INL/CON-25-83082-Revision-0

AI for Interpreting Nuclear Power Plant Documents for Power Uprates

July 2025

Changing the World's Energy Future

John Norman Mapes, Kellen McKay Giraud, Brian M Wilcken, Ahmad Y Al Rashdan, Marcio De Queiroz

Idaho National Laboratory

INL/CON-25-83082-Revision-0

AI for Interpreting Nuclear Power Plant Documents for Power Uprates

John Norman Mapes, Kellen McKay Giraud, Brian M Wilcken, Ahmad Y Al Rashdan, Marcio De Queiroz

July 2025

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

AI for Interpreting Nuclear Power Plant Documents for Power Uprates

Norman John Mapes Jr.*, Kellen Giraud, Brian Wilcken, Ahmad Al Rashdan, Marcio de Queiroz

Idaho National Laboratory, Idaho Falls, Idaho

[leave space for DOI, which will be inserted by ANS]
LRS #: ***INL/CON-25-83082***

ABSTRACT

To reduce the cost and time needed for regulatory compliance, nuclear power plants (NPPs) can utilize artificial intelligence (AI) to assist in interpreting complex and voluminous documents that typically span thousands of pages. Usually, the process of interpreting a plant's technical specifications (TSs) and associated documents is labor intensive. This study aims to understand what processes state-of-the-art large language models (LLMs) can automate and to identify the pitfalls associated with using LLMs to reduce human labor costs and time. This research uses a recent AI technology called retrieval augmented generation (RAG), which retrieves pages of information from TSs and associated documents to assist with NPP power uprates (cleared to produce more power). LLMs are integral to RAG because they create human-like responses based on the retrieved information, aiding in the interpretation and application processes. A baseline case demonstrates how LLMs can operate successfully for a power uprate application. Then five use cases show five types of potential failures: (1) RAG retrieving the incorrect information, (2) RAG misinterpreting the retrieved information, (3) RAG relying on knowledge not contained in the retrieved information, (4) RAG hallucinating, and (5) RAG refusing to answer. The results of the five use cases suggest that automating the human interpretation of TSs and associated documents with AI should be approached with caution. A subject-matter expert reviewed the AI outputs from the five use cases and concluded that an LLM can produce technical information that is needed to produce power uprate applications in certain instances.

Keywords: retrieval augmented generation, nuclear power plant power uprate, nuclear power plant technical specifications, generative AI automation, NRC regulatory compliance

1. INTRODUCTION

The Nuclear Regulatory Commission (NRC) requires commercial nuclear power plants (NPPs) in the U.S. to operate at or below their approved maximum power level. If a plant wants to increase its power production, it requests a power uprate which involves submitting a license amendment request (LAR) and revised technical specifications (TSs) to the NRC for review and approval [1]. Power uprates have been approved in the U.S. since the 1970s, bringing approximately 8 GW of additional power to the existing nuclear fleet. All plant power uprates undergo rigorous safety assessments, and their successful implementation requires a thorough comprehension of the TSs and associated documents, including critical aspects of the plant's life cycle (e.g., design, construction, and maintenance) and operation (e.g., operational requirements, safety limits, control settings, emergency preparedness, quality assurance, radiation protection, and maintenance and testing protocols). These documents delineate permissible actions and regulatory constraints that affect the plant's ability to secure NRC approval for the uprate.

* Norman.Mapes@INL.gov

Not surprisingly, power uprate submissions are time-consuming and costly to prepare. They require extensive interpretation labor to comprehend the TSs and associated documents since it is unrealistic for any one person to have knowledge of all the information. Thus, the goal of this research is to reduce the time needed to draft power uprate submissions and their cost. Specifically, this work investigates using artificial intelligence (AI) models that can assimilate knowledge from TSs to enable the thorough evaluation of the impacts of power uprates. The AI model is meant to streamline the evaluation process by suggesting technical changes and ensuring precise, efficient updates to the TSs and associated documents. Ultimately, it will support the safe and effective implementation of power uprates that meet all regulatory requirements.

