accelerator was configured with gpt-4o as the underlying LLM for this case study. For Claude, Sonnet 4.6 was used and for ChatGPT, gpt-5.2 was selected. The prompt format for section generation is shown below, as derived from the NEI-21-07 format.

The three models used in this evaluation represent different generations of LLM development, as shown in Table 1. Chat GPT's gpt-4o model, released in May 2024, features a 128,000 token context window with a knowledge cutoff of October 2023, making it the oldest and most constrained of the three in terms of both knowledge cutoff date and context capacity. The LLM model gpt-5.2, released December 11, 2025, offers a substantially larger 400,000 token context window and a knowledge cutoff of August 31, 2025, representing a significant generational advancement. Claude Sonnet 4.6, released February 17, 2026, supports a 200,000 token context window, and delivers improvements across coding, long-context reasoning, agent planning, and knowledge work [Artificial Analysis, 2025].

Table 1. Comparison of LLM models.

Attribute	gpt-4o	gpt-5.2	Claude Sonnet 4.6
Developer	OpenAI	OpenAI	Anthropic
Release Date	May 2024	December 2025	February 17, 2026
Context Window	128K tokens	400K tokens	200K tokens
Knowledge Cutoff	October 2023	August 2025	August 2025
Reasoning	Standard	Tiered (Instant/Thinking/Pro)	Adaptive thinking

1.1.4.4. **Major Structures**

Write a concise technical section on the general performance characteristics and description of major features such as reactor building, auxiliary, secondary and support buildings, cooling towers/systems and co-located facilities (e.g., cogeneration, fuel processing and buildings).

Use the following documents as the source of the information: Generic HTGR ConOps section 4.6 through section 4.7, Reactor Core RC SDD section 4.1 through section 4.2, Cooling System CS SDD section 4.2 and Instrumentation and Control IC SDD section 4.1, DOME Containment SDD, ECAR 7264, ECAR6327

Figure 7. Prompt used for section generation in the Microsoft Solution Accelerator and chatbots.

A comparative technical evaluation was performed to assess the effectiveness of the three LLM candidates. The evaluation criteria included factual correctness, units and technical semantics, completeness and comprehensiveness, and document traceability and citation. Overall performance ranked as follows: Claude (Sonnet 4.6) demonstrated the highest overall performance across all evaluated metrics, followed by ChatGPT (gpt-5.2), with the Microsoft Solution Accelerator (gpt-4o) producing less reliable results. Table 2 shows the scores for each candidate for the key AI quality metrics, and an example for each metric is mentioned below.

1. Factual correctness: The original design document mentions the integrated leak rate limit as 5,500 cc/s by pressurizing the containment building to 10 psig. Claude and ChatGPT mentioned this accurately as, "*The containment structure shall maintain a total leak rate of less than 5,500 cc/sec at 10 psig when fully isolated*," and "*Integrated leak rate testing limit ≤ 5,500 cc/sec at 10 psig*," respectively. However, the Microsoft AI tool mentioned "...*ensure leak-tight performance with total leakage rates below 10,500 cc/sec at 10 psig*." This is nearly double the correct value and unsupported by any source document.
2. Technical semantics: In the subsection describing the reactor protection system (RPS), Claude detailed the voting logic, trip parameters, and other technical information as, "*employs {2-out-of-4} voting logic across independent channels, monitoring trip parameters including high power, high outlet temperature, high and low pressure, startup rate anomalies, and loss of primary or secondary coolant flow. Upon trip, the RPS actuates control drums and the central shutdown rod. Three independent and isolated channels (A, B, and C) are housed in separate racks within the IC CONEX*." However, ChatGPT did not mention the RPS at all. The Microsoft AI tool described the RPS at a functional level only, lacking design-specific description as, "*Monitors critical parameters such as core temperature and neutron flux; automatically triggers SCRAM functions during abnormal conditions*."
3. Completeness and comprehensiveness: In the subsection about the cooling system, Claude was able to describe the complete four-subsystem cooling architecture (including the primary cooling system, the intermediate heat exchanger, the secondary cooling system, and the cooling tower system) with specific parameters for each. ChatGPT was able to identify the cooling architecture at a high level but omitted specific parameters and brief descriptions. The Microsoft AI tool limited its description of the cooling system to generic statements without any subsystem breakdown and no quantitative parameters.
4. Document traceability and citation: Claude cited specific document identifiers and section numbers inline for every piece of technical information retrieved (e.g., "*Ref. 5, Section 3.6.1*"). ChatGPT referenced source documents at the document level in a formal reference list but did not link individual technical claims to specific sections inline. The Microsoft AI Tool mentioned document identifiers inline within the text (e.g., "...*as described in ECAR6327*"), but provided no consolidated reference list.

