

# Light Water Reactor Sustainability Program

## The Use of Artificial Intelligence to Interpret Technical Specifications for Regulatory Knowledge Capture and Drafting License Amendment Requests



September 2025

U.S. Department of Energy

Office of Nuclear Energy

**DISCLAIMER**

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed 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 herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does 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.

**The Use of Artificial Intelligence to Interpret Technical Specifications for  
Regulatory Knowledge Capture and Drafting License Amendment Requests**

Norman (John) Mapes, Ahmad Al Rashdan (Principal Investigator), Brian Wilcken,  
Marcio de Queiroz and Kellen Giraud

Idaho National Laboratory

September 2025

**Prepared for the  
U.S. Department of Energy  
Office of Nuclear Energy**

*Page intentionally left blank*

## EXECUTIVE SUMMARY

For any procedural changes, tests, or modifications at a nuclear power plant, a detailed self-assessment must determine whether a license amendment is necessary per 10 CFR 50.59 criteria. License amendments are mandatory when changes exceed these criteria and require substantial time and effort due to extensive documentation and analysis to ensure safety. The process includes preparing a license amendment request (LAR) and responding to the NRC's requests for additional information. The overarching hypothesis of this work is that artificial intelligence, particularly large language models (LLMs), can streamline this process by efficiently interpreting technical specifications, previous LARs, and plant-specific documents, thus expediting approvals, enhancing safety, and reducing preparation costs.

Given the substantial energy capacity of nuclear fuel, current nuclear reactors are prime candidates for expanding electrical power output through a process known as power uprate. However, power uprate applications are usually classified as modifications necessitating license amendments, which demands an in-depth understanding of nuclear power plant (NPP) technical specifications (TS), previously approved LARs, and plant-specific documentation. Successful LLM deployment can potentially (1) contribute to the safe operation of an NPP by highlighting impacts on safety systems, (2) facilitate approval from the NRC, and (3) reduce LAR preparation and drafting costs by reducing the need for human labor.

This research applied multiple methods to interpret regulatory documents pertinent to power uprates and assessed their efficacy in knowledge capture and LAR generation. These methods encompassed the use of retrieval-augmented generation, fine-tuning models with relevant datasets, and several forms of prompt engineering. The performance of these methods was evaluated by a subject matter expert, a novel method to autonomously evaluate LLMs, statistical tests, or by performing similarity analysis between the returned answer by an LLM and the expected one.

With respect to knowledge capture, the research found that frontier models achieved accuracies as high as 100%, but the models were not robust. If the prompt given to the LLM is changed slightly, the LLM will give incorrect answers. However, these models typically require 8 minutes per question, which somewhat limits their usefulness. Additionally, 450 questions and answers were automatically generated from a power uprate LAR and several non-frontier LLMs were tested and graded automatically. None of the non-frontier LLMs exceeded a 60% comprehension score. This indicates that models currently in use have failed to absorb all of the information necessary to generate a LAR.

With respect to LAR creation, the research found that retrieval augmented generation does not benefit drafting LARs through statistical tests. The fine-tuning of an LLM was evaluated and was found to benefit the quality of AI-drafted LARs. The primary method to generate a LAR is structured few-shot prompting (SFP) which means that text data is structured in a database and then used with a specialized prompt to automatically generate much of the LAR. For prompt engineering, the development of SFP was found to aid in generating LARs, but has the same limitations all LLMs have, such as hallucinations.

In the next phase of this research, four improvements to LLMs for drafting LARs have been identified and will be executed. These improvements include enhancing document retrieval capabilities, verifying source text accuracy, increasing reliance on retrieved materials when requested rather than depending on internal knowledge, and reducing the tendency to generate hallucinations. Additionally, future work will be devoted to documenting the workflow of a LAR drafter so that the LLM can replicate their workflow. From a process perspective, future research will engage the NRC to map the LAR review process in detail, enabling more targeted use of LLMs. These combined efforts will close the gap between current capabilities and the desired performance level, ultimately leading to more accurate and reliable LLM usage for knowledge capture of regulatory documents and LAR generation.

*Page intentionally left blank*

## **ACKNOWLEDGMENTS**

The authors wish to thank the Light Water Reactor Sustainability (LWRS) Program for funding this work. The authors also wish to thank the Nuclear Regulatory Commission (NRC) for making their valuable data available through the Agencywide Documents Access and Management System (ADAMS) database. Generative AI such as Idaho National Laboratory's AiVA and Microsoft Copilot was used to improve the wording of this document. This research made use of the resources of the High-Performance Computing Center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the U.S. Department of Energy and the Nuclear Science User Facilities under Contract No. DE-AC07-05ID14517.



# CONTENTS

EXECUTIVE SUMMARY .....	iii
ACKNOWLEDGMENTS .....	vi
ACRONYMS.....	xi
1. INTRODUCTION.....	1
1.1 Problem Statement .....	1
1.2 Power Uprate Use Case.....	1
1.3 Objective .....	3
1.4 Organization of the Report.....	3
2. DATASET .....	4
2.1 Technical Specifications .....	4
2.2 Previous Power Uprate LARs .....	6
2.3 Updated Final Safety Analysis Reports .....	7
3. METHODS.....	9
3.1 Knowledge Capture and Drafting Methods .....	9
3.1.1 Retrieval Augmented Generation.....	9
3.1.2 Fine-Tuning LLMs.....	11
3.1.3 Structuring.....	12
3.1.4 Few-Shot Prompting .....	12
3.1.5 Structured Few-Shot Prompting.....	13
3.2 Evaluation Method and Model Selection Rationale.....	14
3.2.1 Comprehension Measurement in Nuclear Texts .....	14
3.2.2 Model Selection .....	15
4. USE CASES.....	17
4.1 Application of Evaluation Methods to Use Cases.....	18
4.2 Use Case 1: Farley CRDM Knowledge Capture.....	19
4.2.1 Failure Type 1—Incorrect Retrieved Document.....	21
4.2.2 Failure Type 2— Failure to Verify Source Text Accuracy.....	21
4.2.3 Failure Type 3—Internal LLM Knowledge Used Instead of Retrieved Knowledge .....	22
4.2.4 Failure Type 4—LLM Hallucinations .....	23
4.3 Use Case 2: Byron and Braidwood CRDM Knowledge Capture .....	23
4.4 Use Case 3: Farley Entire LAR Knowledge Capture.....	25
4.5 Use Case 4: Drafting the Farley Uprate Application CRDM Section.....	26
4.5.1 LLM Knowledge from Proprietary OpenAI Training.....	27
4.5.2 LLM Knowledge from RAG.....	28
4.5.3 LLM Knowledge from Fine-Tuning .....	28

4.6	Use Case 5: Farley Recreate Entire LAR.....	29
4.6.1	Cosine Similarities of Autogenerated LAR to Actual LAR.....	30
4.6.2	SFP Limitations .....	35
4.7	Frontier LLM Evaluations.....	37
5.	CONCLUSIONS .....	39
6.	REFERENCES .....	39
7.	APPENDIX A – LAR Excerpts for 8 NPPs.....	44

## FIGURES

Figure 1.	ADAMS search parameters for alphabetical list of current TS.....	4
Figure 2.	A Farley Nuclear Plant TS ( <i>left</i> ) and a Westinghouse STS ( <i>right</i> ) sample with nearly identical numbering.....	5
Figure 3.	Locating previously approved uprate applications by using the NRC database.....	6
Figure 4.	Locating UFSARs in the NRC ADAMS database.....	8
Figure 5.	RAG flowchart.....	10
Figure 6.	SFP process for entire LAR auto-generation.....	13
Figure 7.	Process diagram for CMINT.....	14
Figure 8.	Comparison of mean cosine similarity values for SFP and no SFP for four LLMs.....	32
Figure 9.	Heat map of the cosine similarities of the actual LAR to autogenerated LARs.....	34
Figure 10.	Example TS figure and table.....	36
Figure 11.	PDF text that produces OCR artifacts (marked with red boxes).....	36

## TABLES

Table 1.	Sample of structured data from unstructured NRC-approved LARs (unstructured data are not shown here).....	12
Table 2.	Characteristics of LLMs used in this effort.....	16
Table 3.	LLM methods applied to specific cases.....	17
Table 4.	Instruct and embedding LLMs evaluated during each use case.....	18
Table 5.	Evaluation methods applied to each use case.....	19
Table 6.	The failure types for RAG.....	21
Table 7.	Average LLM comprehension score on the 450 test questions.....	26
Table 8.	Cosine similarity of autogenerated Farley LAR to actual Farley LAR comparing SFP to no SFP.....	31
Table 9.	Cosine similarity of autogenerated Farley LAR to actual Farley LAR with SFP comparing RAG to no RAG.....	32
Table 10.	Cosine similarity of autogenerated Farley LAR to actual Farley LAR without SFP comparing RAG to no RAG.....	33

Table 11. Cosine similarity of autogenerated Farley LAR to actual Farley LAR comparing fine-tuning to no fine-tuning. ....	33
Table 12. Frontier model outputs compared to SME-verified answers. ....	37

## ACRONYMS

ADAMS	Agencywide Documents Access and Management System
AC	alternating current
AI	artificial intelligence
AOR	analysis of record
API	application programmer interface
ASME	American Society of Mechanical Engineers
B/B	Byron and Braidwood
BLEU	bilingual evaluation understudy
BWR	boiling-water reactor
CANDU	Canadian deuterium-uranium
CFR	<i>Code of Federal Regulations</i>
CRDM	control rod drive mechanism
CSS	containment spray system
DOE	Department of Energy
EPU	extended power uprate
FP	failure point
FT	failure type
GE	General Electric
GPT	generative pretrained transformer
INL	Idaho National Laboratory
LAR	license amendment request
LCO	limiting condition for operation
LEFM	leading edge flow meter
LLaMA	large language model Meta AI
LLM	large language model
LOCA	loss of coolant accident
MSLB	main steam link break
MUR	measurement uncertainty recapture
MUR-PU	measurement uncertainty recapture–power uprate
NPP	nuclear power plant
NRC	Nuclear Regulatory Commission
NSSS	nuclear steam supply system
OCR	optical character recognition

PDF	portable document format
PTC	performance test code
QPTR	quadrant power tilt ratio
RAG	retrieval augmented generation
RCCA	rod cluster control assembly
RCS	reactor coolant system
RIS	regulatory information summary
ROUGE	recall oriented understudy for gisting evaluation
SBO	station blackout
SFP	structured few-shot prompting
SME	subject-matter expert
STS	standard technical specifications
TF-IDF	term frequency–inverse document frequency
TS	technical specifications
UFSAR	updated final safety analysis report
URL	uniform resource locator

# 1. INTRODUCTION

## 1.1 Problem Statement

For any change in procedures, tests, or modifications at a nuclear power plant, the plant is required to perform a documented and detailed self-assessment to determine whether a license amendment is necessary. This self-assessment evaluates whether the proposed changes meet the criteria listed in Title 10 of the Code of Federal Regulations (10 CFR) Section 50.59, "Changes, tests, and experiments." License amendments become mandatory when a plant modification crosses a threshold based on these criteria. Specifically, 10 CFR 50.59 outlines the conditions under which a plant can make changes without prior NRC approval and delineates the circumstances that necessitate a formal license amendment, which is processed through a license amendment request (LAR).

The challenge with license amendments is that they typically require significantly more time and effort to prepare compared to plant modifications that do not exceed the 10 CFR 50.59 criteria thresholds. This is because the process involves a detailed evaluation of the proposed changes, detailed documentation, and extensive analysis to demonstrate that the modifications will not adversely affect plant safety or operations. The preparation of a LAR is a time-consuming process that requires a thorough understanding of regulatory standards, the plant technical specification, and plant documentation to present evidence that the plant can safely implement the proposed changes.

During the license amendment process the NRC could issue requests for additional information (RAIs). These RAIs seek to clarify or expand upon the details provided in the LAR. The process of responding to RAIs can be time-consuming, as it involves collecting the necessary information, conducting additional analyses, and preparing responses that address the NRC's concerns comprehensively. This iterative process can extend the overall timeline for obtaining the necessary approvals to implement the plant modifications.

Artificial intelligence (AI), and specifically large language models (LLMs), have the potential to significantly streamline the license amendment process for nuclear power plants. An LLM is defined as a model designed to understand and generate text like a human, based on the vast amount of data used to train it. For example, OpenAI's GPT-4 language model successfully completed the entire uniform bar exam, including both the multiple-choice and essay components (Katz et al. 2024). This achievement indicates that LLMs are capable of understanding and processing highly intricate legal language, suggesting their potential effectiveness in assimilating and interpreting the complex knowledge found in *technical specifications (TS) and associated documents* (to be defined in the Section 2).

LLMs can efficiently understand and interpret TS and associated documents. By leveraging LLMs as support tools, it is possible to capture the vast amount of knowledge within these data sources and present it to users in a coherent and useful manner. Furthermore, LLMs can assist in preparing parts or even entire LARs, thereby enhancing the overall efficiency of the process.

As a result, LLMs can contribute to the safe operation of a nuclear power plant (NPP) by highlighting potential impacts on safety systems. They can quickly pinpoint areas where proposed modifications might affect critical safety parameters. Additionally, LLMs have the potential to accelerate the approval process from the Nuclear Regulatory Commission (NRC) by producing well-structured and thoroughly analyzed LARs. LLMs can also significantly reduce LAR preparation and drafting costs by minimizing the need for extensive human labor. Traditionally, the preparation of a LAR involves a team of experts who must meticulously review and draft the necessary documentation. LLMs may be able to automate many of these tasks, thereby freeing up human resources to focus on higher-level decision-making and oversight.

## 1.2 Power Uprate Use Case

The recent surge in AI, combined with substantial investments in data centers to support computationally heavy LLM applications, has led to an accelerated demand for energy that presents

significant logistical challenges. These challenges include the need to enhance existing infrastructure to accommodate the increased energy consumption, which is both a complex and resource-intensive endeavor. Consequently, a more achievable approach is to explore methods for harnessing additional energy from the existing energy generation infrastructure. Given the substantial energy capacity of nuclear fuel, current nuclear reactors emerge as a prime candidate for enhancing electrical power output through a process known as power uprate.

Power uprates involve modifications to the reactor's systems and operations to safely increase the reactor's power output, thereby generating more electricity. The NRC requires commercial NPPs to operate at or below the approved maximum power level. The power level is used to establish the safety of the plant and is available in the license and TS for the plant. The TS and associated documents for an NPP encompass critical aspects of the plant's operation, including operational requirements, safety limits, control settings, emergency preparedness, quality assurance, radiation protection, and protocols for maintenance and testing. All plant power uprates undergo rigorous safety assessments. Therefore, a thorough understanding of the TS and associated documents is crucial for securing a power uprate approval from the NRC.

Since the 1970s, 172 power uprates have been approved, bringing approximately 8.5 GW of additional electric power to the existing fleet. This is the equivalent of approximately eight large reactors. Some plants have had multiple power uprates. This indicates that this is a feasible approach to generate more energy from the operating reactors. Power uprates are also aligned with the recent executive order 14302 "Reinvigorating the Nuclear Industrial Base," issued on May 23, 2025. This executive order sets a goal of 5 GW of power uprates by 2030. For context, 5 GW is the equivalent output of five large nuclear reactors out of the 94 that are currently operating in the U.S. fleet.

In general, there are two types of uprates that have been approved: (1) measurement uncertainty recapture power uprate (MUR-PU) and (2) extended power uprates (EPUs). With an MUR-PU, the uncertainty of the power level of the reactor is reduced, allowing a corresponding increase in the power level of the plant—typically less than 2%. With an EPU, NPP components, including the pipes, valves, pumps, heat exchangers, electrical transformers, and generators, require significant modifications and are evaluated to determine whether they can handle higher steam, water, and electric flow through the systems in question. Typically, EPUs are increases of 7–20%.<sup>a</sup> When licensees modify, replace, or increase the stress and loads on components, they must analyze the changes and demonstrate that the new configuration is safe and protects the health and safety of the public. These analyses are complex and are reviewed by NRC technical, legal, and management experts before the uprate is approved.<sup>b</sup> Implementing a power uprate, especially an EPU, is not an easy task. It involves a comprehensive analysis of plant equipment specifications, operating conditions, technical documentation, and regulatory requirements. Extensive documentation must be reviewed and updated to reflect the changes.