To support this solution, retrieval augmented generation (RAG), a recent AI tool, was explored. RAG is commonly used for drafting documents, answering questions, and analyzing text documents. RAG responds to prompts and includes information from relevant documents [2]. Typically, a subject-matter expert (SME) or a prompt engineer provides the prompt. The relevant documents in this work are the pages of TSs and associated documents for the power uprate. By incorporating RAG, the AI can retrieve and utilize large amounts of technical data to address regulatory guidelines. This technical data is extracted to form the LAR that responds to the regulatory guidelines. Five use cases have been created that mimic typical questions that a LAR drafter would need answered to perform his/her work. These questions exemplify common issues such as design pressures, system functions, and acronym definitions. The five use cases also demonstrate the following limitations: (1) RAG uses an incorrect document, (2) RAG locates the correct document but interprets the document incorrectly, (3) RAG does not use the document to answer the question but instead relies on its internal knowledge, (4) RAG hallucinates, and (5) RAG declines to answer.

The large language model (LLM) GPT-4[1] passed the entire uniform bar exam, including the multiple choice and essay sections [11]. This has demonstrated that LLMs can interpret notoriously complex legal language successfully and therefore may prove effective for absorbing and interpreting the complex knowledge available in TSs. GPT-4o, an upgraded version of this LLM, was used in this work. It is suggested here that RAG be used as an input along with the user query to the LLM. This allows the RAG system to use the relevant portions of the TSs and other documents to provide an answer. It has been observed that LLM models that use RAG create more specific, diverse, and factual language than models without RAG [2].

RAG was used in the nuclear domain in two recent works. Reference [3] detailed how RAG was used to retrieve relevant information for operators of advanced nuclear reactors. This work is similar to the present paper in that it discussed the idea of retrieving relevant information from thousands of pages of TSs and associated documents. The authors of Reference [3] concluded that RAG “displayed appropriate reasoning, comprehension, information retrieval, and extraction capabilities” (p. 340). Reference [4] discussed RAG being used for NPP diagnostics, and its authors concluded that “using RAG techniques can improve accuracy and reliability, allowing for the application of LLMs in specialized areas, even when those areas aren’t extensively covered in their initial training” (p. 4). This conclusion is relevant to the current work because not all the LLMs discussed here were trained on all the TS and associated documents. This paper differs from References [3] and [4] in that these address specific applications of discrepancy checking and diagnostics for nuclear power use cases, whereas the present work focuses on power uprate drafting. The contribution of this work is that it identifies what can and cannot be achieved with AI in reducing the cost and time needed to draft a power uprate application.

[1] GPT-4 is a specific version of an LLM released by OpenAI and is different than ChatGPT.

2. DATA SOURCE

For this effort, current NPP TSs and related data from 2024 were used to inform the AI tools. The NRC's Agencywide Documents Access and Management System (ADAMS) provides access to over 2 million bibliographic citations [5] including TSs for all or most of the operating NPPs in the United States. Additionally, there were six standard technical specifications (STSs) available for each standard plant design including Babcock & Wilcox, Combustion Engineering, General Electric BWR/4, General Electric BWR/6, Westinghouse AP1000, and other Westinghouse plants [6]. In the past licensees justified changes to TSs by comparing the proposed changes to the STSs related to their plant design [7]. Also relevant to these documents are TSs bases and updated final safety analysis reports (UFSARs). Therefore, the TSs bases and UFSARs were used to inform the AI tools.