Table 2. Quality evaluation of AI-generated results.

General Quality Metric	Microsoft Solution Accelerator	ChatGPT	Claude
Factual correctness	2	5	5
Technical semantics	2	3	5
Completeness and comprehensiveness	3	3	5
Document traceability and citation	3	2	5
Average Score	2.5	3.3	5
Scoring rubric: 1 = Poor / significant errors or omissions; 2 = Below expectations / notable gaps; 3 = Adequate / meets minimum expectations; 4 = Good / minor gaps; 5 = Excellent / fully meets expectations with no significant deficiencies.			

While the Microsoft AI tool scored lower on the evaluation metrics shown in Table 2, it offers several workflow and infrastructure advantages over the standalone LLM chatbots (ChatGPT and Claude) including:

1. Structured user interface custom-built for document generation, with configurable parameters such as chunk window sizes and retrieval settings.
2. Larger document database capacity, enabling retrieval across an entire project document set rather than being limited by session context.
3. Document memory retained across sessions, maintaining continuity of the document being generated.
4. Multi-user collaboration, allowing multiple team members to work on the same document simultaneously.
5. Larger token windows for processing extensive engineering document sets.
6. Ease of usability through a guided interface, reducing the technical barrier for non-expert users.

The primary limitation for the Microsoft AI tool identified in this evaluation is the use of gpt-4o as the underlying LLM. Once the Microsoft AI Tool pushes the scheduled LLM upgrade to gpt-5.1, a more equitable performance comparison across all three candidates will be possible.

2.2.2 Parametric Analysis

The Microsoft AI tool has several parameters that can be modified by users to tune the quality of the AI-generated text. Therefore, an assessment was conducted to identify how key retrieval parameters impact the tool's output. The methodology is inspired by the classic sensitivity analysis approach. However, due to the stochastic nature of AI-generated content, it is difficult to conduct a quantitative assessment. Therefore, a qualitative evaluation is presented, focusing on the observed patterns and interpreted trends.

This assessment was conducted for a single section of the target chapter. NEI-21-07 section 1.1.4.4 *Major Structures* was selected because it has an abundance of factual technical information and uses data from multiple source documents. This section provides an overview of the primary structural elements, such as:

- Reactor building
- Auxiliary, secondary, and support buildings
- Cooling towers/systems
- Co-located facilities (e.g., cogeneration, fuel processing, and buildings).

Five key retrieval parameters were selected for this analysis, and each one was assigned a baseline value, as listed below.

1. Chunk size: 1200 tokens
2. Chunk overlap: 180 tokens
3. Number of neighboring chunks: 1
4. Maximum number of chunks per document: 4
5. Minimum relevance: 80%.

First, the benchmark section was generated using baseline parameters. Then, the impact of each parameter was assessed by varying only one parameter while keeping all the other parameters constant at their respective baseline values. The following parameters were kept constant for this analysis:

- Maximum source documents: 20
- AI completion service: Agentic
- Chunking mode: Semantic.

The key observations for each parameter are summarized in the following subsections and summarized in Table 3.

2.2.2.1 CHUNK SIZE

Chunk size is the amount of source text grouped into each retrievable unit, typically measured in tokens. It determines how much context is contained in a single unit (chunk) that can be retrieved and fed to the LLM. For parametric analysis:

- Baseline: 1200 tokens
- Variations: 300, 600, 1800, and 2400.
- **Observations:**
  - Smaller chunks (300 tokens): Tends to be repetitive, reintroducing the section scope and then going deep on the first structure (i.e., reactor building) while under-covering the rest of the structures (cooling, auxiliary, co-located) that appear in the NEI-21-07 section outline.
  - Mid-size chunks (600 tokens): Produces the densest, most technical output, with noticeably more standards, IDs, and design-basis specifics (e.g., American Societ of Mechanical Engineers (ASME)/American Concrete Institute (ACI)/Nuclear Qaulaity Assurance (NQA)-1, ECAR references, limit states). This suggests chunking is helping retrieval stay “on-topic” and pack in high-specificity details rather than broad narrative.
  - Baseline (1200 tokens): Provides a balanced but broad coverage across all major structures, with consistent headings, but less hard-edged technical traceability than the 600-token variant.