Regulatory compliance is a crucial aspect of implementing a power uprate. The plant must adhere to stringent guidelines set by the NRC. The preparation of power uprate LARs demands an in-depth understanding of NPP TS and associated documents. This involves demonstrating the plant's ability to safely operate at the higher power level. When seeking a power uprate, the plant may submit a LAR and

---

<sup>a</sup> "Types of Power Uprates," U.S. Nuclear Regulatory Commission, last updated March 10, 2020, <https://www.nrc.gov/reactors/operating/licensing/power-uprates/type-power.html>

<sup>b</sup> "Background on Power Uprates," U.S. Nuclear Regulatory Commission, last updated March 10, 2020, <https://www.nrc.gov/reactors/operating/licensing/power-uprates/about-power.html>.

revised TS and associated documents to the NRC for approval.<sup>c</sup> The LAR must address all safety concerns, including thermal-hydraulic analysis, structural integrity, and radiation protection. Crafting a LAR requires meticulous attention to detail and a thorough grasp of regulatory standards, as well as the ability to anticipate and address any potential safety concerns that may arise from the proposed modifications. As such, the process is both rigorous and essential, ensuring that the increased energy output does not compromise the operational integrity of the nuclear facility. The regulatory review process is complex and resource-intensive, demanding careful execution to ensure streamlined licensing.

Given the extensive volume of approximately 10,000 pages of TS and associated documents, it is impractical for an individual to possess comprehensive knowledge of the entire content. Therefore, interpreting and comprehending this vast amount of information, and producing the documents that meet regulatory requirements, requires extensive labor, which is costly and time-consuming. This made this use case a great candidate for the demonstration of LLM capability in assisting this process.

### **1.3 Objective**

This work investigates using generative LLM to assimilate knowledge from TS and associated documents to enable thorough evaluations of the impacts of plant power uprates. The LLM is meant to streamline the evaluation process by suggesting technical content to create the LAR. Ultimately, the goal is to use the LLM to support the effective development of a LAR for plant power uprates that meet all regulatory requirements. This decreases the effort needed by NPP staff to generate a complex LAR.

### **1.4 Organization of the Report**

This work is executed through four main tasks: (1) conduct a thorough review of the TS and associated documents; (2) create a novel LLM and methods to absorb and interpret data from the TS and associated documents, focusing on regulatory constraints and operational implications for power uprates; (3) validate recommendations through collaboration with subject-matter experts (SMEs), to support safe power uprate implementation; and (4) use the LLM and methods to assess the impact of plant power uprates on LARs and make recommendations aligned with regulatory guidelines.

This report is structured as follows. Section 2 delineates the dataset available for developing LLM methods. Section 3 details the methods utilized to extract insights from the data, including the retrieval augmented generation (RAG) method, the fine-tuning method to create a novel generative-AI LLM, the prompt engineering and the use of structuring method, the few-shot prompting method, and the novel structured few-shot prompting (SFP) method. It also discusses the comprehension measurement in nuclear texts (CMINT) method, and the basis for selecting the used models. The focus then shifts toward the use cases in Section 4. This comparison is the linchpin for the conclusions drawn from the study in Section 5, which encapsulate the efficacy of the methods.

---

<sup>c</sup> “Power Uprates,” U.S. Nuclear Regulatory Commission, last updated January 6, 2025, <https://www.nrc.gov/reactors/operating/licensing/power-uprates.html>.

## 2. DATASET

In this section, the different publicly available documents that are used as input data to the LLMs are described. Specifically, plant TS, standard technical specifications (STS), bases of TS, previously accepted LARs, and updated final safety analysis reports (UFSARs) are discussed, including their source and significance to the LAR process. Throughout this document the term “TS and associated documents” includes all of these documents, however in certain cases the associated documents used varies and is noted.

### 2.1 Technical Specifications

The NRC defines TS as “Part of an NRC license authorizing the operation of a nuclear production or utilization facility. A *Technical Specification* establishes requirements for items such as safety limits, limiting safety system settings, limiting control settings, limiting conditions for operation, surveillance requirements, design features, and administrative controls.”<sup>d</sup> To support this research, the latest NPP TS data from 2024 was utilized to guide the LLMs (and in some cases to train the LLM). The NRC’s Agencywide Documents Access and Management System (ADAMS) offers access to more than 2 million bibliographic citations,<sup>e</sup> including the latest TS used in this work. The query for finding NPP TS on ADAMS is shown in Figure 1. The search interface in Figure 1 has two recommended settings: (1) the documents must contain the keyword “technical,” as in “technical specifications,” and (2) the document titles must contain the phrase “current facility operating license” because the TS are attached to the licenses. TS for most of the operating NPPs in the U.S. appear to be available.

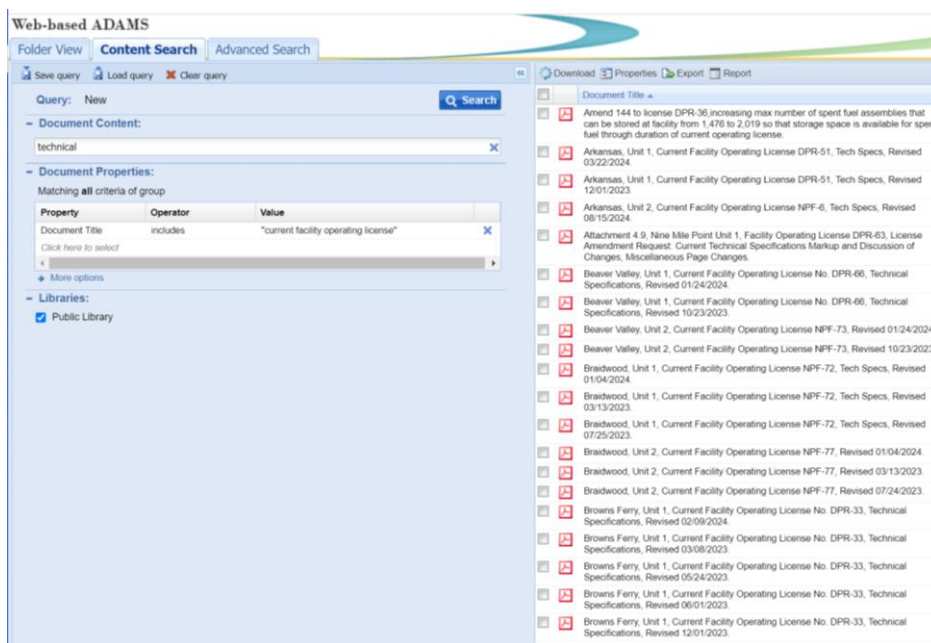


Figure 1. ADAMS search parameters for alphabetical list of current TS.

<sup>d</sup> “Glossary,” U.S. Nuclear Regulatory Commission, last updated August 26, 2025, <https://www.nrc.gov/reading-rm/basic-ref/glossary.html>.

<sup>e</sup> “Web-Based ADAMS,” U.S. Nuclear Regulatory Commission, accessed September 5, 2025, <https://adams.nrc.gov/wba>.

STS are a nuclear reactor designer’s TS specific for a particular reactor design. NPPs use STS to create a plant-specific TS. Six STS are available, one for each of Babcock & Wilcox, Combustion Engineering, General Electric (GE) BWR/4, GE BWR/6, Westinghouse AP1000, and non-AP1000 Westinghouse plants.<sup>f</sup> Licensees can make changes to their TS by showing that proposed changes align with the STS related to their plant design. For example, this excerpt is from a regulatory document justifying a TS change using the STS as a basis: “The licensee compares the proposed 24-hour completion times to the completion times approved in the Improved Standard Technical Specifications (STS) for other reactor designs. Although the referenced STS are not directly applicable to Catawba, the NRC finds the justification for approval in the STS would also apply to the Catawba proposed change because the control room cooling function is largely independent of the reactor designs.” (U.S. NRC 2024)

In terms of generated text, the value of this work is drafting a LAR rather than editing the TS for several reasons. First, deriving the specific numerical changes required for the TS, such as allowed operating power, fuel operating temperature, and reactor coolant system flow rate, is challenging without example calculations. Since this information is not publicly available, LLMs cannot replicate the calculations accurately. Second, drafting the LAR, which often exceeds 150 pages, is an extensive writing task. For instance, during the 2020 Joseph M. Farley power uprate, the TS experienced only minor verbiage changes, limited to fewer than 20 words, while the LAR required significant content development.

TABLE OF CONTENTS		TABLE OF CONTENTS		Page
1.0	USE AND APPLICATION	1.1-1	ABSTRACT	iii
1.1	Definitions	1.1-1	B 2.0 SAFETY LIMITS (SLs)	B 2.1.1-1
1.2	Logical Connectors	1.2-1	B 2.1.1 Reactor Core SLs	B 2.1.1-1
1.3	Completion Times	1.3-1	B 2.1.2 Reactor Coolant System (RCS) Pressure SL	B 2.1.2-1
1.4	Frequency	1.4-1	B 3.0 LIMITING CONDITION FOR OPERATION (LCO) APPLICABILITY	B 3.0-1
2.0	SAFETY LIMITS (SLs)	2.0-1	B 3.0 SURVEILLANCE REQUIREMENT (SR) APPLICABILITY	B 3.0-17
2.1	SLs	2.0-1	B 3.1 REACTIVITY CONTROL SYSTEMS	B 3.1.1-1
2.2	SL Violations	2.0-1	B 3.1.1 SHUTDOWN MARGIN (SDM)	B 3.1.1-1
3.0	LIMITING CONDITION FOR OPERATION (LCO) APPLICABILITY	3.0-1	B 3.1.2 Core Reactivity	B 3.1.2-1
3.0	SURVEILLANCE REQUIREMENT (SR) APPLICABILITY	3.0-3	B 3.1.3 Moderator Temperature Coefficient (MTC)	B 3.1.3-1
3.1	REACTIVITY CONTROL SYSTEMS	3.1.1-1	B 3.1.4 Rod Group Alignment Limits	B 3.1.4-1
3.1.1	SHUTDOWN MARGIN (SDM)	3.1.1-1	B 3.1.5 Shutdown Bank Insertion Limits	B 3.1.5-1
3.1.2	Core Reactivity	3.1.2-1	B 3.1.6 Control Bank Insertion Limits	B 3.1.6-1
3.1.3	Moderator Temperature Coefficient (MTC)	3.1.3-1	B 3.1.7 Rod Position Indication	B 3.1.7-1
3.1.4	Rod Group Alignment Limits	3.1.4-1	B 3.1.8 PHYSICS TESTS Exceptions - MODE 2	B 3.1.8-1
3.1.5	Shutdown Bank Insertion Limits	3.1.5-1	B 3.2 POWER DISTRIBUTION LIMITS	B 3.2.1A-1
3.1.6	Control Bank Insertion Limits	3.1.6-1	B 3.2.1A Heat Flux Hot Channel Factor ( $F_{Q(Z)}$ ) (CAOC- $F_{QZ}$ Methodology)	B 3.2.1A-1
3.1.7	Rod Position Indication	3.1.7-1	B 3.2.1B Heat Flux Hot Channel Factor ( $F_{Q(Z)}$ ) (RAOC-W(Z) Methodology)	B 3.2.1B-1
3.1.8	PHYSICS TESTS Exceptions—MODE 2	3.1.8-1	B 3.2.1C Heat Flux Hot Channel Factor ( $F_{Q(Z)}$ ) (CAOC-W(Z) Methodology)	B 3.2.1C-1
3.2	POWER DISTRIBUTION LIMITS	3.2.1-1	B 3.2.2 Nuclear Enthalpy Rise Hot Channel Factor ( $F_{EN}^{(H)}$ )	B 3.2.2-1
3.2.1	Heat Flux Hot Channel Factor ( $F_{Q(Z)}$ )	3.2.1-1	B 3.2.3A AXIAL FLUX DIFFERENCE (AFD) (Constant Axial Offset Control (CAOC) Methodology)	B 3.2.3A-1
3.2.2	Nuclear Enthalpy Rise Hot Channel Factor ( $F_{EN}^{(H)}$ )	3.2.2-1	B 3.2.3B AXIAL FLUX DIFFERENCE (AFD) (Relaxed Axial Offset Control (RAOC) Methodology)	B 3.2.3B-1
3.2.3	AXIAL FLUX DIFFERENCE (AFD)	3.2.3-1	B 3.2.4 QUADRANT POWER TILT RATIO (QPTR)	B 3.2.4-1
3.2.4	QUADRANT POWER TILT RATIO (QPTR)	3.2.4-1	B 3.3 INSTRUMENTATION	B 3.3.1-1
3.3	INSTRUMENTATION	3.3.1-1	B 3.3.1 Reactor Trip System (RTS) Instrumentation	B 3.3.1-1
3.3.1	Reactor Trip System (RTS) Instrumentation	3.3.1-1	B 3.3.2 Engineered Safety Feature Actuation System (ESFAS) Instrumentation	B 3.3.2-1
3.3.2	Engineered Safety Feature Actuation System (ESFAS) Instrumentation	3.3.2-1	B 3.3.3 Post Accident Monitoring (PAM) Instrumentation	B 3.3.3-1
3.3.3	Post Accident Monitoring (PAM) Instrumentation	3.3.3-1	B 3.3.4 Remote Shutdown System Instrumentation	B 3.3.4-1
3.3.4	Remote Shutdown System Instrumentation	3.3.4-1	B 3.3.5 Loss of Power (LOP) Diesel Generator (DG) Start Instrumentation	B 3.3.5-1
3.3.5	Loss of Power (LOP) Diesel Generator (DG) Start Instrumentation	3.3.5-1	B 3.3.6 Containment Purge and Exhaust Isolation Instrumentation	B 3.3.6-1
3.3.6	Containment Purge and Exhaust Isolation Instrumentation	3.3.6-1	B 3.3.7 Control Room Emergency Filtration System (CREFS) Actuation Instrumentation	B 3.3.7-1
3.3.7	Control Room Emergency Filtration/Pressurization System (CREFS) Actuation Instrumentation	3.3.7-1	B 3.3.8 Fuel Building Air Cleanup System (FBACS) Actuation Instrumentation	B 3.3.8-1
3.3.8	Penetration Room Filtration (PRF) System Actuation Instrumentation	3.3.8-1	B 3.3.9 Boron Dilution Protection System (BDPS)	B 3.3.9-1
			B 3.4 REACTOR COOLANT SYSTEM (RCS)	B 3.4.1-1
			B 3.4.1 RCS Pressure, Temperature, and Flow Departure from Nucleate Boiling (DNB) Limits	B 3.4.1-1
			B 3.4.2 RCS Minimum Temperature for Criticality	B 3.4.2-1

Figure 2. A Farley Nuclear Plant TS (left) and a Westinghouse STS (right) sample with nearly identical numbering.

<sup>f</sup> “Standard Technical Specifications – Operating and New Reactors – Current Versions,” U.S. Nuclear Regulatory Commission last updated April 4, 2023, <https://www.nrc.gov/reactors/operating/licensing/techspecs/current-approved-sts.html>.

The *bases* section of a TS includes technical, regulatory, and safety explanations for the requirements given in the document. The bases are written alongside every section of the TS with a mostly identical numbering system, as seen in Figure 2. Typically, the NRC does not require bases to be published and therefore they are not usually publicly available documents, except for the STS bases that are available from the NRC. STS bases are critical for the LLMs to understand the TS for power uprates. For example, in the Farley Nuclear Plant application for a 1.7% power uprate there is a subsection, “Condensate Inventory for Decay Heat Removal,” in section “V.1.B Station Blackout Equipment,” that refers to the TS Limiting Condition for Operation (LCO) 3.7.6 and a figure of 102% rated thermal power. This 102% figure is found in the STS bases under section 3.7.6 “LCO” but does not appear to be in the TS (U.S. NRC ADAMS Accession No. ML19308A763 2019 and ML21259A159 2021). Therefore, a LAR drafter needs the bases document in addition to the TS.

## 2.2 Previous Power Uprate LARs

Since 1977, the NRC has approved 172 uprate applications, as recorded in an NRC database as shown in Figure 3.<sup>g</sup> This database categorizes the applications by the percentage of the uprate, the type of uprate, and the date of the uprate. Recent applications typically exceed 100 pages and include amendments to the TS. Each power uprate application also contains a safety evaluation conducted by the NRC. This evaluation is divided into sections for each plant aspect affected, such as containment systems and pressure-temperature limits. These sections generally include four components: (1) regulatory evaluation, (2) technical evaluation, (3) licensee statements, and (4) NRC conclusions. The licensee statements are usually derived from the LAR associated with the power uprate application, which contains a detailed list of the licensee’s statements on the uprate’s potential impact on the plant, backed by technical information.

The LAR for the uprate application is available on ADAMS, with its document number listed in the references section of the NRC’s uprate application. Moreover, there are two specific guidance documents for power uprate LARs: one for MUR-PU’s (U.S. NRC 2002) and another for EPU’s (U.S. NRC 2003). These guidance documents are crucial for understanding power uprates.