The NRC has approved 172 power uprate applications since 1977, as shown in an NRC database [8]. The database is organized by the percentage uprate, type of uprate, and uprate date. Recent applications are typically more than 100 pages and include amendments to TSs. Also included in the power uprate application is the safety evaluation written by the NRC. The safety evaluation has a section for each affected plant aspect, including containment systems and pressure-temperature limits. Each aspect typically contains four items: regulatory evaluation, technical evaluation, licensee statements, and NRC conclusions. The licensee statements typically originate from the LAR corresponding to the power uprate application. The LAR contains a list of licensee statements on the impact of the uprate on the plant supported by technical background information. Additionally, there are two NRC guidance documents for power uprate LARs, one for MUR[2] (measurement uncertainty reduction) power uprates [9] and one for EPUs[3] (extended power uprates) [10]. These guidance documents may prove valuable for the AI to understand power uprates.

3. BASELINE CASE

This baseline case shows how the system works when it successfully retrieves the correct information and presents it to the user correctly. The Joseph M. Farley NPP was randomly selected from recent MUR uprates to avoid selection bias. RAG is expected to produce facts to help locate information from a collection of documents. There are 20 different PDF documents totaling 6,662 pages from the following three references: TSs, STS bases, and the UFSAR for the Joseph M. Farley NPP. A passage was taken from the STS bases and converted into a question. This passage was then successfully retrieved using the RAG process. A prompt was used to convert the question and context into a query that the Microsoft Phi-3-small LLM used to formulate a response. The LLM worked as expected, providing the correct response to the query according to the SME.

QUESTION: *Does the QPTR limit ensure that the gross radial power distribution remains consistent with the design values used in the safety analyses?*
RETRIEVED CONTEXT: *…The QPTR limit ensures that the gross radial power distribution remains consistent with the design values used in the safety analyses. …*
PROMPT: *You are an AI assistant specifically designed to support a nuclear power plant. Your replies must be directly related to the plant and the user's inquiries. Keep a professional tone, and ensure your answers are precise and useful. Closely follow the user's questions and give pertinent information. If you are unsure of the answer, simply say, "I don't know." Do not invent information. Here is some context to guide your response: [RETRIEVED CONTEXT] Question: [QUESTION]*

---

[2] An MUR uprate is typically about a 2% increase and involves minor instrumentation upgrades.
[3] An EPU uprate is typically greater than a 7% increase and involves significant modifications to equipment.

RESPONSE: *Yes, the QPTR (Quadrant Power Tilt Ratio) limit ensures that the gross radial power distribution remains consistent with the design values used in the safety analyses…*

4. USE CASES OF RAG FAILURES

RAG was demonstrated in these use cases for the tasks of design pressure determination, system function and definition of terms. However, it is important to understand the pitfalls of RAG systems so that the LAR drafters can work effectively with the AI assistance that interprets the TSs and associated documents. In these 5 use cases, five failure types (FTs) were identified in the RAG system: FT 1–FT 5. These failures, which are similar to the seven failure points (FPs) observed in Reference [12], are given in Table I. Neither the FTs nor FPs are likely to be comprehensive. Rather they are specific observations from the use cases presented in this work and from the case studies of Reference [12].

Table I. The failure types and use cases for RAG.

Section	Failure Number	Details	Use Case Task
4.1	FT 1	The retriever retrieves incorrect document.	Design pressure of the containment.
4.2	FT 2	The LLM incorrectly interprets the retrieved document.	Design pressure of the containment.
4.3	FT 3	The LLM uses its internal knowledge to answer the question instead of using the retrieved context.	CSS[4] limits (decreases) the peak containment pressure.
4.4	FT 4	The LLM hallucinates and gives incorrect facts.	CSS limits (decreases) the peak containment pressure.
4.5	FT 5	LLM declines to answer or says, "I don't know."	Define and describe the acronym MSLB[5].

To demonstrate these FTs, the following use cases are given. The Joseph M. Farley NPP LAR contains facts that are used in all five types of failure. The goal is to reproduce each fact using RAG, just as a drafter of the LAR would be expected to. As in Section 3, the same question, retrieved context, prompt and response are given and the same dataset and LLM are used. Importantly, the LAR fact is not given to the RAG system. By not feeding the LAR fact to the RAG system, there is a fair evaluation of what a LAR drafter would expect to have available (or unavailable) when creating the LAR. In each FT, a portion of the LAR fact, question, etc. has been put in bold text to emphasize the intent of the LAR drafter question.