  - Larger chunks (1800–2400 tokens): Trends toward narrative redundancy and repetition. The 2400-token sample repeats the reactor building section almost verbatim near the end, indicating that very large chunks can increase looping/duplication and reduce "new information per token."

2.2.2.2 CHUNK OVERLAP

Chunk overlap refers to the portion of text that is intentionally repeated between adjacent chunks, usually measured in tokens or as a percentage of chunk size. Overlap helps preserve continuity so key definitions or assumptions near a boundary are not lost. For the parametric analysis:

- Baseline: 180 tokens (15% of baseline chunk size)
- Variations: 0, 60, 120, 240, 360.
- **Observations:**
  - Zero overlap (0 tokens): Noticeable "dropout" as the text progresses. The output strongly covers the reactor building and then moves into auxiliary buildings, but it doesn't reliably reach later structures (cooling/containment/co-located) that are covered in the baseline sample.
  - Low overlap (60 tokens): Observable increase in specificity. In fact, in some instances, the text becomes too specific. For example, the reactor building section is highly detailed and standards-heavy (materials specs, wall thicknesses, Nuclear Regulatory Commission (NRC)/ASME references), suggesting overlap is helping carry over very specific retrieved facts from adjacent text in the reference documents.
  - Moderate overlap (120 tokens): Improved continuity into sub-subsections. For example, it flows more naturally from "Reactor Building" into a focused deep dive (e.g., "Containment Structure"), indicating overlap helps the model keep context across boundary transitions rather than restarting at each chunk.
  - Baseline overlap (180 tokens): Appears most "balanced." It gives a complete inventory of major structures and then proceeds through them in a structured manner (Reactor → Auxiliary → Cooling → Co-located) in alignment with the NEI-21-07 guidelines.
  - Higher overlap (240 tokens): Trends toward redundancy, with a few instances of repetitive explanation of the same concepts with less new content per paragraph.
  - Very high overlap (360 tokens): Seems to trigger outright repetition/looping. The 360-token variant repeats the opening "major structures" framing and overview near the end, which may be an indication of too much shared context across chunks.

2.2.2.3 NUMBER OF NEIGHBORING CHUNKS

This parameter defines how many chunks immediately before and/or after a matched (retrieved) chunk are also included as additional context. This expands local context around the hit to capture surrounding definitions, derivations, or conclusions. For parametric analysis:

- Baseline: 1
- Variations: 0, 2.

Because of the low number of data points, each variation was repeated three times. The insights are based on the average quality of the three samples for each variation.

- **Observations:**
  - Zero neighboring chunks: Shows high variance and more generic language. Across the three 0-neighbor attempts, the content frequently leans into broadly true nuclear-facility descriptions (materials, seismic resilience, generic systems) with fewer distinctive, source-tethered specifics.
  - One neighboring chunk (baseline): Best mix of completeness and concrete detail across three samples. The benchmark section provides full-structure coverage (Reactor → Auxiliary → Cooling → Containment → Co-located) and includes multiple impactful specifics (e.g., VFD-controlled cooling cells, explicit wind/snow load values, CONEX container mention).
  - Two neighboring chunks: Smoother coherence but tends to dilute specificity. The two-neighbor attempts read like an executive summary, but contain some speculative placeholders (e.g., “[*DETAIL: Dimensions to be confirmed from ECAR-7264*]”) and speculative language (e.g., “*Future sequences will incorporate quantitative data such as specific dimensions, capacities, and regulatory benchmarks extracted directly from ECAR documents and SDDs*.”), which indicate quality regression compared to the baseline.