NO.	PLANT	% UPRATE	MWT	DATE APPROVED	TYPE	ACCESSION #
157	Catawba 1	1.7	58	04/29/16	MU	<a href="#">ML16081A333</a>
158	Catawba 2	1.7	58	04/29/16	MU	<a href="#">ML16081A333</a>
159	Columbia Generating Station	1.66	58	05/11/17	MU	<a href="#">ML17095A117</a>
160	Browns Ferry 1	14.3	494	08/14/17	E	<a href="#">ML17032A120</a>
161	Browns Ferry 2	14.3	494	08/14/17	E	<a href="#">ML17032A120</a>
162	Browns Ferry 3	14.3	494	08/14/17	E	<a href="#">ML17032A120</a>
163	Peach Bottom 2	1.66	65	11/15/17	MU	<a href="#">ML17286A013</a>
164	Peach Bottom 3	1.66	65	11/15/17	MU	<a href="#">ML17286A013</a>
165	Hope Creek 1	1.6	62	04/24/18	MU	<a href="#">ML18096A542</a>
166	Farley 1	1.7	46	10/09/20	MU	<a href="#">ML20121A283</a>
167	Farley 2	1.7	46	10/09/20	MU	<a href="#">ML20121A283</a>
168	Watts Bar 2	1.4	48	10/21/20	MU	<a href="#">ML20226A444</a>
169	Oconee 1	1.64	42	01/26/21	MU	<a href="#">ML20335A001</a>
170	Oconee 2	1.64	42	01/26/21	MU	<a href="#">ML20335A001</a>
171	Oconee 3	1.64	42	01/26/21	MU	<a href="#">ML20335A001</a>
172	Millstone 3	1.6	59	11/09/21	MU	<a href="#">ML21262A001</a>

Figure 3. Locating previously approved uprate applications by using the NRC database.

<sup>g</sup> “Status of Power Uprate Applications,” U.S. Nuclear Regulatory Commission, last updated March 10, 2020, <https://www.nrc.gov/reactors/operating/licensing/power-uprates/status-power-apps.html>.

Over the last 23 years, there have been numerous MUR-PU's due to the availability of instrumentation produced by Cameron, formerly known as Caldon. Cameron's device reduces the reactor power level measurement uncertainty by approximately 2%, and therefore power uprates of approximately 2% were approved by the NRC. For this work, it was desired to focus the research on one type of power uprate. Since MUR-PU's follow a similar pattern and there are eight uprate applications (i.e., eight LARs), their unique language, despite their similarities, makes them an ideal use case for testing the proposed methods in Section 4.

NPP LAR sections are numbered per the NRC Regulatory Issue Summary (RIS) 2002–3 (U.S. NRC 2002). To test an LLM's ability to generate a LAR section, a section of the LAR must be identified that is not excessively plant-specific. For instance, the installed location of the Caldon instrumentation is so plant-specific that the LLM would not be able to generalize the plant's responses to other plants. On the other hand, a section that is boilerplate is not desired because such sections are mostly copied and pasted. From LARs examination, Section IV.1.A.iii, "Control Rod Drive Mechanisms (CRDM)," of an MUR-PU was found to be neither overly plant-specific nor boilerplate. In Appendix A, the LAR CRDM sections are organized by utility, plant name, ADAMS accession number, and date of application.

### 2.3 Updated Final Safety Analysis Reports

UFSARs document a nuclear facility's design, safety analyses, and operational limits, ensuring compliance with regulatory requirements and providing a reference for ongoing safe plant operation. Updated regularly, UFSARs give regulators and plant staff a complete, accurate record of facility safety and performance. By consolidating technical data, risk assessments, and licensing basis, UFSARs support informed decision-making, enhance transparency, and strengthen confidence in a nuclear facility's long-term safety.

UFSARs contain the description of every plant to be used as a reference for present and future safety analyses needed by the plant, the NRC, and other interested parties. Importantly, the NRC refers to the UFSAR for regulatory oversight, such as for evaluating power uprate LARs. The public can also retrieve UFSARs to gain an understanding of the plant and its operation. An UFSAR contains text, tables, and diagrams that explain the design bases, safety analyses, and facility operation for normal conditions, accidents, anticipated internal and external events, and natural phenomena under which the plant is designed to function (Nuclear Energy Institute 1999). An UFSAR contains roughly 18 chapters, and a typical chapter is 100+ pages. "FSAR" is mentioned 182 times in the Farley 1.7% power uprate LAR, while "TS" is mentioned only 50 times. Therefore, UFSARs must be part of the information the LLM uses to inform power uprates.

UFSARs can be accessed through a content search of the web-based NRC ADAMS database, as shown in Figure 3.<sup>h</sup> Each UFSAR can be downloaded by exporting the results and then automating a download of each URL for the 18 chapters and appendices. For this effort the Byron and Farley NPP UFSARs were downloaded. The query name in Figure 4 is called "New"; this is the default name to save to and does not reflect the actual search that was performed. The wildcard character "\_" was used to indicate that any text is acceptable as long as the document title contains three elements: "Byron," "updated final safety analysis report," and "revision 19." Revision 19 was chosen through trial and error incrementing the revision numbers until no later revisions appeared.

---

<sup>h</sup> "Web-Based ADAMS," U.S. Nuclear Regulatory Commission, accessed September 10, 2025, [https://adams.nrc.gov/wba/?data=\(mode:sections,sections:\(filters:\(public-library:!t\),properties\\_search:!\('\\$title',infolder,byron,\),'!\('\\$title',infolder,'updated+final+safety+analysis+report',\),'!\('\\$title',infolder,'revision+19',\)\),single\\_content\\_search:\\_\)&qn=New&tab=content-search-pars&z=0](https://adams.nrc.gov/wba/?data=(mode:sections,sections:(filters:(public-library:!t),properties_search:!('$title',infolder,byron,),'!('$title',infolder,'updated+final+safety+analysis+report',),'!('$title',infolder,'revision+19',)),single_content_search:_)&qn=New&tab=content-search-pars&z=0).

## Web-based ADAMS

Folder View **Content Search** Advanced Search

Save query Load query Clear query

Query: New Search

- Document Content:  
-

- Document Properties:  
Matching all criteria of group

Property	Operator	Value	
Document Title	includes	byron	X
Document Title	includes	updated final safety analysis report	X
Document Title	includes	revision 19	X

[Click here to select](#)

+ More options

- Libraries:  
 Public Library

Figure 4. Locating UFSARs in the NRC ADAMS database.

## 3. METHODS

In this effort, several LLM methods were explored for knowledge capture and LAR drafting, including RAG (Lewis et al. 2020), fine-tuning (Devlin et al. 2019), structuring (Al Rashdan et al. 2024), and few-shot prompting (Brown et al. 2020). These methods are described below along with a literature review of their prior uses in related technical fields, particularly nuclear. A new LAR drafting method was also introduced: *structured few-shot prompting* (SFP), which combines structuring and few-shot prompting. Then a new knowledge capture evaluation method was introduced: *comprehension measurement in nuclear texts* (CMINT), which validates the nuclear knowledge of the LLMs. Finally, a discussion of the selection of models is provided.

### 3.1 Knowledge Capture and Drafting Methods

#### 3.1.1 Retrieval Augmented Generation

RAG retrieves relevant documents to help an LLM respond to user prompts. The relevant documents in this work are the pages of a TS and associated documents. RAG is commonly used for document drafting, question answering, and other types of analysis of text documents, as is reflected in the literature. By incorporating RAG, the LLM can retrieve and utilize large amounts of technical data. It has been observed that RAG models create more specific, diverse, and factual language than models without RAG (Lewis et al. 2020).

RAG has been used in the nuclear domain in at least four recent works. For instance, Anwar et al. (2024) used RAG to interpret Canadian deuterium-uranium (CANDU) reactor information, which is similar to the work explored here to interpret TS and associated documents. Athe, Lin, and Dinh (2024) investigated using RAG to retrieve relevant information for operators of advanced nuclear reactors. This is analogous to the current work to retrieve relevant information from approximately 10,000 pages of TS and associated documents. Additionally, Athe, Lin, and Dinh (2024) concluded that RAG “displayed appropriate reasoning, comprehension, information retrieval, and extraction capabilities” (p. 340). Reeves et al. (2024) further explored using RAG for NPP diagnostics. They 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 latter point is important to the current work because not all of the LLMs used have been trained on all the TS and associated documents. Mehta et al. (2023) used RAG to create a copilot for searching documents and assisting with operations and question answering. The current work is similar in that it seeks to aid LAR drafters during the power uprate application process.

RAG has been used in the technological and research domains for question answering in microchip design (Liu et al. 2023), generating 6G network specifications (Tang and Guo 2024), automating literature reviews (Han, Susnjak, and Mathrani 2024), and composing documents (Chen and Tran 2024). RAG has been deployed in other fields that are notoriously complex, including medical and legal. NPP documents are no less complex, thus these previous works are relevant to the current work. In the medical field, RAG was used to analyze electronic health records (Alkhalaf et al. 2024), to interpret clinical guidelines (Krešević et al. 2024), to answer biomedical questions (Xiong et al. 2024), and to generate x-ray reports (Ranjit et al. 2023). In the legal field, Chouhan and Gertz (2024) used RAG to draft definitions in legal documents, Chu et al. (2024) used it to analyze patent texts, and Hou et al. (2024) used it to write legal documents.

The current work differs from these prior works because of its application focus. While Anwar et al. (2024) and Athe, Lin, and Dinh (2024) focused on interpreting reactor information and retrieving relevant information for reactor operators, this work specifically aims to aid LAR drafters during the power uprate application process. Moreover, the current work advances the state of knowledge by using the Atomic

Canyon Fermi-1024 model.<sup>i</sup> This model is important because it was specifically designed for the nuclear sector and incorporates specialized nuclear vocabulary that potentially improves the precision and relevance of document retrieval and question answering. The Atomic Canyon Fermi-1024 model was used for the RAG system to convert text into a numerical representation (i.e., sparse embeddings). The model was trained using the Microsoft MiniLM model, which is a common choice for searching documents in a RAG framework (W. Wang et al. 2020). This work did not use only Atomic Canyon Fermi-1024. Even though Atomic Canyon Fermi-1024 extends the state of knowledge, this work compares Atomic Canyon Fermi-1024 results to other methods. This work also used MiniLM. Additionally, term frequency–inverse document frequency (TF-IDF) from Spärck Jones (1972) was used because it performs the same function of converting text to an embedding. Manual retrieval was used (e.g., manually selecting a chapter from thousands of pages of documents). In manual retrieval the LLM still examines the documents, it is just aided by a human narrowing the selection of the documents (e.g., to a single chapter)

RAG facilitates the interaction between the plant personnel drafting the LAR (labeled “LAR drafter” in Figure 5) and the specific information that needs to be included in the LAR (“uprate information”). The LLM denoted “embedding LLM” is used twice. The first use, denoted as “previously assembled” in the gray box, converts the “Tech Specs, UFSAR, previous LARs 1,000+ pages” into “document text embeddings” (an embedding is a numerical representation of a document). The document text embeddings are stored in a database denoted “vector database” (vector is another word for an embedding, hence the name) that allows efficient searching for documents. The second use of the “embedding LLM” is to convert the “LAR drafter” “question” into a “question text embedding.” Next, the “question text embedding” is compared to every document in the vector database. Using embeddings for searching was performed with cosine similarity (Singhal 2001). This results in the most relevant “retrieved pages of TS, UFSAR, LARS.” The retrieved pages are those with the highest similarity to the question, hence the name “similarity comparison.” The second LLM in the gray box called “pre-trained instruct LLM” was trained on billions of documents gathered from the internet and denoted “huge training data set.” This second LLM has 3 inputs: the training documents, the question, and the retrieved pages. The second LLM then responds to the LAR drafter question to produce information that the drafter needs to complete the LAR.

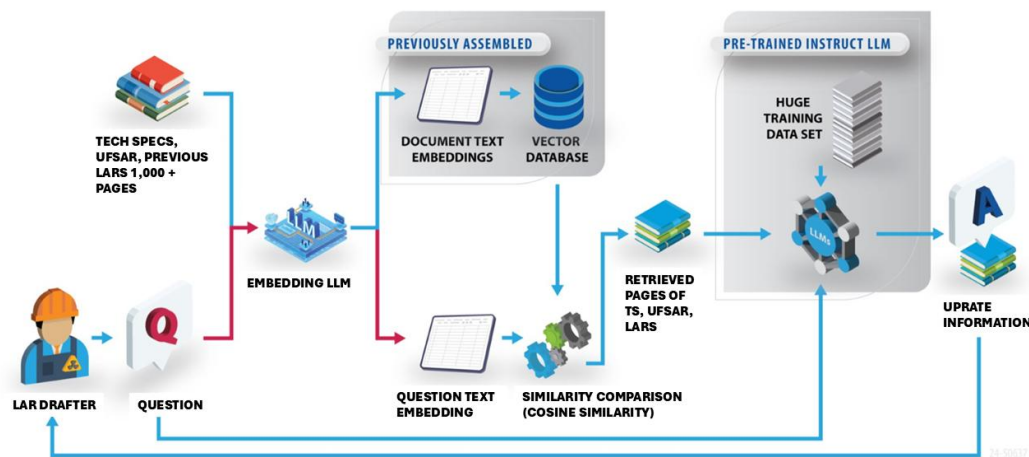


Figure 5. RAG flowchart.

<sup>i</sup> <https://huggingface.co/atomic-canyon/fermi-1024>

### 3.1.2 Fine-Tuning LLMs

Fine-tuning is the process of training a previously trained LLM further on a smaller and more specialized dataset. It allows the model to perform more specific tasks, such as generating information and documents for power uprates based on the TS and associated documents. Fine-tuning typically improves a model's performance in a targeted application while allowing it to retain the understanding of language and reasoning it learned in the initial pretraining. For this task, an abundance of data from the NRC ADAMS is available in the TS, associated documents, and even other nuclear-specific documents to train the LLM. The goal of this work was to use a fine-tuned LLM to build a nuclear expert model that can distill knowledge into a LAR.

Fine-tuned LLMs have been used in the nuclear industry already, including for fault-tree analysis (Xiao et al. 2024), for nuclear fuel cycle classification (Burke et al. 2021), and for classifying safety-related station condition records at nuclear power stations (de Costa et al. 2024), and both a nuclear domain LLM (Bouhoun et al. 2024) and a nuclear-fusion-dialogue LLM (Wang et al. 2025) have been developed. Like RAG, fine-tuned LLMs have also been applied in related domains such as technological (Kan et al. 2024), medical (Dorfner et al. 2025), and legal (Wei et al. 2023). Bouhoun et al. (2024) noted a noticeable gap in the knowledge that current LLMs have for nuclear, suggesting a need for the development of specialized models in this field. This observed necessity has been addressed by this work through the fine-tuning of a LAR-specialized model.

Before an LLM can be fine-tuned, preprocessing must be carefully performed to ensure the data is in an optimal condition for training so the model will competently generate results that meet expectations. The first step involves extracting textual data from PDFs of the LARs. This phase includes parsing and analyzing the text to determine optimal line separations that ensure critical grammatical features are preserved. Paragraph and sentence structures must be preserved across page breaks. Another aspect of this phase is capturing table structures in a way that the LLM can learn from and thus reproduce equivalent structures in its output. Deciding on a means for interpreting tables and reflecting them as consistent textual constructs is important in this process because tables are common in the ADAMS training data. Other textual preprocessing includes applying various regular expressions, textual classifiers, and other techniques to identify and redact or replace certain features that could confuse the model or bias it in an unsatisfactory way. For instance, in this work it was important to redact email headers and footers, which are problematic features that appear in thousands of ADAMS documents. Also, removing repetitive text such as page numbers, headers, and footers ensured that the text flowed seamlessly and prevented biasing the model into reproducing unhelpful output. Finally, before an LLM can be trained in this work, the text data must be represented in a form the model can understand, which means further preprocessing to convert the text to tokens that conform with the vocabulary the model understands. The text tokens in each document are then grouped into chunks comprising 4,096 tokens each. These batches of tokens are then used to train the LLM.

The Microsoft Phi-3.5 mini instruct model<sup>j</sup> was fine-tuned for this work. All the weights in the model were unlocked during fine-tuning (often called full fine-tuning). There are 3.82 billion parameters, and 100% of the parameters were updated during training. This contrasts with training a small fraction of the weights. This effort required Idaho National Laboratory's (INL's) high-performance computing resources. It typically took 3 weeks to fine-tune the LLM on the NRC ADAMS database; this was performed about 20 times<sup>k</sup>. The LLM was also trained for a few hours to fine-tune it on the TS and associated documents. The potential advantage of full fine-tuning is higher performance metrics.

---

<sup>j</sup> <https://huggingface.co/microsoft/Phi-3.5-mini-instruct>

<sup>k</sup> Sometimes the training was halted early when problems were noticed.