4.1. Failure Type 1—Incorrect Retrieved Document

The relevant portion of the LAR fact in this use case was that the containment design pressure is 54 psig.

LAR FACT: *The CSS operates to limit peak containment pressure to less than the **design pressure of 54 psig** during a LOCA or MSLB in order to maintain containment structural integrity.*
QUESTION: *What is the peak **design pressure** of the containment?*

[4] CSS means containment spray system.

[5] MSLB means main steam line break.

RETRIEVED CONTEXT: *The maximum containment pressure resulting from the worst case LOCA, [44.1] psig, does not exceed the containment* ***design pressure, [55] psig****…*
PROMPT: *[SAME AS SECTION 3]*
RESPONSE: *The peak* ***design pressure*** *of the containment* ***is 55 psig****.*

In this case an incorrect answer was given (55 psig instead of 54) because the RAG engine failed to identify the correct document for answering the question. Specifically, it used the STS bases document and not the UFSAR document. The UFSAR page 6.2-6 is the correct document and supersedes the STS bases because it is more specific to the plant. The square brackets were added by the STS bases drafters. The square brackets around numbers in the retrieved context likely indicate that these numbers should be replaced with ones specific to the plant. The STS bases were included in the dataset searched because they contain facts that are not found in the other reference documents. The overarching idea is that the LAR fact and the retrieved context do not match.

4.2 Failure Type 2—Incorrect LLM Interpretation of Document

As in Section 4.1, the relevant portion of the LAR fact in this use case is that the containment design pressure is 54 psig. However, the wording of the question is different from that of Section 4.1, resulting in a different output. This was done to demonstrate an FT that is different from the one in Section 4.1. The "question" in this section is not actually a question but a statement, and the failure could have been avoided with a different prompt. Nevertheless, this type of failure is expected to occur even when a user uses better prompts to interact with the system.

LAR FACT: *[SAME AS SECTION 4.1]*
QUESTION: ***Peak design pressure*** *containment*
RETRIEVED CONTEXT: *The limiting LOCA was reanalyzed as described in paragraph 6.2.1.3.3, and the resultant peak pressure is below the* ***design pressure of 54 psig****, as shown in table 6.2-6.*
PROMPT: *[SAME AS SECTION 3]*
RESPONSE: *The peak design pressure for the containment of the nuclear power plant is* ***54 psig, as indicated in table 6.2-6****.*

The mistake in this example is not that the psig value is incorrect, but that the LLM reports that Table 6.2-6 contains the peak design containment pressure, which it does not. It appears that the LLM failed to verify the statement by double-checking the table. Therefore, this is an incorrect interpretation instead of a hallucination. Due to this incorrect response, a potential LAR drafter would not be able to use this RAG system without reviewing the outputs to verify that they are factual. So, the mistake was that the LLM stated that the information was retrieved from Table 6.2-6 rather than page 6.2-6. This is because the correct text was retrieved from the UFSAR page 6.2-6.

4.3 Failure Type 3—Internal LLM Knowledge Used Instead of Retrieved Knowledge

The relevant portion of the LAR fact in this use case is that the CSS operates to limit peak containment pressure to less than the design pressure of the containment.

LAR FACT: *[SAME AS SECTION 4.1]*
QUESTION: *Does the* ***CSS operate to limit peak containment*** *pressure to less than the design pressure?*
RETRIEVED CONTEXT: *[IRRELEVANT TEXT]*
PROMPT: *[SAME AS SECTION 3]*

RESPONSE: ***The CSS (Containment Spray System) is designed to limit the peak containment pressure*** *by providing cooling to the reactor containment structure and the reactor coolant system (RCS) during a loss-of-coolant accident (LOCA). The CSS operates to maintain the containment pressure below the design pressure by condensing steam and reducing the volume of gas within the containment...*