2.2.2.4 MAXIMUM NUMBER OF CHUNKS PER DOCUMENT

This parameter puts a cap on how many chunks the retriever is allowed to take from any single document for a given query. It prevents one long document from dominating the retrieved context and encourages diversity of evidence (sources) across documents. For parametric analysis:

- Baseline: 4
- Variations: 1, 2, 6, 8.
- **Observations:**
  - Low max chunks (1): Stays coherent but trends generic. The generated text reads like a standard template description (broad features, few concrete parameters/IDs), suggesting not enough source material is being surfaced to inject distinctive details or descriptions.
  - Moderate max chunks (2): Shows some structure duplication. The output text includes a repeated “Introduction and Overview” later in the text, suggesting boundary effects or re-priming when fewer chunks are allowed and the model re-anchors itself.
  - Mid-range (4, baseline): Maintains full section coverage and includes concrete quantitative/implementation details (e.g., VFD-controlled cooling tower cells; wind/snow load numbers).
  - Large max chunks (6): Shows increased risk of self-referential metatext and repetition. This variant contains first-person commentary (e.g., “*I recommend focusing first on the reactor building, covering its design, safety features, and integration within the facility's operations*”), which is highly undesirable in a final document.
  - Very large max chunks (8): Incorporates details such as system decomposition. For example, the generated text explicitly lists and describes the functions/interfaces of subsystems like Primary Coolant System (PCS) and Intermediate Heat Exchanger (IHX) rather than only describing cooling towers at a high level. However, the text also includes placeholders like “[*Facility Name*],” which can be a side effect of pulling in more generic template-like context.

2.2.2.5 MINIMUM RELEVANCE

Minimum relevance is a cutoff score or threshold that a candidate chunk must meet to be included in retrieval results. Higher minimum relevance yields fewer, more tightly matched chunks; lower minimum relevance admits more chunks but can increase noise. For parametric analysis:

- Baseline: 80%
- Variation: 20%, 40%, 60% 100%.
- **Observations:**
  - Low minimum relevance (20%): Includes more “generic-but-plausible” nuclear content. The 20% sample expands into broadly applicable features (e.g., decontamination facilities, secondary control centers, N+1 redundancy) that are often true for nuclear applications but are not clearly anchored to the specific reference documents for this case study.
  - Low-medium relevance (40%): Appears more tailored to the specific application instance and cites concrete engineering artifacts/values (e.g., ASME BPVC, pressure ranges mentioned in the ECAR, SDC-3 limit states).
  - Mid-to-high relevance (60%–80%): Follows the expected outline. These consistently progress through Reactor → Auxiliary/Support → Cooling → Co-located, matching the intended NEI-21-07 section guideline.
  - Very high minimum relevance (100%): Appears overly “source-faithful” but loses some of the key details. For example, the 100% sample is coherent and system-structured yet omits some of the baseline’s distinctive specifics (e.g., the explicit wind/snow load numbers and mention of CONEX container), suggesting a possible tradeoff where strict filtering narrows what is eligible/acceptable and smooths output toward safer, more traceable phrasing.

Table 3. Observations from sensitivity analysis.

Parameter	Observation
Chunk size	Varying chunk size allows balancing of coverage vs. depth. Smaller chunks lead to good local coherence, but poor section completeness. Meanwhile, very large chunks result in a more inclusive narrative but with higher repetition risk and diluted signal.
Chunk overlap	Varying overlap creates a tradeoff between continuity and repetition. As overlap increases, the text is more likely to stay “on track” across chunk boundaries (i.e., better transitions and fewer abrupt resets), but it also becomes more likely to repeat itself.
Neighboring (preceding and following) chunks	Difficult to identify patterns with a limited dataset, but traceability is not seen to monotonically improve with more neighbors. Adding at least one neighbor appears beneficial for grounding and complete coverage.
Maximum chunks per document	Higher max chunks lead to more technical breadth but introduce an increased risk of “retrieval noise.” With more chunks available, the text can pull deeper details. However, with lower constraints, the generated output can lack specificity in certain instances.
Minimum relevance	As minimum relevance rises, specificity and traceability increase. Low relevance results in richer breadth and creativity, but a higher hallucination risk. High relevance provides stronger grounding and fewer “made-up” additions but can reduce “intelligent” connections that enrich the text with details.

2.2.3 Subject Matter Expert Evaluation

The technical quality of any AI-generated content is best evaluated by SMEs in that domain. Therefore, the PDSA chapter generated by the Microsoft AI tool was examined by nuclear safety analysts who are well-versed with the regulatory requirements, acceptance criteria, and writing safety basis documents. SMEs provided qualitative comments on the generated PDSA. This section contains a discussion of the comments provided as well as an interpretation of their implications on the Microsoft Solution Accelerator.