### 3.1.3 Structuring

Structuring involves organizing data from unstructured sources such as documents, figures, and tables. Structured text is defined as text data formatted in standardized, machine-readable formats that explicitly follow rules, allowing for consistent extraction, analysis, and automated processing. In contrast, unstructured text is defined as textual data that lack a predefined, standardized format; it is typically presented as plain, free-form language without explicit structural elements, which makes automated processing and analysis challenging. An example of structured data is a database that contains a list of NRC requirements and information pertaining to how plants meet the requirements, as shown in Table 1. This table contains structured information from hundreds of pages of LARs that were unstructured.<sup>1</sup> Since most data in the nuclear industry are unstructured, Al Rashdan et al. (2024) proposed using structuring for nuclear knowledge. Brach, Košťál, and Ries (2025) observed that research in structuring offers “new possibilities for automated structured data generation across various domains, from medical records to technical documentation, potentially transforming the way organizations process and utilize unstructured information.”

Table 1. Sample of structured data from unstructured NRC-approved LARs (unstructured data are not shown here).

NRC Requirement	Byron	Joseph M. Farley	Other Plants
Effect of power uprate on emergency diesel generators	The onsite (emergency) alternating current (AC) power system for each unit consists of two diesel generators...	The onsite emergency AC power supply for Farley Units 1 and 2 consists of five diesel generators...	...
Effect of power uprate on station blackout equipment (SBO)	The MUR power uprate has no impact on the current SBO coping duration of four hours ...	Evaluation of the SBO licensing bases revealed that the licensing bases for SBO continue to be met under MUR-PU conditions ...	...
Effect of power uprate on environmental qualification of electrical equipment	The Byron...electrical equipment will function, as required, under normal, abnormal, and/or accident environmental conditions...	The equipment qualification program equipment accommodates MUR-PU conditions without exceeding electrical equipment qualification margins...	...
Effect of power uprate on grid stability	Two grid studies have been completed to support the proposed uprate...	A stability analysis did not identify any stability-related transmission constraints...	...

### 3.1.4 Few-Shot Prompting

Few-shot prompting demonstrates to an LLM how it should respond using examples. Few-shot prompting is closely related to RAG but differs in that RAG finds relevant documents but does not give an example of how to respond. Similar prompts were used for both few-shot prompting and RAG. The bold font in the following highlights the differences. For few-shot prompting, the prompt was: “Here are 3 document excerpts: [excerpt 1], [excerpt 2], and [excerpt 3]. Create a new document excerpt **like these** specific to the Joseph M. Farley nuclear power plant.” For RAG, the prompt was: “Here are 3 document

<sup>1</sup> However, the presence of section headers means that the data is technically semistructured.

excerpts: [excerpt 1], [excerpt 2], and [excerpt 3]. Create a new document excerpt **on the topic of [topic]** specific to the Joseph M. Farley nuclear power plant.” Sometimes RAG gave an example in the retrieved excerpts, but this is not the way RAG typically works. Typically, RAG retrieves documents with text that is highly similar to the topic that are not example outputs, but few-shot prompting literally shows what an acceptable output is like.

Few-shot prompting has been used in technical fields that are of comparable difficulty to nuclear regulatory documents such as medical (Nachane et al. 2024), legal (Yu, Quartey, and Shilder et al., 2022), and technological (Shao et al. 2024). Yu, Quartey, and Shilder (2022) found that using approaches with few-shot and fine-tuning led to good and consistent results. The current research expands on previous efforts by evaluating few-shot prompting effectiveness for creating LARs. This work found that few-shot prompting may perform better than fine-tuning when there are very few examples, because fine-tuning attempts to memorize the few examples, whereas few-shot prompting relies on the comprehension capabilities of the model to learn the examples. In this study, fewer than 10 examples of previously accepted MUR LARs from the last 10 years were used.

### 3.1.5 Structured Few-Shot Prompting

A novel contribution of this work is combining structuring (Section 3.1.3) and few-shot prompting (Section 3.1.4) into what is called “structured few-shot prompting” (SFP).

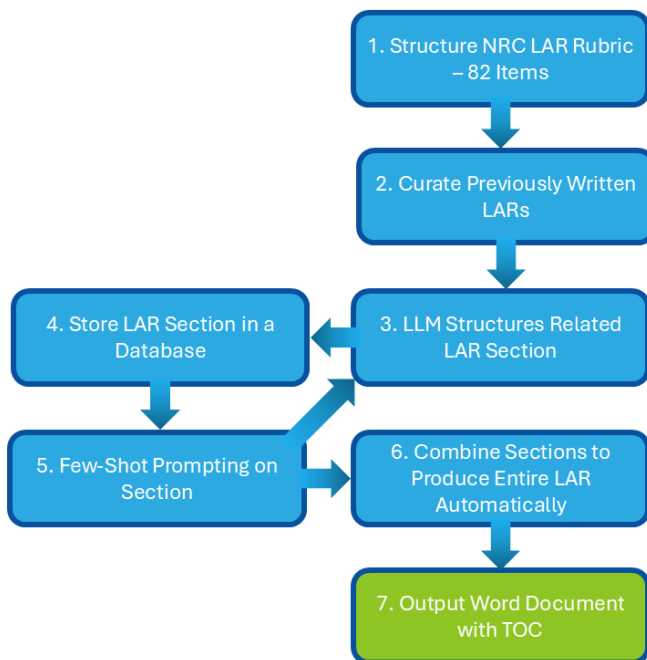


Figure 6. SFP process for entire LAR auto-generation.

Figure 6 illustrates the seven-step SFP process for drafting a LAR. First, the 82 sections of the NRC rubric are converted into a machine-readable format from the U.S. NRC (2002) with structuring. Second, several LARs are curated, which requires some SME effort to deduce which previously written LARs are most relevant (e.g., all MUR updates in the last 10 years). Third, each LAR is then structured by a structuring LLM (e.g., AiVA long context<sup>m</sup>). Fourth, the structured sections are stored in a database (see Section 3.1.3 for an example of structured data in a database). Fifth, few-shot prompting is then applied, which means similar LARs are used as templates (i.e., shots) to generate an unseen LAR. RAG, fine-

<sup>m</sup> AiVA uses GPT-4o-mini for long context. In this case 144 pages were input requiring the long context model.

tuning, and pretrained knowledge<sup>n</sup> provide the ability to autogenerate the LAR based on plant-specific data. Thus, the two sources of data are combined: the previous LARs and the plant-specific TS and associated documents. This process is repeated until the entire LAR is created, hence the cycle for Steps 3, 4, and 5. Sixth, the autogenerated sections are combined. Finally, a Microsoft Word document is created automatically with a table of contents. The process in Figure 6 is termed “structured few-shot prompting” because document sections are structured (Steps 1 and 3) for use by few-shot prompting (Step 5). The goal is to generate entire documents by organizing them according to the section headers that will be structured into a database.

## 3.2 Evaluation Method and Model Selection Rationale

### 3.2.1 Comprehension Measurement in Nuclear Texts

Comprehension scores are essential for assessing how well LLMs understand licensing text, this work proposes comprehension measurement in nuclear texts (CMINT). The evaluation method involves using weaker LLMs, such as the AiVA long context model, to extract facts from a set of documents. These facts are then used to create questions where the facts serve as the answers. The questions are presented to the evaluated model, referred to as the student model. The student model's answers are then compared to the answers generated by the weaker models. This comparison is conducted by a stronger model, like the Meta LLaMA 3.3 70B model. The stronger model provides a score to determine the student model's performance. This evaluation process is illustrated in Figure 7. In this effort, 450 questions and corresponding answers were extracted from each page of a licensing document. These 450 facts establish the derived standard for testing a student LLM.

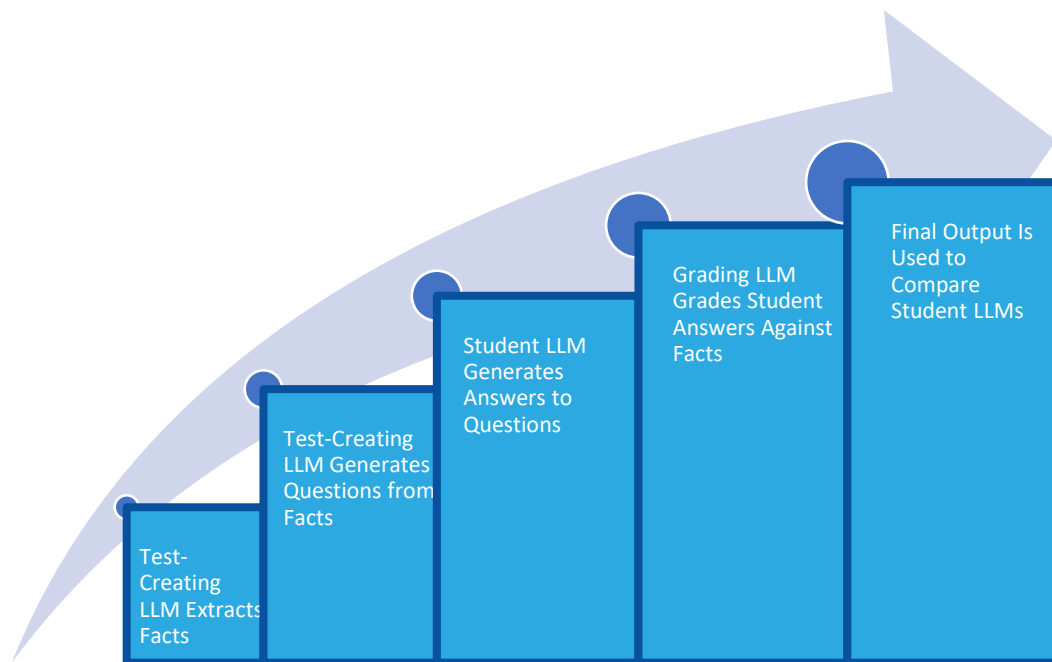


Figure 7. Process diagram for CMINT.

The method measures the LLM’s comprehension much like a reading test measures a human’s understanding. To better illustrate this process, a demonstration of the prompts of the LAR (150 pages) is provided below:

<sup>n</sup> Pretrained knowledge is the information that was provided—i.e., by Microsoft or OpenAI—to the LLM before INL received it.

Interaction with test-creating LLM:

Prompt 1: Give me 3 facts from each LAR page.

Response 1: 1. LEFM uncertainty calculation is based on ASME Performance Test Code (PTC) 19.1.

2. ...

3. ...

450. ...

Prompt 2: Turn these into questions.

Response 2: 1. On what standard is the LEFM uncertainty calculation based?

2. ...

3. ...

450. ...

Interaction with student LLM:

Prompt 3: 1. On what standard is the LEFM uncertainty calculation based?

2. ...

3. ...

450. ...

Response 3: 1. The LEFM (Leading Edge Flow Meter) uncertainty calculation is typically based on the ASME (American Society of Mechanical Engineers) Performance Test Code (PTC) 19.1-1985, "Measurement Uncertainty." This standard provides guidelines for calculating the overall uncertainty of measurements, including those related to flow meters used in nuclear power plants.

2. ...

3. ...

450. ...

Interaction with grading LLM:

Prompt 4: 1. How close are these two answers to a question:(1) [LLM ANSWER.] (2) [FACT]? Please only give a single number that is a number between 0 and 100.

2. ...

3. ...

450. ...

Response 4: 1. 95.

2. ...

3. ...

450. ...

### 3.2.2 Model Selection

Choosing which model to use primarily depends on whether to use local models or cloud-based models, based on the security nature of the data. If the TS and associated documents are publicly

available, then data security may not be an issue and cloud-based models may be used. If non-public TS and associated documents are used, then data security or privacy could be an issue, and one may choose to use a local model. A cloud-based model may generate security or privacy concerns but is usually more sophisticated than a local model. For this effort, Phi-3.5 was frequently used for the local model while INL’s internally hosted generative-AI model AiVA, which is based on ChatGPT-4o, was used as the cloud-based model. Cloud based frontier models such as OpenAI o1 were only experimented with on a limited basis due to limited access at the time of writing.

There is a distinction between LLMs, one type is the instruct LLM and the other type is the embedding LLM. The instruct LLM takes ordinary language inputs and the embedding LLM finds relevant documents. The instruct LLMs used include: AiVA, Phi-3, Phi-3.5 fine-tuned, Phi-3.5 and LLaMA-3.2 3B. None of these models were fine-tuned except for Phi-3.5 fine-tuned that was developed at INL. The embedding LLMs used were: Atomic Canyon Fermi-1024, TF-IDF, and MiniLM fine-tuned. All of these models were not fine-tuned except for MiniLM. TF-IDF is not actually an LLM; instead, it finds documents by word matching. More details on each LLM are given in Table 2. All local instruct models were approximately 3 billion parameters; this was the maximum LLM parameter count that would be efficiently trainable on INL high-performance computing. Embedding LLMs typically are much smaller, in the millions of parameters range.

Table 2. Characteristics of LLMs used in this effort.

Model \ Characteristic	Cloud	Local	Instruct	Embedding	Fine-Tuned	Parameter Count
AiVA	X		X			100B+
Phi-3 Default		X	X			3.8B
Phi-3.5 fine-tuned		X	X		X	3.8B
Phi-3.5 Default		X	X			3.8B
LLaMA 3.2 3B Default		X	X			3.2B
Atomic Canyon Fermi-1024		X		X		110M
TF-IDF		X		X	N/A	N/A
MiniLM		X		X	X	22.7M

## 4. USE CASES

Five use cases of increasing complexity were identified to establish the feasibility of using LLMs to generate LARs; these use cases are listed in Table 3 along with the LLM methods applied to each one. Although each of the use cases focus on power uprate LARs, it may be possible to extrapolate the results to other types of LARs. Use Case 1 concerns the Joseph M. Farley NPP with a focus on the CRDM section of the LAR. Use Case 2 concerns the Byron and Braidwood plants with a focus on the CRDM section of the LAR. Use Case 3 demonstrates comprehension of the entire technical content (attachment 4) of the LAR for Farley. Together, Use Cases 1, 2 and 3 evaluate the *knowledge capture* of the LLMs, which means the ability to reproduce elements of the LARs. Use Cases 4-5 are focused on *content generation*. Use Case 4 recreates one section of the CRDM for Farley, with a focus of demonstrating the human-machine synergy possible with RAG. Use Case 5 demonstrates use of each method to generate an entire LAR for Farley—including SFP, which was specifically designed to be used with the generation of the technical content of an entire LAR. In Table 3 the methods used are provided. Note that SFP was only used for one use case, this was the most complex use case that required the SFP method specifically. Use Cases 1 and 2 were completed prior to the development of a fine-tuned instruct LLM and therefore do not have an X in the fine-tuning method.

Table 3. LLM methods applied to specific cases.

Use Cases Methods	Knowledge Capture			Content Generation	
	(1) Farley CRDM	(2) Byron and Braidwood CRDM	(3) Farley Entire LAR	(4) Farley CRDM Section	(5) Farley Entire LAR
RAG	X	X	X	X	X
Fine-Tuning			X	X	X
SFP					X
Pretrained	X	X	X	X	X

Table 4 lists the specific embedding and instruct<sup>o</sup> LLMs used for all use cases. It is important to note that due to the rapidly evolving nature of LLMs, different models were used throughout the project, allowing researchers to use the most capable system available at the time of evaluation. Each use case explored the two major classes of instruct models—local and cloud-based. The cloud-based model was always AiVA, and the local model was Phi-3<sup>p</sup> (small) for Use Cases 1 and 2, and Phi-3.5 (mini) for Use Cases 2-5 (Use Case 2 was performed with both versions).

All use cases used the TF-IDF embedding model because of its effectiveness, as noted in Al Rashdan et al. (2024). Additionally, the first and second use case experimented with the MiniLM fine-tuned model because of its effectiveness noted in Al Rashdan et al. (2024). Use Cases 3 and 5 switched to the newly released Atomic Canyon Fermi-1024 model because it was specifically designed for the nuclear sector and incorporates specialized nuclear vocabulary; therefore, it was assumed that the Atomic Canyon Fermi-1024 model would perform better than MiniLM. To test the improvement of fine-tuning Phi-3.5

<sup>o</sup> An instruct model can converse in natural language, while an embedding model cannot.

<sup>p</sup> <https://huggingface.co/microsoft/Phi-3-small-8k-instruct>

for drafting LARs, Use Case 5 also tested the Phi-3.5 default model. Meta’s LLaMA 3.2 3B was also implemented as an additional comparison for the benefits of fine-tuning of Phi-3.5 for Use Cases 3 and 5. This particular model was chosen because it was developed at around the same time as the Phi-3.5 model, and it had similar parameter counts. Meta’s LLaMA 3.3 70B was used in Use Case 3 to perform an ancillary function (i.e., testing other LLMs), it was not evaluated for knowledge capture nor drafting a LAR and therefore is not listed in the table.

Table 4. Instruct and embedding LLMs evaluated during each use case.