Here, the retrieved context is irrelevant to the question per an SME. This happened because either there was no good match for the question or the RAG pipeline failed to retrieve the correct document. This can occur when the reference documents do not contain the answer, or the retrieval system does not correctly understand the question or the reference documents. UFSAR, page 6.2-37, contains the following: "The function of the containment spray system is to spray water into the containment atmosphere, when appropriate, in the event of a LOCA to ensure that containment peak pressure is below its design value." Yet the model failed to retrieve this correct document page according to an SME. An additional and more noteworthy failure is that the model relied on its internal knowledge acquired during training to answer the question and not on the reference document. This can be problematic because the intended use of RAG is to understand the reference documents and answer according to those documents rather than relying on knowledge acquired during training by Microsoft. By relying on internal knowledge, RAG did not give an appropriate citation to support its response.

4.4 Failure Type 4—LLM Hallucinations

The relevant portion of the LAR fact in this use case is that the CSS limits (decreases) the peak containment pressure.

LAR FACT: *[SAME AS SECTION 4.1]*
QUESTION: *Does the CSS increase or **decrease** peak containment pressure during a LOCA or MSLB.*
RETRIEVED CONTEXT: *The maximum containment pressure resulting from the worst case LOCA, [44.1] psig, does not exceed the containment design pressure, [55] psig. The containment was also designed for an external pressure load equivalent to [-2.5] psig. The inadvertent actuation of the Containment Spray System was analyzed to determine the resulting **reduction** in containment pressure.*
PROMPT: *[SAME AS SECTION 3]*
RESPONSE: *The Containment Spray System (CSS) is designed to **decrease** peak containment pressure during a Loss of Coolant Accident (LOCA) or a **Minor Stationary Leak (MSLB).** The CSS cools the steam and reduces the pressure within the containment by condensing the steam back into water.*

This question is nearly the same as the one used in Section 4.3 but worded differently to demonstrate a different type of failure—that is, a hallucination. Hallucinations differ from the incorrect interpretation of data (Section 4.2) in that they involve giving false information in response to the question. In this example, MSLB stands for "main steam line break"; however, the LLM hallucinated the meaning to be "minor stationary leak." Nevertheless, the question was answered correctly (decreases pressure), but the response could not go into a potential LAR because of the factual inaccuracies in the hallucinated response.

4.5 Failure Type 5—LLM Declines to Answer

The relevant portion of the LAR fact in this use case is the meaning of MSLB. Here, the question is the definition of the acronym MSLB.

LAR FACT: *[SAME AS SECTION 4.1]*
QUESTION: *Please define and describe the acronym **MSLB**.*

RETRIEVED CONTEXT: *[IRRELEVANT TEXT]*
PROMPT: *[SAME AS SECTION 3]*
RESPONSE: ***I don't know. The provided context does not contain information about the acronym MSLB.***

Sometimes the LLM correctly reads the context retrieved but concludes that it cannot answer the question posed. This is not surprising because the prompt specifically requests the LLM to output "I don't know" when it does not know the answer and to not invent information (The prompt was: *You are an AI assistant specifically designed to support a nuclear power plant. Your replies must be directly related to the plant and the user's inquiries. Keep a professional tone, and ensure your answers are precise and useful. Closely follow the user's questions and give pertinent information. If you are unsure of the answer, simply say, "I don't know." Do not invent information. Here is some context to guide your response: [RETRIEVED CONTEXT] Question: [QUESTION]*.) This inability to answer may seem confusing because the LLM hallucinated an answer in a different test case (Section 4.4). However, the outcome of the current use case differs from a hallucination because the statement the LLM gives is a true response based on the retrieved context available. However, there is another failure at work here; MSLB is available in the reference texts (UFSAR page 6.2-31), so the RAG system failed to retrieve the correct document. In this specific case finding the relevant context can be presented more efficiently using a text search instead of the more sophisticated RAG system. As a future work, incorporating a combination of text search and sophisticated search may prove beneficial.