2.2.3.1 SME REVIEW COMMENTS AND NOTES

Comments received from the SME review were grouped into topic areas. A summary of the key SME findings and selected review comments is contained in the following subsections.

SCOPE AND LEVEL OF DETAIL

SME reviewers found that the generated documents consistently overelaborated or provided excess detail in sections meant to be high-level overviews. Repeated comments were made noting that content was too verbose and contained unnecessary “fluffy” language. Sections regularly contained a summary of the type of content that should be in the section rather than a summary of the content generated for that section. Select comments for this topic area are as follows:

- "*This section should provide a high-level methodology of how the safety case will be developed. It should not describe specific systems."*
- *"Should provide a high-level description of how the fundamental safety functions will be satisfied, only noting the systems that provide that function. Does not describe how the system does that."*

TECHNICAL ACCURACY AND APPLICABILITY

The narrative repeated knowledge that was not specific to the HTGR design under evaluation. In several instances, the model hallucinated by including information applicable to other reactor technologies, such as chemical shim systems (boric acid control) and Emergency Core Cooling Systems (ECCS), which are not credited design features of the generic HTGR evaluated in this project. SMEs noted occasional hallucinations where values were presented from typical HTGR designs rather than the generic HTGR used in this project.

Some critical facility-specific information was missing from the output, and few numerical parameters were provided. Additionally, the technical information presented in the sections was formatted entirely as text, whereas SMEs would expect tables and figures. Select comments for this topic area are as follows:

- *"600MWth isn't correct for this use and maybe from it generalizing HTGR's but not as microreactors."*
- *"The way it talks about modularity may not be applicable to this microreactor. It may be thinking of HTGRs in general here. Also talking about hydrogen production seems like unnecessary information."*
- *“The document lacks quantitative nuclear design parameters and fails to clearly delineate safety classification or system boundaries.”*

FORMATTING AND REDUNDANCY

The Microsoft Solution Accelerator results tended to largely reiterate chapter requirements within a section and repeat information across sections. Issues with section-to-section consistency were noted throughout the generated chapters. Frequently, information contained in one section was repeated within another. Acronyms were defined multiple times as if they were being used for the first time. Formatting and referencing of information and documents were inconsistent across sections and did not conform to a specific style. Bullet points, section numbering, and illustrations were noted as incorrect in multiple places. The output was noted to create subsections for generated content without any numbering. Select comments for this topic area are as follows:

- "This section may be redundant to the one above."
- "We've added these subsections without giving section numbers or differentiating them from regular text."

PROMPTS AND SOFTWARE ARTIFACTS

A pattern of AI artifacts appearing in the final document was noted by reviewers. Reviewers identified prompt instructions, self-referential commentary, and meta-discussion that should have been removed by AI agents prior to document finalization. The AI also exhibited technical limitations related to technical writing style, first-person statements, and consistency in terminology. Several instances of the conversations between AI agents were included in the final document. Select comments for this topic area are as follows:

- "Is this left over from a prompt or is this standard to see in this type of document. It seems unusual."
- "Probably shouldn’t use I statements in this document. It seems like it's commentating on its own work before finishing the section."

2.2.3.2 INTERPRETATION OF SME COMMENTS

SMEs noted that, while the generated content was incomplete, it did provide a starting point that outlined the document and content to accelerate development. SMEs estimated that the document provided for review could result in approximately 50% savings in time to generate the DSA compared to manual authoring. Additionally, these savings were accomplished by a project team that did not consist of nuclear safety specialists.

Issues related to the input prompts for each section were identified in the comments made via SME review. Prompts were not developed using SME input and likely did not capture the expected level of detail or key topics that SMEs expected. Refinement of the generated text via multi-shot prompting using the Microsoft Solution Accelerator’s included chat interface was not performed. Use of these post-generation tools could refine sections to remove the unnecessary verbiage noted by SMEs.

Reference documents provided to the Microsoft Solution Accelerator were limited due to the hypothetical nature of the reactor design. Information for required values such as effective delayed neutron fraction (β_eff), shutdown margin, prompt neutron lifetime, and peak-to-average power ratios were not included, as these values are not yet finalized in supporting engineering documentation. This lack of supporting information contributed to the missing critical design information that would be expected of a submission.