Use Case LLMs	Knowledge Capture			Content Generation	
	(1) Farley CRDM	(2) Byron and Braidwood CRDM	(3) Farley Entire LAR	(4) Farley CRDM Section	(5) Farley Entire LAR
Instruct LLMs					
AiVA	X	X	X	X	X
Phi-3 Default	X	X			
Phi-3.5 Fine-Tuned		X	X	X	X
Phi-3.5 Default					X
LLaMA 3.2 3B Default			X		X
Embedding LLMs					
Atomic Canyon Fermi-1024			X		X
TF-IDF	X	X	X	X	X
MiniLM Fine-Tuned	X	X		X	

## 4.1 Application of Evaluation Methods to Use Cases

Four different evaluation methods were used to evaluate the quality of the LLM-generated content: SME evaluation, LLM-assigned scores, statistical tests, and cosine similarity. The first evaluation method was performed by SMEs with years of experience working with NPP documentation. Their focus was to evaluate how closely the AI-generated content aligned with the NRC-approved documents. The chosen SMEs also had LLM experience and were familiar with common LLM pitfalls, such as hallucinations. For the second evaluation method, CMINT (introduced in Section 3.2.1), an LLM judged how close an answer was to the solution and assigned a score. The LLM was instructed to give a score from 0 to 100, with a higher score indicating the answer was closer to the solution. The third evaluation method

measured if the results were either statistically significant or not statistically significant. The last evaluation method used cosine similarity to measure how similar two documents are. The cosine similarity scores generally ranged from 0 to 1 with a higher score indicating a greater degree of overlap between the LLM-generated content and the NRC-approved documents.

Table 5 shows how each evaluation method was assigned to the different use cases. Use Cases 1, 2 and 4 were only evaluated by an SME. The SME was the most appropriate choice to catalog the nature of the comprehension failures in Use Case 1. SME review was selected for Use Case 2 due to the relatively limited scope and volume of the material, which did not necessitate the use of an automated review method. For Use Case 3, an automated means was used to comprehensively evaluate an LLM's, evaluating 450 LAR facts, as discussed in Section 3.2.1. Use Case 4 was also evaluated by an SME because the workload was manageable. A combination of manual and automated evaluation methods was used for Use Case 5: (1) an SME reviewed the quality of the responses, (2) cosine similarity of an LLM-generated LAR to approved LARs measured the quality of the responses, and (3) statistical tests. The reason for 3 methods is because the workload of Use Case 5 was large and consisted of comparing approximately 150 pages of a LAR that was automatically drafted to an actual LAR of the same approximate length. An SME spot-checked this use case to identify limitations during the autodrafting of an entire LAR.

Table 5. Evaluation methods applied to each use case.

Use Cases Methods	Knowledge Capture			Content Generation	
	(1) Farley CRDM	(2) Byron and Braidwood CRDM	(3) Farley Entire LAR	(4) Farley CRDM Section	(5) Farley Entire LAR
SME Evaluation	X	X		X	X
LLM-Assigned Scores			X		
Cosine Similarity					X
Statistical Tests					X

## 4.2 Use Case 1: Farley CRDM Knowledge Capture

RAG is expected to produce facts to help the LAR drafter locate information from a collection of documents (i.e., a corpus). Use Case 1 was developed to test the ability of RAG to accurately extract information from a corpus in support of a power uprate. The 2020 Farley NPP power uprate was randomly selected from recent MUR uprates (ensuring that no selection bias was present) for this demonstration. The corpus comprised nearly two dozen different PDF documents totaling about 7,000 pages from the following three references for the Joseph M. Farley NPP: (1) TS, (2) STS bases, and (3) an UFSAR. Note that the actual LAR was not included in the corpus, as the ultimate goal is to show that RAG can be used to create the LAR. Additionally, other LARs were not used because they do not contain Farley-specific information necessary for this use case.

A passage related to the CRDM was taken from the STS bases and manually converted into a question as seen below. This passage was then successfully retrieved using the embedding LLM in the

RAG process, which is shown in the retrieved context section below. A prompt was then used to convert the question and context into a query. The Microsoft Phi-3 (Abdin et al. 2024) LLM answered the query to formulate a response. The example below is one that worked correctly; more complex scenarios are demonstrated in the next section along with the failures that occurred.

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 (embedding LLM): *...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]*

Response (instruct LLM): *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...*

Having established the ability to properly retrieve information related to a power uprate, the focus shifted to identifying the different types of knowledge capture failures that could be expected. To accomplish this, a single sentence was selected from the associated uprate LAR: "The CSS [containment spray system] operates to limit peak containment pressure to less than the design pressure of 54 psig during a LOCA [loss of coolant accident] or MSLB [main steam-line break] in order to maintain containment structural integrity." This sentence was selected because it contains many facts that pose problems to an LLM attempting to draft a LAR automatically. Furthermore, some of these facts are not immediately obvious to a LAR drafter—e.g., the containment peak (maximum) design pressure of 54 pounds per square inch gauge (psig). It was important to include non-obvious facts in this analysis to elicit knowledge capture failures. Ultimately, the LLM may function as a junior LAR drafter that provides recommendations to the human LAR drafter, and it is critical to understand the potential pitfalls.

The goal of this exercise was to use an advanced generative-AI technique (i.e., RAG) to absorb the information in the TS and associated documents to generate the facts contained in this single sentence. The RAG selected the appropriate pages of the TS, STS, and UFSAR. Next, a prompt was devised to convert the retrieved pages into something a LAR drafter will understand. Finally, the AI-generated facts were compared to the facts contained in the single sentence using SME review.

In this demonstration, four failure types (FTs) were identified in the RAG system and are listed in Table 6. These failures are similar to the seven RAG failure points observed in Barnett et al. (2024). Neither the FT nor Barnett's failure point lists are likely to be comprehensive, rather they are specific observations from the failures identified in this work and from the case studies of Barnett et al. (2024).

Table 6. The failure types for RAG.

Failure Number	Explanation
FT 1	The retriever retrieves the incorrect document (Section 4.2.1).
FT 2	The LLM fails to verify source text accuracy (Section 4.2.2).
FT 3	The LLM uses its internal knowledge to answer the question instead of using the retrieved context (Section 4.2.3).
FT 4	The LLM hallucinates and gives incorrect facts (Section 4.2.4).

For each failure type, the boldface used for a portion of the LAR fact, question, or response emphasizes the intent of the LAR drafter question.

#### 4.2.1 Failure Type 1—Incorrect Retrieved Document

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 Farley nuclear plant containment?*

Retrieved context: *[Irrelevant document]*

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]*

Response: *I don’t know.*

The relevant portion of the LAR fact in this case is that the containment design pressure is 54 psig, therefore an attempt was made to reproduce this fact using RAG to mimic the question a LAR drafter may have. The LLM response did not retrieve the correct UFSAR document. Therefore, an incorrect answer was found because the RAG engine failed to identify the correct document for the question. However, AiVA, a non-RAG approach, was experimented with and it responded correctly. The use of AiVA will be explored as an alternative method in Section 4.7. Bear in mind that because AiVA did not utilize RAG for its answer and instead relied on the training documents that went into training AiVA, there are no reference documents that justify the LLM response.

#### 4.2.2 Failure Type 2— Failure to Verify Source Text Accuracy

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: ***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: *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]

Response: *The peak design pressure for the containment of the nuclear power plant is **54 psig**, as indicated in table 6.2-6.*

The response correctly identified that the containment design pressure is 54 psig. However, the LLM incorrectly reports that Table 6.2-6 in the UFSAR (the specific document was denoted via metadata in the returned response, but is not seen in the text here) contains the peak design containment pressure; this is not the case. This failure is due to noise in the dataset—the document itself incorrectly indicates that the table contains the value. The LLM should have verified the statement by checking that the table contained the relevant value. Hence, the LLM returned some incorrect information, but did not hallucinate. This is an indication that a prospective LAR drafter would need to review the RAG system’s outputs to ensure their accuracy before using them. Nevertheless, the correct value was retrieved from page 6.2-6 in Chapter 6 of the UFSAR.

Note that the wording of the question is different from that used in Section 4.2.1, resulting in a different output. This was done to evoke a different response, potentially demonstrating a different failure type. The question in this section is not actually a question, and the failure might have been avoided with a different prompt. However, this type of failure is expected to occur as a user interacts with the system, even if they use better prompts.

### 4.2.3 Failure Type 3—Internal LLM Knowledge Used Instead of Retrieved Knowledge

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: *Does the CSS operate to limit peak containment pressure to less than the design pressure?*

Retrieved context: [Irrelevant text]

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]*

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....*

The relevant portion of the LAR fact in this case is that the CSS operates to limit peak containment pressure to less than the design pressure of the containment. This fact was posed as a question, and the dataset was searched using RAG. Here the retrieved context is irrelevant to the question per an SME. There were low cosine similarities of the question to the text data, and therefore the RAG pipeline failed to retrieve relevant documents. This can occur because the answer is not contained in the reference documents, or the retrieval system does not correctly understand the question or understand the reference documents. For this specific example, the correct document is page 6.2-37 in Chapter 6 of the UFSAR, which contains the following passage: “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 the correct document page. A more significant issue is that the model depended on its internal knowledge from training rather than use the reference document to respond to the prompt. This is problematic because RAG is designed to comprehend and answer based on reference documents, not on pretraining knowledge.

Rewriting the question to *Does the containment spray system (CSS) operate to limit peak containment pressure to less than the design pressure?* improved the RAG’s ability to search the UFSAR because the page containing the information uses the full unabbreviated name for CSS instead of the acronym. However, when this modified question was used, the RAG still failed to identify the correct document. This failure is to some extent understandable because the RAG had to search thousands of pages to locate the one that contained this fact.

#### 4.2.4 Failure Type 4—LLM Hallucinations

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: *Does the CSS increase or decrease peak containment pressure during a LOCA or MSLB.*

Retrieved context: *...The inadvertent actuation of the Containment Spray System was analyzed to determine the resulting **reduction** in containment pressure...*

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]*

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.*

The relevant portion of the LAR fact in this case is that the CSS limits (decreases) the peak containment pressure. This question is nearly the same as the one given in Section 4.2.3, but it is worded slightly differently to demonstrate a different type of failure—i.e., hallucinations. Interacting with the RAG system has revealed that it sometimes gives false information when responding to questions. In this example, MSLB stands for “main steam-line break”; however, the LLM hallucinated the meaning as “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 included in the hallucinated response. While this hallucination was due to an acronym, not all acronyms are hallucinations.

### 4.3 Use Case 2: Byron and Braidwood CRDM Knowledge Capture

For Use Case 2, the RAG system was tested—based on the Byron and Braidwood (B/B) NPP power uprate application—to compare the impact of using automatically retrieved documents vs manually retrieved documents on the prompt response. For the chosen question, the relevant section of the TS and associated documents was identified via both a human-provided lookup and an automated lookup. Then an AI-generated response was given to the LAR drafter’s question based on each set of retrieved pages. As a result, the LLM was tested independently of the retrieval mechanism. If the LLM can produce a correct response based on non-specific human-retrieved content such as a page range, this would mean that the LLM can reduce the workload by sifting through the details of multiple pages in a moment. A reduced workload results in more efficient licensing processes for power uprates.

B/B was chosen instead of the Farley NPP because there are specific parameters mentioned in the CRDM portion of the Byron LAR for but not for Farley. These parameters form the basis for this use case. Ten pages on the subject of control rod drive mechanisms (CRDM) were retrieved using automated RAG. A separate manual search of the B/B UFSAR identified 14 pages that potentially contained the information needed by the power uprate application drafter. Both the manually retrieved and automatically retrieved documents overlapped, and more importantly both contained the relevant page: 3.9-82.

The manually retrieved and automatically retrieved information were separately fed to the generative-AI portion of the RAG system. The LLM then extracted the design pressure and temperature (i.e., parameters) needed by the power uprate application drafter. The LLM was successful for both the manually retrieved information and the automatically retrieved information. The prompt used for both the manually and automatically retrieved information is based on the actual B/B LAR with the parameter values of interest removed.

*Prompt: Please fill in the parameters [parameter 1] and [parameter 2] from the following LAR: [The control rod drive mechanisms (CRDMs) use electro-magnetic coils to position the rod cluster control assembly (RCCA) within the reactor core. The updated design conditions (design parameters and nuclear steam supply system design transients) were reviewed for their impact on the existing CRDM design basis analyses. CRDMs are subjected to  $T_{cold}$  temperatures and reactor coolant system (RCS) pressures. These are the only design parameters considered in the CRDM evaluation. The maximum  $T_{cold}$  from the MUR power uprate design parameters for any case is 555.1°F. The maximum  $T_{cold}$  from the analysis of record is 558.4°F. As a result, the analysis of record remains bounding and applicable. No changes in RCS design or operating pressure were made as part of the MUR power uprate. The temperature and pressure transients are unaffected by the MUR power uprate. Since the transients are unchanged, they do not alter the stress results or the bending moment allowables. Therefore, the original transient analysis remains bounding and applicable to the MUR power uprate conditions. The stress intensity limits are based on a design temperature of [parameter 1] and a pressure of [parameter 2], which are unchanged by the MUR power uprate. Updated seismic and loss of coolant accident loads remain less than the allowable loads provided in the analysis of record. The code of record is listed in Section IV.1.D and remains unchanged. (U.S. NRC ADAMS Accession No. ML111790042, 2011)] Please use these details from the following UFSAR to generate the parameters above. [retrieved context.]*

Because the original LAR has this information, it is possible to check the accuracy of the RAG system. The correct answer for Parameter 1 is 650°F, and the correct answer for Parameter 2 is 2,500 psia per the B/B power uprate application.

### **Generation based on manually retrieved documents**

The manual search identified 14 pages of the UFSAR, which were then provided to the cloud-based model, AiVA; AiVA then returned the correct parameters. Manual retrieval for the RAG system does not guarantee the system will work as expected, but it does remove a possible failure to retrieve the correct document, such as the failure seen in Section 4.2.1. This is relevant because often the human may know approximately where the content is located in the documentation, but needs the LLM to identify the specific information. Only the AiVA model coupled with embedding model MiniLM was successful; AiVA coupled with the TF-IDF embedding failed, as did Phi-3 and Phi-3.5 fine-tuned models with either MiniLM or TF-IDF.

*Retrieved context: [14 manually selected relevant pages of NRC ADAMS Accession Number ML24103A228, pages 3.9-80 to 3.9-92.]*

*Response: Based on the information provided in the UFSAR, the parameters for the stress intensity limits are as follows:*

*[parameter 1]: 650°F (This is the design temperature for the pressure vessel components)*

*[parameter 2]: 2500 psia (This is the design pressure for the pressure vessel components)*

*So the complete sentence would be:*

*“The stress intensity limits are based on a design temperature of 650°F and a pressure of 2500 psia, which are unchanged by the MUR power uprate.”*

### **Automated Retrieval for RAG**

The automated retrieval system for RAG was used to locate 10 pages from 10,530 pages. This process also resulted in a correct answer, as shown here. As seen in Section 4.2.1, automated retrieval does not guarantee the correct result, but in some cases—like the one shown here—it can be successful. Notably, RAG provides references that a human can see. This is a notable advantage in that the references could help the human verify that the information is correct while drafting a power uprate request. Only the AiVA model coupled with the MiniLM embedding was successful; the Phi-3 and Phi-3.5 fine-tuned models failed. All models coupled with the TF-IDF embedding failed as well.

Retrieved context: *[10 AI-retrieved pages from 27 documents of TS, TS bases, and UFSAR, including page 3.9-82 that contains the correct information.]*

Response: *Based on the provided documents, the parameters you are looking for are:*

- *[parameter 1]: 650°F*
- *[parameter 2]: 2500 psia*

*These values are derived from Document 4, which states that “The mechanism internals are designed to operate in 650°F reactor coolant. The pressure vessel is designed to contain reactor coolant at 650°F and 2500 psia.”*

The LLM identified Document 4 as the source of information, which corresponds to the 4th page returned from the automated retrieval system, which was verified to be page 3.9-82 in the USFAR. The LLM replied that the values are derived from Document 4 because the actual formatting of the retrieved context is: *Document 1: [document 1 content], Document 2 [document 2 content],...Document 10: [document 10 content].*

#### **4.4 Use Case 3: Farley Entire LAR Knowledge Capture**

LLM comprehension was measured because of the known limitation of hallucinations. Hallucinations are phenomena in which LLMs perceive nonexistent or irrelevant patterns in the data, often returning results that appear accurate but are factually incorrect. Hallucinations are often subtle and difficult to detect (e.g., providing a well formatted citation of a publication that does not exist). 450 derived standard facts were extracted from the Joseph M. Farley LAR (see Section 3.2.1) to create an automated comprehension test for the LLMs.