5. CONCLUSIONS

This research demonstrated that generative AI has the potential to assist personnel drafting a LAR for a NPP power uprate in certain instances. Since the documents and page numbers are provided when using RAG, this provides a degree of explainability, as demonstrated by the baseline case in Section 4. However, as shown in the other 5 uses cases, there are pitfalls associated with the system, such as incorrect document retrieval, LLM misinterpretations, reliance on internal knowledge, hallucinations, and refusal to answer. Knowing these limitations provides realistic expectations for how AI can and cannot assist in the power uprate drafting process. The limitations also chart the path for future work. Overall, the use cases indicate that RAG tools may present a new opportunity for increased productivity during power uprate drafting.

As previously mentioned, AiVA—an INL tool based on the OpenAI GPT-4o model—was used in this study. The latest OpenAI models, such as GPT-4o, are gaining traction across disciplines and are expected to be found in future work, including in the nuclear industry. Additionally, we used a local Microsoft Phi-3-small model to demonstrate testing for sensitive data that cannot be put into the cloud. By using these two well-known models, future researchers can duplicate our results and build upon our efforts to automate the interpretation and utilization of large bodies of nuclear-specific text. Further research is needed to fully realize these abundant opportunities to assist with document-interpretation labor.

ACKNOWLEDGMENTS

We gratefully acknowledge the support and guidance of Ahmad Al Rashdan, the project lead for this work. This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory under the Light Water Reactor Sustainability (LWRS) program. Neither the U.S. Government, nor any agency thereof, nor any of their employees 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 privately owned rights.

REFERENCES

1. U.S. NRC, "Power Uprates," U.S. Nuclear Regulatory Commission, https://www.nrc.gov/reactors/operating/licensing/power-uprates.html (2025).
2. P. Lewis, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, **33**, pp. 9459–9474 (2020).
3. P. Athe, L. Lin, and N. Dinh, "Using Generative AI to Implement the Discrepancy Checker for a Nearly Autonomous Management and Control System for Advanced Reactors," 2024 *International Congress on Advances in Nuclear Power Plants ICAPP* (2024).
4. T. E. Reeves, et al., *Improving Reliability of Large Language Models for Nuclear Power Plant Diagnostics*, No. INL/EXP-24-79591-Rev000, Idaho National Laboratory (2024).
5. U.S. NRC, "Web-Based ADAMS," U.S. Nuclear Regulatory Commission, https://adams.nrc.gov/wba/ (2024).
6. U.S. NRC, "Standard Technical Specifications–Operating and New Reactors–Current Versions," U.S. Nuclear Regulatory Commission, https://www.nrc.gov/reactors/operating/licensing/techspecs/current-approved-sts.html (2023).
7. U.S. NRC, "Catawba Nuclear Station, Units 1 and 2–Issuance of Amendment Nos. 319 and 315 to Technical Specification 3.7.11," ML24017A065, U.S. Nuclear Regulatory Commission, https://www.nrc.gov/docs/ML2401/ML24017A065.pdf (2024).
8. U.S. NRC, "Status of Power Uprate Applications," U.S. Nuclear Regulatory Commission, https://www.nrc.gov/reactors/operating/licensing/power-uprates/status-power-apps.html (2020).
9. U.S. NRC, "Regulatory Information Summary (RIS) 2002-03," U.S. Nuclear Regulatory Commission, https://www.nrc.gov/docs/ML0135/ML013530183.pdf (2002).
10. U.S. NRC, "Review Standard for Extended Power Uprates," U.S. Nuclear Regulatory Commission, https://www.nrc.gov/reactors/operating/licensing/power-uprates/rs-001-rev-0-dec2003.pdf (2003).
11. D. M. Katz, et al., "GPT-4 Passes the Bar Exam," *Philosophical Transactions of the Royal Society A*, **382** (2270), (2024).
12. S. Barnett, et al., "Seven Failure Points When Engineering a Retrieval Augmented Generation System," *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI* pp. 194–199 (2024).