The embedding model, text-embedding-ada-002, lacked the ability to process images or tables, which also represents an area for improvement. Many of the reference documents contained images of system interfaces or tables of parameters that were noted by SMEs as missing from the final product. The inability to reuse or reproduce information taken from these tables and figures contributed to comments on the technical accuracy of the output.

Software issues with the Microsoft Solution Accelerator's deployment also contributed to SME comments. Limits on the number of turns the AI agents had to discuss content contributed to instances where AI agents were interrupted mid-conversation to produce the output document. Additional issues with the unavailability of the document validation pipeline had an impact on the Microsoft Solution Accelerator's ability to maintain section-to-section consistency and reduce redundancy.

Many of the issues with text style, depth of discussion, technical referencing, and language can be attributed to the gpt-4o LLM. Gpt-4o has limitations on the context windows and reasoning that were apparent in all AI document generations. As discussed in Section 2.2.1, the gpt-4o model underperforms compared to more contemporary peers such as gpt-5.2 or Sonnet-4.5. This performance is reflected in SME comments related to the reiteration of inputs and lack of technical knowledge. While the gpt-4o model hindered performance, the Microsoft Solution Accelerator does allow for use of other models hosted on the Microsoft Foundry. The next approved model for the U.S. government cloud will be gpt-5.1, which could enhance the output of the Microsoft Solution Accelerator without changing other parameters. However, the use of any of these AI models is estimated to reduce document generation time by more than 50% compared to fully manual authoring based on this gpt-4o case study.

3. CONSIDERATIONS FOR ADOPTION OF AI-GENERATED SAFETY DOCUMENTATION AND NEXT STEPS

3.1 Software Maturation

Version numbering of the Microsoft Solution Accelerator is not performed due to its in-development nature. The version of the Microsoft Solution Accelerator used in this evaluation is identified by its repository fork date of January 23, 2026. The continued development of this software is expected to further refine its capabilities to fix currently identified issues and provide better end results. It is acknowledged that this study utilizes a pre-release version of the accelerator and, as such, there are software issues that will be fixed in future releases.

A feature that would be beneficial to a multi-section generation tool is cross-section validation. The Microsoft Generative AI for Permitting Solution Accelerator contains provisions that can perform full document validation. However, this feature was not functioning at the time of this review. As larger document sizes are generated, a detailed document validation feature would be necessary to reduce redundancy and increase consistency of technical data.

3.2 Contemporary LLM and Embedding Model Use

The gpt-4o model is being deprecated on the Azure Government Cloud on March 31, 2026. The replacement model, gpt-5.1 (released November 2025), has higher token context windows to provide enhanced processing and reasoning of complex prompts and inputs. Upgrading to gpt-5.1 is anticipated to result in more detailed and comprehensive generations of regulatory document sections with less need for human post-processing and refinement based upon the findings in Section 2.2.1 with the similar gpt-5.2 model.

An additional area of AI model improvement is the embedding pipeline. The embedding model used, text-embedding-ada-002 (released December 2022), can only process text from reference documents. Utilizing a more advanced embedding model could improve the generated content's referencing. Additionally, an agentic approach to embedding, where text, figures, or tables are handled by individual specialized models, could improve the ability of the generative AI to accurately summarize the safety case using source information.

3.3 Nuclear Domain-specific Evaluation Metric Development

This case study evaluated the generation of safety basis information using SME evaluations and metrics developed within INL for the evaluation of human-generated safety documentation. This process is labor intensive and provides limited review objectivity. Additionally, it does not translate well to AI document generation due to the propensity for AI to provide variations in output with an identical input prompt. Development of a standard nuclear domain-specific metric for generated content would allow more accurate assessment of AI-generated nuclear documentation. The metric should address the prevalence of hallucinations, quality of technical writing content, and have the ability to accurately identify and link source material used as references.

It is recommended to examine how other federal regulatory agencies are approaching artificial intelligence governance to guide the development of AI evaluation metrics for nuclear licensing. The Federal Aviation Administration (FAA), for example, has published a roadmap for AI safety assurance, which outlines guiding principles for integrating AI into aviation systems [FAA]. Similarly, agencies such as the U.S. Food and Drug Administration (FDA), Occupational Safety and Health Administration (OSHA), and the National Highway Traffic Safety Administration (NHTSA), have begun developing frameworks or issuing guidance on the responsible use of AI in their respective domains. Including these agencies in a cross-agency review can provide valuable insights into best practices for transparency, validation, human oversight, and regulatory acceptance. Many of these principles may be directly applicable to the nuclear industry's adoption of AI-enabled tools.