This exercise was designed to prevent two sources of bias. The first potential bias was that a model could grade its own answers. This was addressed by not using the grading LLaMA 3.3 70B model as a student model. The second potential bias was that the test-creating LLM should not be the grading LLM. The LLM that generated the questions may implicitly “expect” specific responses, penalizing valid answers. To prevent this, different independent models were used for test creation and grading. It is important to note that the goal was not to evaluate the beliefs of the test-writing or grading models, but to compare the derived standards (i.e., LAR facts) to the student model submissions. Thus, using separate models for generating the test and performing the grading effectively avoided this bias.

The total range of comprehension scores for each question was 0-100. It was also important to establish a baseline to compare the models against. The echo response baseline repeated the question back as the answer and was graded for accuracy, realizing only an average comprehension score of 5, as shown in Table 7. This shows that the common test-taking strategy of repeating the question for partial credit does not result in good scores. Overall, the cloud-based model AiVA performed best. Of the local models, the fine-tuned Phi-3.5 model, which had approximately 2 million NRC ADAMS documents as its training set, performed only slightly better than LLaMA 3.2 3B.

Table 7. Average LLM comprehension score on the 450 test questions.

Embedding LLM \ Instruct LLM	No RAG	RAG TF-IDF	RAG Atomic Canyon	Mean	Multiple over Baseline
AiVA	58.7	52.7	46.3	52.6	10.52 ×
Fine-Tuned Phi-3.5	49.4	49.8	43.9	47.7	9.54 ×
LLaMA 3.2 3B	32.9	43.9	41.3	39.4	7.88 ×
Echo Response Baseline	5.0	N/A	N/A	5.0	1.00 ×

### Memorization, Overfitting, and Data Contamination

It is clear the LLMs have not memorized the ADAMS database because their LAR comprehension scores were below 60%. LLMs are presumably trained on high-quality data such as the ADAMS database that contains LARs, so the scores should have been closer to 90%. This is a limitation of the LLMs. However, there is a related limitation that does not appear to affect the LLMs. The low comprehension scores suggest that the LLMs have not *overfit* the data. Overfitting is learning the details and noise of the training data and performing well on it but not performing well on the testing data. An example is person memorizing homework examples and answers but not concepts, and then failing the exam because the questions differed from the examples they had memorized, and they lacked the conceptual knowledge to generate novel answers.

Another closely related limitation is *data contamination*, which occurs when models are trained on test data (Magar and Schwartz 2022). It is assumed that all the LLMs tested have been trained on ADAMS documents. But the limitation is that the LLMs are tested on ADAMS documents. This is potentially an unfair test because the LLM can memorize the testing data and then output it. This may result in misleading high-performance metrics that may not reflect the LLM’s true ability to generalize to unseen data. However, this limitation is not significant because the LLMs did not memorize the LARs, as shown in their lower-than-60% comprehension scores.

In this situation, it is not possible to separate the training and testing data for two reasons: (1) there were no power uprates approved after the LLM training date, and (2) all power uprate documents are public. To understand these reasons, keep in mind that the goal was to obtain test data that the models were not trained upon. In Reason 1, if there were power uprates available after the training date, then using the newest power uprate for testing would prevent this problem. In Reason 2, if there was a non-public power uprate, then using this non-public power uprate would prevent this problem. In conclusion, it is impractical to separate training and testing data given the current availability of data. A possible workaround is a future industry partner collaboration that provides access to non-public data. Furthermore, in related research Brown et al. (2020) reported that memorization and data contamination did not have a significant effect on their results. Additionally, they reported that there was no correlation between the amount of contamination and performance. It appears that data contamination is a prominent problem that affects more than this work (Magar and Schwartz 2022, and Dodge et al. 2021).

## 4.5 Use Case 4: Drafting the Farley Uprate Application CRDM Section

In this section, the quality of the LLM-generated CRDM section is compared to the actual CRDM section of the Farley 1.7% uprate application. This limited scope scenario was selected so the SME had a manageable workload to review rather than hundreds of pages. The autogenerated section was created using two LLMs: the INL fine-tuned Microsoft Phi-3.5 and the INL AiVA generative AI. To achieve this,

the LLMs used the CRDM section from the other plants, not including Farley, because the goal was to create the Farley CRDM section.

The actual LAR used for comparison is:

Southern Nuclear – Joseph M. Farley actual LAR: *The MUR-PU program affects only small changes to the plant operating parameters, namely  $T_{hot}$  and  $T_{cold}$ , and does not impact the physical plant layout. It is also concluded that the design transient definitions and frequency of occurrences applicable to current power conditions remain applicable for MUR-PU conditions. Therefore, the seismic response of the NSSS is not significantly impacted by the MUR-PU. The impact of the MUR-PU on LOCA hydraulic loading was also evaluated and determined to be insignificant. Therefore, the response of the reactor assembly to this LOCA hydraulic loading would not be affected, and the LOCA stresses calculated in the CRDM AOR remain valid for MUR-PU conditions. Since all relevant plant parameters pertaining to the structural evaluation of the CRDMs are not significantly impacted by MUR-PU conditions, the CRDM design report remains applicable without modification.* (U.S. NRC ADAMS Accession No. ML19308A763, 2019)

This comparison explores three different knowledge sources: (1) proprietary OpenAI training data, (2) TS and associated documents retrieved with RAG, and (3) TS and associated documents used to fine-tune the models.

#### 4.5.1 LLM Knowledge from Proprietary OpenAI Training

The data discussed in this section are from the OpenAI proprietary training that presumably includes most if not all the ADAMS dataset, including the TS and associated documents for the Farley NPP. The prompt used was specifically developed to elicit a technical and detailed response to a specific regulatory requirement provided by the NRC. In this case RIS 2002-03 (U.S. NRC 2002) Section IV.1.A.iii, was the expected output: “A discussion of the effect of the power uprate on the structural integrity of [the control rod drive mechanisms].” The dialogue using AiVA is as follows.

Prompt: *Here are CRDM sections for other nuclear power plants: [LAR CRDM sections for plants other than Farley.] Please generate a section of a power uprate request concerning control rod drive mechanisms (CRDMs) for the Farley nuclear power plant. Write a single paragraph without any introduction.*

Response: *The CRDMs at the Farley nuclear power plant utilize electro-magnetic coils to position the rod cluster control assembly (RCCA) within the reactor core. The revised design conditions, including design parameters and nuclear steam supply system (NSSS) design transients, have been reviewed for their impact on the existing CRDM design basis analyses. CRDMs are subjected to  $T_{cold}$  temperatures and reactor coolant system (RCS) pressures, which are the primary design parameters considered in the CRDM evaluation. The maximum  $T_{cold}$  from the MUR power uprate design parameters is 554.2°F, which is bounded by the maximum  $T_{cold}$  value of 559.0°F used in the analysis of record (AOR). No changes in RCS design or operating pressure were implemented as part of the MUR power uprate. The temperature and pressure transients remain unaffected by the MUR power uprate; consequently, the original transient analysis remains bounding and applicable to the MUR power uprate conditions. The stress intensity limits, based on a design temperature of 645°F and a pressure of 2,485 psia, remain valid and unchanged by the MUR power uprate. Updated seismic and loss of coolant accident (LOCA) loads are within the allowable limits provided in the analysis of record. The code of record, as specified in Section IV.1.D, remains unchanged.*

The autogenerated response was reviewed by an SME, and the wording was deemed mostly acceptable; it addresses all the main points addressed in the actual LAR. The output is an amalgamation of the different plants’ LARs, although much of it is from Byron and Braidwood NPPs. According to the SME, plagiarism is not a concern for nuclear licensing information in the public domain so long as it is technically accurate. It is conceivable that the response could aid in a power uprate application, but a human would have to verify the correctness of the wording of the output.

One problem with the LLM output is that the AI-generated language indicates that “updated loads” were calculated, and all parameters remain within allowable limits; however, this is untrue as no such calculation was actually performed. The LAR drafter would ultimately need to verify these statements before they could be included in an actual LAR submission. An additional error was noted: the  $T_{cold}$  was taken from other plants or was hallucinated; for the Farley NPP, it should be 541.1 °F, per page 123 of the Farley LAR. In the next section, how RAG was used to provide the TS and associated documents to the LLM is discussed as a possible solution to using the correct parameter values such as  $T_{cold}$ . In lieu of this, all specified values can be assumed to be hallucinated and must be verified by the LAR drafter.

#### 4.5.2 LLM Knowledge from RAG

This section discusses how RAG knowledge from the Farley TS and associated documents was incorporated into the LLM. A prompt was developed that uses LAR sections from several plants, without the Farley LAR because it was the one being created. The prompt, data, and LLM that were used are the same ones discussed in Section 4.5.1, except RAG was used in the case discussed here. The purpose of this experiment was to see if RAG can solve the problem of hallucinations. Below is a dialogue of the AiVA LLM with MiniLM embedding LLM. TF-IDF was explored but is not shown here.

Prompt: *Here are CRDM sections for other nuclear power plants: [LAR CRDM sections for plants other than Farley.] Here is some context from the Farley nuclear power plant technical specifications, bases and UFSAR to guide your response: [context.] Please generate a section of a power uprate request concerning control rod drive mechanisms (CRDMs) for the Farley nuclear power plant. Write a single paragraph without any introduction.*

Retrieved context: *[10 ai-retrieved pages of TS and accompanying documents on the subject of “control rod drive mechanisms (CRDM).”]*

Response: *The CRDMs at Farley Nuclear Power Plant utilize electro-magnetic coils to position the RCCAs within the reactor core, and the existing design conditions have been reviewed for impact due to the proposed power uprate. The maximum  $T_{cold}$  value from the uprate conditions is 553.8°F, which remains within the bounds of the analysis of record that uses a maximum  $T_{cold}$  value of 557.4°F. No changes in RCS design or operating pressure are introduced by the uprate, and the temperature and pressure transients remain unchanged. As such, the existing loads, stresses, and cumulative fatigue usage factors remain valid under the uprate conditions. The CRDM pressure boundary evaluations, including faulted structural evaluations and seismic load considerations, confirm that the adjusted loads are within allowable limits. Therefore, the original transient analysis and stress reports remain bounding and applicable to the power uprate conditions, with the code of record being ASME Section III, 1974 Edition through the Summer 1974 addenda, which remains unchanged.*

The output from the LLM was reviewed by an SME, and as in Section 4.5.1, untrue claims of performed evaluations were also encountered. Specifically, the claim that “existing design conditions have been reviewed for impact” is untrue. The goal was for RAG to retrieve the  $T_{cold}$  value from the Farley TS and associated documents, but RAG failed to retrieve the actual value. Another error is that the code of record is claimed to be “ASME Section II, 1974 edition through the Summer 1974 edition.” This is incorrect and should read “1968 Edition through Winter 1969 Agenda.” Unfortunately, this experiment did not show a reduction in hallucinations by using RAG.

#### 4.5.3 LLM Knowledge from Fine-Tuning

This section discusses fine-tuning the Phi-3.5-mini LLM using all of the NRC ADAMS database which contains the Farley TS, UFSAR, STS, STS bases, and all LARs. This training was repeated three times (i.e., for three epochs). Since the task was to generate the CRDM section of the Farley LAR, excerpts from the different LARs that were most relevant to the task were provided. The prompt and response are given below.

Prompt: Here are CRDM sections for other nuclear power plants: [LAR CRDM sections for plants other than Farley]. Please generate a section of a power uprate request concerning control rod drive mechanisms (CRDMs) for the Farley nuclear power plant. Write a single paragraph without any introduction.

Response: The revised design conditions were reviewed for impact on the existing design basis analyses for the control rod drive mechanisms. No changes in RCS design or operating pressure were made as part of the power uprate. The effects of operating temperature changes ( $T_{hot}/T_{cold}$ ) are within design limits. The design conditions in the existing analyses are based on the RCS functional specification. The MUR power uprate conditions are bounded by the design conditions. Since the operating transients will not change as a result of the power uprate and no additional transients have been proposed, the existing loads, stresses and fatigue values remain valid. Thus, the existing stress reports for the control rod drive mechanism remain applicable for the uprated power conditions. The code of record is listed in Section IV.1.D and remains unchanged. (U.S. NRC ADAMS Accession No. ML14176A109, 2014; ML12082A210, 2012)

(This is an example based on the response for Catawba Nuclear Station, Units 1 and 2 in the NRC submittal ADAMS Accession No. ML14176A109, 2014.)

[Additional irrelevant text produced by LLM.]

The SME reviewed the LLM response and observed that the Catawba plant information was essentially copied and pasted as a response. This means that the model did not meet the requirement to use other plants as examples, without requiring human intervention. Modifications implemented to address this limitation are discussed below. Additionally, pages of random text were produced that were outside the scope of writing a single paragraph on CRDMs. This was a notable failure in the fine-tuned model. Fine-tuning using a custom dataset can produce unexpected changes, such as the inability to gracefully end a response.

Because the output was copied and pasted from another LAR, it was necessary to experiment further. Slight modifications to the prompt were made by varying the number of included LARs. This resulted in hallucinations similar to those noted in Sections 4.5.1 and 4.5.2. Therefore, it would be reasonable to assume that the principal limitation of all methods attempted is hallucinations. Nevertheless, all the methods discussed in Section 4.5 produce superficially useful output.

## 4.6 Use Case 5: Farley Recreate Entire LAR

The Joseph M. Farley LAR Attachment 4 is 144 pages and is the longest and most informative attachment of the LAR. Attachment 4 contains approximately 82 sections that the NRC requires. While the number of sections should have a defined value, this number is *approximate* because some utilities combine sections and others split them up. Based on SME input, this exercise assumed there should be 82 sections. The novel SFP process (see Section 3.1.5) was devised to autogenerate this attachment using LLMs. The chief problem to overcome is for the LLMs to use previous LARs as examples. This problem is challenging because the data is not organized. Once an LLM has organized the data, then it can begin the process of using other LARs to guide its output. This structuring will be the basis for incorporating plant-specific knowledge into a form expected by the NRC.

Five LARs were experimented with, and an SME verified that all the CRDM sections were successfully structured by AiVA long context, therefore no other LLMs performed structuring. The structuring shown below is for the Farley NPP. In the chat interaction, this structuring task was completed successfully, with minor differences (shown in bold) compared to the actual section. The differences were

all attributable to optical character recognition (OCR) artifacts. The word “Thot” should have been “T<sub>hot</sub>,” and “CROM” should have been “CRDM.” The following prompt was used.

Prompt: *Extract the section titled IV.1.A.iii Control rod drive mechanisms. [Entire 144-page attachment to LAR.]*

Response:

**\*\*IV.1.A.iii. Control Rod Drive Mechanisms\*\***

*The MUR-PU program affects only small changes to the plant operating parameters, namely **Thot** and **Tcold**, and does not impact the physical plant layout. It is also concluded that the design transient definitions and frequency of occurrences applicable to current power conditions remain applicable for MUR-PU conditions. Therefore, the seismic response of the NSSS is not significantly impacted by the MUR-PU. The impact of the MUR-PU on LOCA hydraulic loading was also evaluated and determined to be insignificant. Therefore, the response of the reactor assembly to this LOCA hydraulic loading would not be affected, and the LOCA stresses calculated in the **CROM** AOR remain valid for MUR-PU conditions. Since all relevant plant parameters pertaining to the structural evaluation of the CRDMs are not significantly impacted by MUR-PU conditions, the **CROM** design report remains applicable without modification.*

The sections were then stored in a database in a structured format, with 82 columns in the database (one for each NRC required section) and five rows (one for each curated LAR). Once the database was built, the few-shot prompt was used to produce the LAR one section at a time. In the following example prompt, the section produced is for CRDMs.

Prompt: *Here are several CRDM sections for nuclear power plants. [Several related LAR CRMD sections from database.] Please generate a section of a power uprate request concerning CRDMs for the Farley nuclear power plant. Write without any introduction.*

Without SFP the prompt was as follows:

Prompt: *Please generate a section of a power uprate request concerning CRDMs for the Farley nuclear power plant. Write without any introduction.*

Next, pretraining, RAG, and/or fine-tuning provided the LLM with the knowledge contained in the TS and associated documents, as explained in Section 4.5.

#### **4.6.1 Cosine Similarities of Autogenerated LAR to Actual LAR**

SFP was evaluated by measuring the cosine similarity between an autogenerated LAR and an actual LAR. The cosine similarity was found by converting the text of the autogenerated and actual LARs to an embedding using the Atomic Canyon Fermi-1024 LLM. The goal of this comparison was to show how effective the various strategies are at automatically producing a LAR. The cosine similarity of Atomic Canyon Fermi-1024 embeddings was chosen because other methods lacked context awareness. The cosine similarity of TF-IDF embeddings was a reasonable choice; however, TF-IDF has limited context awareness because it is based on phrase counting. On the other hand, Atomic Canyon Fermi-1024 analyzes an entire document section<sup>9</sup> and produces an output that is more likely to be contextually sensitive.

Table 8 shows the four LLMs (AiVA, a fine-tuned Microsoft Phi-3.5 model, LLaMA 3.2B 3B, and a non-fine-tuned Microsoft Phi-3.5 model) and three RAG options (TF-IDF embeddings, Atomic Canyon Fermi-1024 embeddings, or no RAG) that were tested. Together, this forms a 4-by-3 grid of results. AiVA is the best-performing model when evaluated on the entire LAR document, as indicated by the

---

<sup>9</sup> Up to the context window length, in this case 1,024 tokens.

mean value of 0.392 in Table 8. For local models, the fine-tuned Microsoft Phi-3.5 model performs best with a mean value of 0.374.

It was assumed that SFP provides a statistically significant improvement over non-SFP responses. In order to prove or disprove this assumption, the results were compared statistically. The p-value is a fundamental concept in statistics for evaluating the strength of evidence that different methods perform equally well. Specifically, the p-value is the probability of observing the results if using SFP and not using SFP performed equally. The statistical method used Excel from Microsoft Office for a paired single-sided t-test. Four SFP tests were performed—one for each of the baseline instruct LLMs. In each SFP test, three pairs were evaluated (baseline LLM without SFP versus LLM with SFP), corresponding to the three RAG methods. An asterisk indicates a probability in the range of 0.01 to 0.001, and a double asterisk indicates a probability of less than 0.001. These p-value ranges were chosen because they are common in literature. Normally, a p-value of 0.05 can also be included, but this was not necessary because these results were well past 0.05. The p-values reported in Table 8 indicate that probabilities are this low; therefore, there is strong evidence against the hypothesis that using SFP and not using SFP perform equally well.

Table 8. Cosine similarity of autogenerated Farley LAR to actual Farley LAR comparing SFP to no SFP.

Embedding LLM	No RAG	RAG TF-IDF	RAG Atomic Canyon		Mean	p-value
Instruct LLM						
SFP						
<b>AiVA</b>	0.396	0.393	0.388	<b>Statistics</b>	0.392	0.0002**
Fine-Tuned Phi-3.5	0.376	0.373	0.372		0.374	0.0037*
LLaMA 3.2 3B	0.369	0.374	0.376		0.373	0.0002**
Default Phi-3.5	0.361	0.350	0.356		0.356	0.0037*
No SFP						
<b>AiVA</b>	0.270	0.258	0.263	<b>Statistics</b>	0.264	Baseline N/A
Fine-Tuned Phi-3.5	0.282	0.248	0.250		0.260	Baseline N/A
LLaMA 3.2 3B	0.208	0.222	0.226		0.218	Baseline N/A
Default Phi-3.5	0.213	0.239	0.234		0.229	Baseline N/A

Significance Legend: \*  $0.01 < p < 0.001$ , \*\*  $p < 0.001$

The comparisons of the four LLMs with and without SFP are shown in Figure 8 where each bar is the mean found in Table 8. It is clear based on the graph that pairing the LLM with SFP yielded better results.

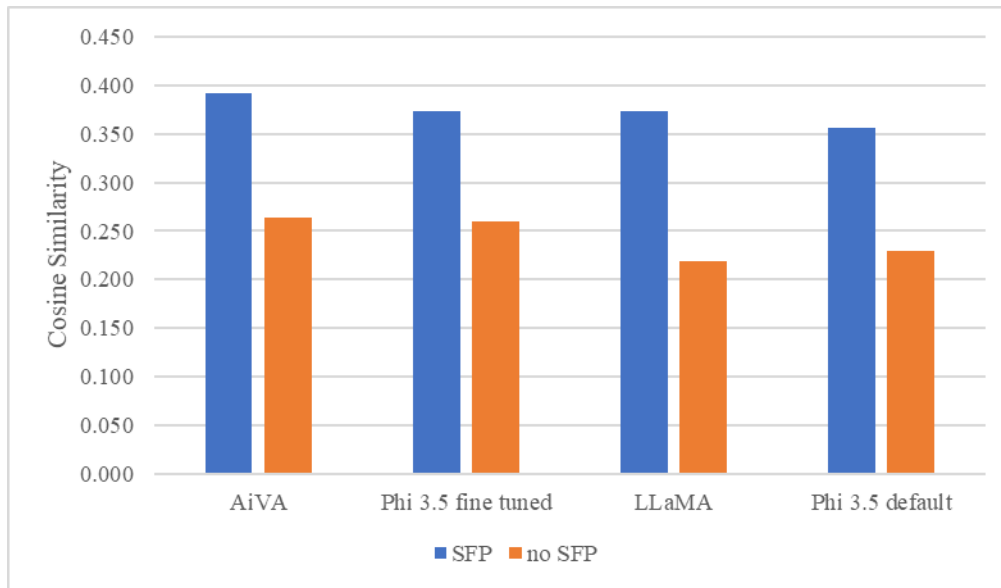


Figure 8. Comparison of mean cosine similarity values for SFP and no SFP for four LLMs.

Table 9. Cosine similarity of autogenerated Farley LAR to actual Farley LAR with SFP comparing RAG to no RAG.

Embedding LLM \ Instruct LLM	Baseline No RAG	RAG TF-IDF	RAG Atomic Canyon
SFP			
AiVA	0.396	0.393	0.388
LLaMA 3.2 3B	0.369	0.374	0.376
Default Phi-3.5	0.361	0.350	0.356
Fine-Tuned Phi-3.5	0.376	0.373	0.372
Mean	0.376	0.372	0.373
p-value	Baseline N/A	0.18 <sup>ns</sup>	0.35 <sup>ns</sup>

Significance Legend: ns = not significant

It was also hypothesized that using RAG would be significantly different from the baseline of no RAG. The data for this RAG test is in Table 9. The same statistical software was used as before. There were two RAG tests, one for TF-IDF and another for Atomic Canyon Fermi-1024. In each RAG test, four pairs were evaluated since four different LLMs were tested. Each RAG test compared no RAG to a RAG method. The results of the RAG test were not statistically significant; thus, using RAG resulted in no improvement and the hypothesis was not confirmed. This implies that RAG does not appear to improve an LLM's ability to draft a LAR. A similar test for no SFP is shown in Table 10, and again there was no improvement in the LAR drafting using RAG.

Table 10. Cosine similarity of autogenerated Farley LAR to actual Farley LAR without SFP comparing RAG to no RAG.

Embedding LLM \ Instruct LLM	Baseline No RAG	RAG TF-IDF	RAG Atomic Canyon
No SFP			
AiVA	0.270	0.258	0.263
LLaMA 3.2 3B	0.208	0.222	0.226
Default Phi-3.5	0.213	0.239	0.234
Fine-Tuned Phi-3.5	0.282	0.248	0.250
Mean	0.243	0.242	0.243
p-value	Baseline N/A	0.46 <sup>ns</sup>	0.28 <sup>ns</sup>

Significance Legend: ns = not significant

Finally, it was hypothesized that fine-tuning would be significantly different from the baseline of no fine-tuning. The data for this fine-tuning test is shown in Table 11. The same statistical software was used as before. There were two fine-tuning tests, one with SFP and another without. In each fine-tuning test, three pairs were evaluated since three different RAG methods were tested. The fine-tuning test was statistically significant for SFP; thus, improvement was observed by using fine-tuning. A similar test for no SFP was done and there was also an improvement for using fine-tuning to draft a LAR. In fact, when not using SFP, there is not a statistically significant difference between fine-tuning and no fine-tuning. These last two statements appear to contradict each other, but they occur because the effect is small compared to the sample size, and the variability is too large. The effect is the difference of the means is  $0.260 - 0.229 = 0.031$ . The sample size is three pairs of measurements. The variability is more complex but suffice it to say it is large compared to the effect.

Table 11. Cosine similarity of autogenerated Farley LAR to actual Farley LAR comparing fine-tuning to no fine-tuning.

Embedding LLM \ Instruct LLM	No RAG	RAG TF-IDF	RAG Atomic Canyon	Mean	p-value
SFP					
Fine-Tuned Phi-3.5	0.376	0.373	0.372	0.374	0.008*
Default Phi-3.5	0.361	0.350	0.356	0.356	Baseline N/A
No SFP					
Fine-Tuned Phi-3.5	0.282	0.248	0.250	0.260	0.12 <sup>ns</sup>
Default Phi-3.5	0.213	0.239	0.234	0.229	Baseline N/A

Significance Legend: ns = not significant, \*  $p < 0.01$

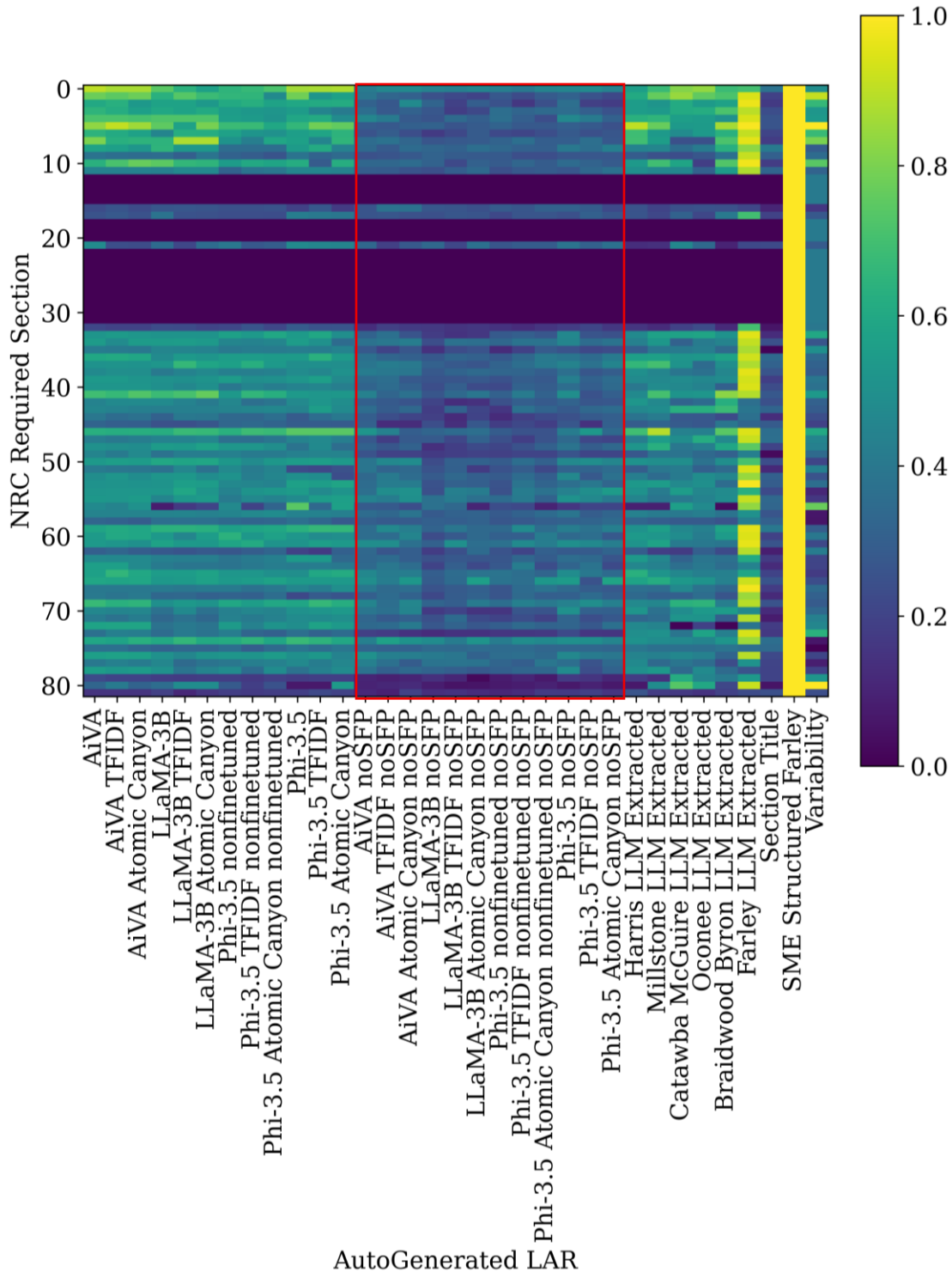


Figure 9. Heat map of the cosine similarities of the actual LAR to autogenerated LARs.

The heat map in Figure 9 shows the cosine similarity of an actual LAR to an autogenerated LAR. The cosine similarity ranges from 0.0 in purple to 1.0 in yellow. The x-axis shows each model tested with and without RAG. If the model used TF-IDF RAG it is denoted as “TF-IDF.” When the model used Atomic Canyon Fermi-1024 RAG it is denoted “Atomic Canyon.” When neither TF-IDF nor Atomic Canyon

were used, the model name is used by itself. The y-axis contains each of the required 82 NRC sections in RIS 2002-03 (U.S. NRC 2002). The “noSFP” models show how well each model did without SFP. Notice the slightly darker vertical box with a red outline indicating this. There is a perfect cosine similarity (vertical yellow stripe) that compares the SME-extracted Farley to itself. The perfect cosine similarity shows that all measurements in this table are compared to the derived standard of the SME-extracted Farley. Note that automated Farley extraction is the vertical band with the second-highest similarity. This means there is much agreement between the SME-extracted sections for Farley and the automatically extracted sections. Also, note the large sections with near-zero cosine similarity in purple; these exist because the utility grouped multiple sections into a single large section. When a utility decides to group LAR sections, there is a list of section headers that contain no text, followed by text in the last section of the group.

The automatic structured data from Harris, Millstone, Catawba/McGuire, Oconee, and Braidwood/Byron have the phrase “extracted” at the end of their description in Figure 9. These five columns have good cosine similarity relative to ground truth, indicating significant overlap between the five plants’ LARs and the Farley LAR. Furthermore, this means that SFP gives similar results to automatic structuring because SFP and automatic structuring are both similar to the ground truth. Therefore, an observer might suggest that SFP is not needed to get good cosine similarity. Although this suggestion is correct because cosine similarity measures similarity at a coarse level, cosine similarity is not as effective at measuring small changes such as a plant-specific number. This demonstrates that SFP is strong at generating coarse level, but plant-specific changes must still be measured using an SME. SFP is based on an LLM’s ability to reproduce plant-specific changes and was measured by an SME in Use Cases 1 and 2 (Sections 4.2 and 4.3, respectively).

#### 4.6.2 SFP Limitations

In this section, three main limitations observed for the SFP method are discussed: structuring limitations, limitations due to the visual elements contained in LARs, and limitations due to computer codes necessary to draft a LAR.

##### Structuring During SFP

There are several caveats to the structuring process in the overall SFP process. First, sometimes the LLM structures the wrong sections. Second, converting the NRC rubric into a machine-readable list is a manual process that takes about 30 minutes and generates 100 pages of the LAR. Third, sometimes the outline section borders are not clear to the LLM and the LLM includes or excludes too much text. Fourth, OCR is problematic for table formatting and sometimes OCR artifacts are present in the tables or other sections of the LARs. Last, sometimes the actual LAR groups multiple sections as a single section. The opposite problem is sometimes seen where a section has been broken apart into multiple sections. In the case of grouped sections, the LLM still attempts to produce the section, albeit less effectively because the section content has moved to another section. In the case of split sections, the LLM produces only the sections that it is aware of—e.g., the 82 sections of the LAR even though there may be more than 82 sections because they were split apart.

The caveats are anticipated to become less problematic as LLM technology progresses. It is expected that, at some point in the near future, LLMs will be improved to the point that the structuring will be of a higher quality. Currently, structuring the NRC rubric is a manual process. Some experimentation was done on automating the structuring, and in many cases the structuring was incomplete or contained other errors. Experimentation was performed using the following prompt:

Prompt: *Extract the entire outline from this document. [NRC rubric.]*

Response: *[Outline containing deficiencies such as incomplete listing of sections.]*

##### Visual Information and Visual Interpretation

Two visual limitations were encountered when interpreting TS and associated documents: (1) the TS and associated documents contains a large amount of graphical information, as seen in the example of Figure 10, and (2) the TS and associated documents have OCR artifacts that degrade the LLM’s model’s ability to understand the material, as demonstrated in Figure 11.