3.4 Next Steps and Early Applications

This case study successfully demonstrated AI-enabled generation of a DSA chapter directly linked to NRIC's digital thread within the constraints of the U.S. Government cloud environment. With this technical foundation established, the program is positioned to transition from proof of concept to structured early adoption. The demonstrated ~50 percent reduction in document development effort—validated through SME review—provides a compelling basis for applying this capability to active projects utilizing the NRIC digital ecosystem.

These efforts will apply the capability to general technical document development, DSA drafting, and structured updates informed by authoritative digital-thread data. By enabling AI assisted synthesis of requirements, hazards, controls, and design artifacts, the tool strengthens alignment between evolving design baselines and safety documentation. This early adoption approach allows NRIC to mature the capability in context, where configuration management, engineering development, and safety basis documentation intersect, while reinforcing traceability and technical consistency.

The NRIC Laboratory for Operations and Testing in the United States (LOTUS) test bed will serve as the first direct NRIC application of this technology. For LOTUS, the AI integration will be leveraged, to the maximum extent practicable, for technical document development, safety basis document generation, and rapid verification queries against the digital thread. This application is intended both to realize projected efficiency gains, up to 50 percent reduction in document development effort, and to provide a real-world demonstration of AI enabled safety documentation within an active nuclear project environment.

Through targeted early adoption and immediate application on LOTUS, NRIC will demonstrate how AI supported, digital-thread-aware workflows can enhance efficiency, consistency, and technical rigor in advanced nuclear technology deployment.

4. SUMMARY AND CONCLUSION

4.1 Summary of Case Study

While there were shortcomings associated with the Microsoft Generative AI for Permitting Solution Accelerator and gpt-4o, the use of the Microsoft Solution Accelerator allowed a multiplication of effort for project team members. There were some instances of hallucinations of various design details as well as areas where the LLM identified the need for human intervention. Gpt-4o's performance was seen as contributing to many of the SME comments received. While this represents a bottleneck, the impending retirement of gpt-4o and the shift to a more contemporary model provides confidence that the Microsoft Solution Accelerator will have better results in future use cases. However, SMEs identified that the generated output would cut their level of effort in half, so effort could be focused on refining and editing the generated content to meet safety documentation needs.

The section-by-section prompt strategy implemented by the Microsoft Solution Accelerator allows many advantages compared to a traditional chat style interface for generative AI. Isolating sections from one another allows the agents to focus on developing larger and more specific information sets for each section without the user having to provide significant back and forth. Defining a document process and initial prompt package as a saved template enables rapid recreation of documents with minimal input from users. This document template approach allows an organization to take advantage of economies of scale for use of the product on subsequent document generations.

4.2 Limitations and Challenges

This case study is limited by the technology and documentation available. The following caveats should be noted when reviewing the results of this study:

- The software used is not a full commercial release software. Some features of the software were unavailable at the time of testing. Additionally, limited documentation was available to the research team on the configuration, use, and features of this tool.
- The case study utilized a relatively new DSA document template with no publicly available prior examples. Additionally, the DOE-approved PDSA for the DOME test facility used an older template (DOE-3009-2014), which is quite different in structure from the new format. Therefore, the evaluation effort did not have the benefit of a benchmarking document for direct section-wise or chapter-wise comparison.

- The evaluation focused not only on an AI generation of the PDSA, but one with *this specific software*. As such, emphasis was placed on the settings and configuration of the Microsoft *Generative AI for Permitting Solution Accelerator*. Analysis of these settings may not be applicable to all publicly available software.
- AI models available for use with the software are restricted to models available on the U.S. Azure Government cloud. At the time of testing this study was performed with **gpt-4o** for content generation and **text-embedding-ada-002** for document embedding.
- The current evaluation of performance and results of generative AI for nuclear safety documentation is qualitative. Because of the stochastic nature of AI, the output is not reproducible or predictable. While scoring and benchmarking between AI models can be performed, there is no objective standard for evaluating the performance of an AI when developing nuclear-specific documentation.
- Data used for the basis of a facility and reactor PDSA in this evaluation may be incomplete compared to a full regulatory submission. The reactor design used is hypothetical and generic and may be missing key analyses that inform the safety case.

4.3 Conclusion

This case study demonstrated, for the first time within NRIC's digital ecosystem, successful integration of generative AI document development with a live digital thread architecture. The Microsoft Generative AI for Permitting Solution Accelerator was connected to authoritative project data through DeepLynx Nexus, establishing a functional pathway for AI-assisted synthesis of safety basis documentation from a single source of truth.

While performance was constrained by the limitations of pre-release software and the restricted model availability within the U.S. Government Azure cloud environment, SME evaluation confirmed that the generated DSA chapter would reduce document development effort by approximately 50 percent compared to traditional manual authoring. This reduction in effort represents a substantial improvement in efficiency while preserving the critical role of expert review and technical validation.

These results validate both the technical feasibility and the strategic value of AI-enabled, digital-thread-aware safety documentation. Continued maturation of models, embedding pipelines, and validation metrics is expected to further improve output quality and increase realized efficiency gains. The demonstrated capability positions NRIC to lead the responsible deployment of AI-supported safety basis development within the nuclear sector.

5. REFERENCES

[1] Artificial Analysis. (2025). *LLM Leaderboard – Comparison of over 100 AI models from OpenAI, Google, DeepSeek & others*. ArtificialAnalysis.ai. https://artificialanalysis.ai/leaderboards/models [Accessed Feb 24, 2026].

[2] Bergman, T. (2021). *Lessons-Learned from the Design Certification Review of the NuScale Power, LLC Small Modular Reactor*. (No. LO-0221-74035). NuScale Power, Rockville, MD. https://www.nrc.gov/docs/ML2105/ML21050A431.pdf

[3] Darrington, J. W., J. M. Browning, and C. S. Ritter. (2020). *Deep Lynx: Digital Engineering Integration Hub*. Idaho National Laboratory. https://github.com/idaholab/Deep-Lynx

[4] FAA. (2024). *FAA Roadmap for Artificial Intelligence Safety Assurance, Version I*. US Department of Transportation. https://www.faa.gov/aircraft/air_cert/step/roadmap_for_AI_safety_assurance [Accessed Aug 26, 2025].

[5] Lund, M. L., K. M. O'Rear, J. S. Rymer, K. J. Terrill, and P. A. Suyderhoud. (2024). *Framework for a digital documented safety analysis* (No. INL/RPT-24-78848-Rev000). Idaho National Laboratory (INL), Idaho Falls, ID. https://inldigitallibrary.inl.gov/sites/STI/STI/Sort_112773.pdf

[6] Mandelli, D., C. Wang, J. Cogliati, and V. Agarwal. (2023). "Data Fusion of Numerical and Textual Equipment Reliability Data: A Knowledge-Graph-based Approach." In *Proceedings of the Probabilistic Safety Assessment and Management (PSAM) 2023 Topical Conference*. https://www.ideals.illinois.edu/items/128993

[7] Patel, D. (2025). "Comparing Traditional OCR with Generative AI-Assisted OCR: Advancements and Applications." *International Journal of Science and Research* (IJSR), 347–351. https://doi.org/10.21275/sr25603211507

[8] Ritter, C., M. Rhodes. (2023). "Incorporating Digital Twins in Early Research and Development of Megaprojects to Reduce Cost and Schedule Risk." *INCOSE Insight*., vol. 26, issue 3. https://doi.org/10.1002/inst.12457

[9] Roy, S. S., A. M. Petrovic, H. K. Sandhu, K. M. O'Rear, P. A. Suyderhoud, and J. Rymer. (2025). *Evaluation of AI-enabled Digital Documented Safety Analysis* (No. OSTI ID: 3013976). Idaho National Laboratory (INL), Idaho Falls, ID. https://doi.org/10.2172/3013976

[10] Swartzendruber, Tracy. (2025). "What is a Digital Twin?" *GE Vernova Electrification Software* (blog). August 1, 2025. https://www.gevernova.com/software/blog/what-digital-twin [Accessed Aug 26, 2025].

[11] U.S. DOE. (2016). *Review and Approval of Nuclear Facility Safety Basis and Safety Design Basis Documents*. (No. DOE-STD-1104-2016). Retrieved from https://www.standards.doe.gov/standards-documents/1100/1104-astd-2016/@@images/file [Accessed Aug 26, 2025].