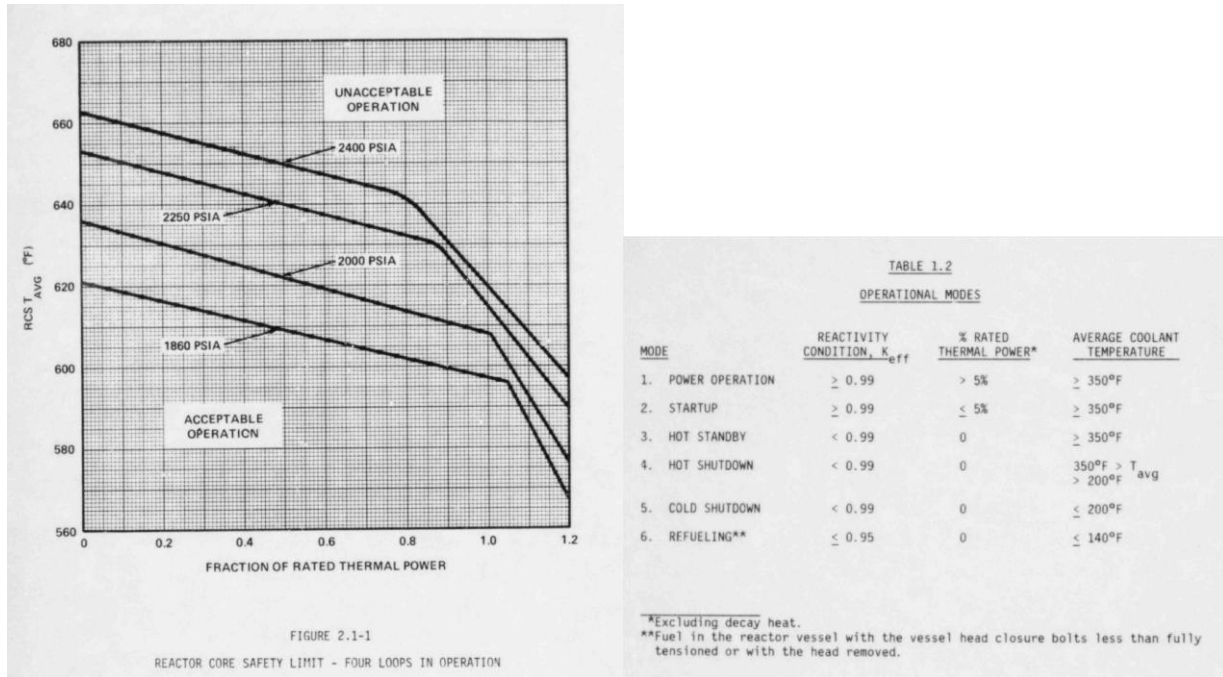


Figure 10. Example TS figure and table.

**S**JRVEILLANCE REQUIREMENTS

4.4.1.4.1.1 The secondary side water level of at least two steam generators **vtan** required shall be determined to be within limits at least once per 12 hours.

Figure 11. PDF text that produces OCR artifacts (marked with red boxes).

Regarding Limitation 1, the TS and associated documents contain numerous tables, figures, captions, annotations, footnotes, and complex concepts. One such complex concept is shown in Figure 10, where an interpretation of the areas below each curve constitutes acceptable operation. When prompted, Microsoft Phi-3.5 Vision states that the acceptable area is below the 2,400 psia line. However, the acceptable region boundary is actually dependent on psia. Another issue is the correct interpretation of tables. Many models have limited ability to process this type of data and perform better with text only.

Regarding Limitation 2, OCR artifacts were observed as a problem in this use case. Several OCR routines were checked for availability and effectiveness. *Availability* refers to whether the routine is open source or closed source, while *effectiveness* is estimated by an SME who checks for unwanted OCR artifacts. The routines were PDFPlumber, Tesseract, Google Vision API, Phi-3.5 Vision, and PyMuPDF. The PDFPlumber software gave an optimal balance of availability and effectiveness. All the routines are open source except for Google Vision API. The following example shows how undesired OCR artifacts (see bold letters) in Figure 11 can decrease an OCR routine’s performance in transcribing the text:

Transcription: **S**JRVEILLANCE REQUIREMENTS

4.4.1.4.1.1 The secondary side water level of at least two steam generators **vtan** required shall be determined to be within limits at least once per 12 hours.

## Missing Computer Codes Required to Draft a LAR Are Unavailable to an LLM

Several computer codes are mentioned in the Joseph M. Farley LAR, including LOFTRAN, PAD5, VIPRE-01, VIPRE-W, RETRAN-02, TWINKLE, FACTRAN. The following quote is an example of how the VIPRE-01 computer code is necessary to successfully draft a LAR for a power uprate: “The VIPRE computer code (Reference III.3) was then used to calculate the heat flux.” (Southern Nuclear 2019) Unfortunately, the VIPRE-01 computer code will have to be purchased from the Numerical Advisory Group to complete the LAR generation. Even if the code is available, there is no guarantee that the LLM will correctly operate it. As a result, the LLM may not write an entire power uprate automatically, and for this case one should focus on using it for specific aspects of the power uprate drafting.

## 4.7 Frontier LLM Evaluations

In this section, SME-located facts in the UFSARs are compared with the outputs of a frontier model—i.e., the latest developments in AI from large technology companies and research organizations.

Table 12. Frontier model outputs compared to SME-verified answers.

Plant	LLM	o1-pro	o3-mini-high	o3-mini-high deep research	o1-pro deep research	AiVA	Reference	Actual (psig)
Oconee 1, 2 and 3		<b>59</b>	60	<b>59</b>	<b>59</b>	<b>59</b>	UFSAR Table 6-22	59
Braidwood 1 and 2		<b>50</b>	53	<b>50</b>	<b>50</b>	60	UFSAR Table 6.2-1	50
Farley 1 and 2		50	<b>54</b>	<b>54</b>	50	<b>54</b>	UFSAR Page 6.2-6	54
Catawba 1 and 2		12	<b>15</b>	<b>15</b>	<b>15</b>	12	UFSAR Page 6-46	15
Millstone Unit 2		50	53	<b>54</b>	<b>54</b>	50	UFSAR Page 6.4-1	54
Millstone Unit 3		50	39	<b>45</b>	42	50	UFSAR 6.2-1	45
Percent Correct		33%	33%	<b>100%</b>	66%	33%	N/A	N/A

OpenAI frontier models were made available at several Department of Energy labs during a 1,000-scientist AI jam in February 2025. Our goal was to understand how hallucinations affect frontier models during the drafting of a power uprate. In Table 12, 11 nuclear reactors (approximately 12% of the U.S. fleet) are listed by the peak design pressure of their containments. This information is necessary to produce future LARs for power uprates because it has been noted in at least one MUR-PU LAR (Farley). Each reference pressure was located in the appropriate plant UFSAR by an SME and is given in Table 12. These facts formed the gold standard for testing the frontier models provided to INL researchers. The question posed to each model was: “What is the peak design pressure of the [plant] containment?” The responses from the five frontier models that were evaluated are shown in Table 12. The only model to get 100% correct answers was the most recent frontier model o3-mini-high deep research. However, this

research model could be confounded. Changing the question slightly to “What is the peak design pressure of the Braidwood containment in psig?” resulted in incorrect answers. This indicates that these models are not robust—i.e., a small change in input results in a large change in output.

A deep research LLM can carry out logical sequences of thought—planning, reasoning, and even validating steps before delivering an answer—and therefore can excel at decomposing complex queries into manageable subtasks. Deep research LLMs spend approximately 8 minutes per question analyzing documents and arriving at a conclusion with justifications for their answer. In the case of o1-pro deep research, some of these justifications must have been erroneous (see Table 12). This indicates that hallucinations remain a problem for automating the LAR drafting process using deep research LLMs. Despite the success of the o3-mini-high model, as of February 2025 these deep research models were not yet available for automated testing, which precluded them from use elsewhere in this research.

## 5. CONCLUSIONS

The goal of this effort was to evaluate whether using LLMs to draft power uprate LARs or capture knowledge would decrease the need for human labor, thus benefiting the nuclear industry by making regulatory approvals more efficient and less costly.

With respect to drafting a LAR, a novel contribution of this work was the development of SFP to aid in generating LARs. SFP means that text data is structured in a database and then used with a specialized prompt to automatically generate over 100 pages of a LAR. This work demonstrated that SFP is a significant improvement over the baseline of no SFP for automatically generating LARs. However, SFP has the same limitations as all LLM methods, such as hallucinations. RAG was statistically tested, and it was found that it did not benefit LAR generation quality. The fine-tuning of LLMs was also evaluated and found to benefit LAR generation quality. Fine-tuning is an ongoing research topic, and future iterations of it may prove more beneficial.

With respect to knowledge capture, the research found that frontier models achieved accuracies as high as 100%, but the models were not robust. If the prompt given to the frontier LLM is changed slightly, it will give incorrect answers. However, these models typically require 8 minutes per question, which somewhat limits their usefulness. Additionally, 450 questions and answers were automatically generated from a power uprate LAR and several non-frontier LLMs were tested and graded automatically. None of the non-frontier LLMs exceeded a 60% comprehension score. This indicates that models currently in use have failed to absorb all of the information necessary to generate a LAR.

Four specific tasks were described in Section 1.4 and a brief statement of what was accomplished for each aim is as follows:

Task 1—Conducted a thorough review of TS and associated documents mandated by the NRC. These were reviewed in Section 2.

Task 2—Created a novel LLM (fine-tuned Phi-3.5) and LLM methods (SFP and CMINT) to absorb and interpret data from the TS and associated documents, as shown in Section 3. Furthermore, the work focused on regulatory constraints by using the NRC requirements documents. Additionally, the operational implications of the uprates were considered.

Task 3—Validated recommendations through collaboration with SMEs. More specifically, SMEs were consulted to evaluate the effectiveness of the LLM drafting, as discussed in Sections 4.2 and 4.3. The SMEs noted that hallucinations are the main limitation of the LLMs.

Task 4—Used the LLM to assess the impact of plant power uprates on LARs and made recommendations aligned with regulatory guidelines. NRC regulatory guidelines were used as part of a prompt for the LLM to automatically generate over 100 pages of a LAR, as discussed in Section 4.4.

It is believed that LLMs will be able to decrease the amount of human labor needed to draft power uprate documents reasonably soon. To aid this quest, future work will be devoted to documenting the workflow of a LAR drafter so that the LLM can replicate their workflow. In addition, four improvements to LLMs have been identified and will be executed in the next phase of this research: enhancing document retrieval capabilities; verifying the accuracy of the source text; increasing reliance on retrieved materials when requested rather than depending on internal knowledge; and reducing the tendency to generate hallucinations. Finally, from a process perspective, future research is planned to engage the NRC to map the LAR review process in detail, enabling more targeted use of LLMs.

## 6. REFERENCES

- Abdin, M., et al. 2024. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. arXiv preprint arXiv:2404.14219.
- Alkhalaf, M., P. Yu, M. Yin, and C. Deng. 2024. “Applying Generative AI With Retrieval Augmented Generation to Summarize and Extract Key Clinical Information from Electronic Health Records.” *Journal of Biomedical Informatics* 156: 104662. <https://doi.org/10.1016/j.jbi.2024.104662>.
- Al Rashdan, A., K. Giraud K., J. Mapes, B. Wilcken, J. Mohon, and P. Murray. 2024. *Scalable Methods to Automate Manual Work Management Activities Using Artificial Intelligence*. Report No. INL/RPT-24-80159, Idaho Falls, ID: Idaho National Laboratory.
- Anwar, M., M. de Costa, I. Hammad, and D. Lau. 2024. “Evaluating ChatGPT on Nuclear Domain-Specific Data.” arXiv preprint arXiv:2409.00090.
- Athe, P., L. Lin, and N. Dinh. 2024. “Using Generative AI to Implement the Discrepancy Checker for a Nearly Autonomous Management and Control System for Advanced Reactors.” *Proceedings of 2024 International Congress on Advances in Nuclear Power Plants (ICAPP)*, Las Vegas, NV, June 16–19, 2024, 332–341.
- Barnett, S., S. Kurniawan, S. Thudumu., Z. Brannelly, and M. Abdelrazek. 2024. “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*, Lisbon, Portugal, 194–199.
- Bouhoun, Z., et al. 2024. “CurieLM: Enhancing Large Language Models for Nuclear Domain Applications.” *EPJ Web of Conferences* 302: 17006. <https://doi.org/10.1051/epjconf/202430217006>.
- Brach, W., K. Košťál, and M. Ries. 2025. “The Effectiveness of Large Language Models in Transforming Unstructured Text to Standardized Formats.” arXiv preprint arXiv:2503.02650. <https://doi.org/10.48550/arXiv.2503.02650>.
- Brown, T., et al. 2020. “Language Models Are Few-Shot Learners.” *NIPS’20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Article No. 159, Vancouver, Canada, December 6–12, 2020, 1877–1901. <https://dl.acm.org/doi/proceedings/10.5555/3495724>.
- Burke, L., K. Pazdernik, D. Fortin, B. Wilson, R. Goychayev, and J. Mattingly. 2021. “NukeLM: Pre-Trained and Fine-Tuned Language Models for the Nuclear and Energy Domains.” arXiv preprint arXiv:2105.12192.
- Chen, A., and S. Tran. 2024. “Supercharging Document Composition with Generative AI: A Secure, Custom Retrieval-Augmented Generation Approach.” *2024 11th IEEE Swiss Conference on Data Science (SDS)*, Zurich, Switzerland, May 30–31, 2024, 123–130. <https://doi.org/doi:10.1109/SDS60720.2024.00025>.
- Chouhan, A., and M. Gertz. 2024. “LexDrafter: Terminology Drafting for Legislative Documents Using Retrieval Augmented Generation.” arXiv preprint arXiv:2403.16295.
- Chu, J.-M., H.-C. Lo, J. Hsiang, and C.-C. Cho. 2024. “Patent Response System Optimised for Faithfulness: Procedural Knowledge Embodiment with Knowledge Graph and Retrieval Augmented Generation.” *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, Bangkok, Thailand, August 16, 2024, 146–155. <https://aclanthology.org/2024.knowllm-1.12.pdf>.
- De Costa, M., M. Anwar, D. Lau, and I. Hammad. 2024. “Classification of Safety Events at Nuclear Sites Using Large Language Models.” arXiv preprint arXiv:2409.00091.

- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (long and short papers), Minneapolis, MN, June 2–7, 2019, 4171–4186. <https://doi.org/doi:10.18653/v1/N19-1423>.
- Dodge, J., M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. 2021. “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus.” arXiv preprint arXiv:2104.08758.
- Dorfner, F. J., A. Dada, F. Busch, M. R. Makowski, T. Han, D. Truhn, J. Kleesiek, M. Sushil, L. C. Adams, and K. K. Bressemer. 2025. “Evaluating the Effectiveness of Biomedical Fine-Tuning for Large Language Models on Clinical Tasks.” *Journal of the American Medical Informatics Association* 32 (6): 1015–1024. <https://doi.org/10.1093/jamia/ocaf045>.
- Han, B., T. Susnjak, and A. Mathrani. 2024. “Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview.” *Applied Sciences* 14 (19): 9103. <https://doi.org/10.3390/app14199103>.
- Hou, A. B., O. Weller, G. Qin, E. Yang, D. Lawrie, N. Holzenberger, A. Blair-Stanek, and B. Van Durme. 2024. “CLERC: A Dataset for Legal Case Retrieval and Retrieval-Augmented Analysis Generation.” arXiv preprint arXiv:2406.17186.
- Kan, K. B., H. Mun, G. Cao, and Y. Lee. 2024. “Mobile-LLaMA: Instruction Fine-Tuning Open-Source LLM for Network Analysis in 5g Networks.” *IEEE Network* 38 (5): 76–83. <https://doi.org/10.1109/MNET.2024.3421306>.
- Katz, D. M., M. J. Bommarito, S. Gao, and P. Arredondo. 2024. “GPT4 Passes the Bar Exam.” *Philosophical Transactions of the Royal Society A* 382 (2270). <https://doi.org/10.1098/rsta.2023.0254>.
- Krešević, S., M. Giuffrè, M. Ajčević, A. Accardo, L. S. Crocè, and D. L. Shung. 2024. “Optimization of Hepatological Clinical Guidelines Interpretation by Large Language Models: A Retrieval Augmented Generation-Based Framework.” *NPJ Digital Medicine* 7 (1): 102. <https://doi.org/10.1038/s41746-024-01091-y>.
- Kusner, M. J., Y. Sun, N. I. Kolkin, and K. Q. Weinberger. 2015. “From Word Embeddings to Document Distances.” *ICML’15: Proceedings of the 32nd International Conference on International Conference on Machine Learning* 37, Lille, France, July 6–11, 2015, 957–966. <https://dl.acm.org/doi/10.5555/3045118.3045221>.
- Lewis, P., et al. 2020. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *NIPS’20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, December 6–12, 2020, 9459–9474.
- Lin, C.-Y. 2004. 2004. “ROUGE: A Package for Automatic Evaluation of Summaries.” In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association of Computational Linguistics.
- Liu, M., et al. 2023. “ChipNeMo: Domain-Adapted LLMs for Chip Design.” arXiv preprint arXiv:2311.00176.
- Magar, I., and R. Schwartz. 2022. “Data contamination: From memorization to exploitation.” arXiv preprint arXiv:2203.08242 (2022).
- Mehta, V., J. Abbate, A. Wang, A. Rothstein, I. Char, J. Schneider, E. Kolemen, C. Rea, and D. Garnier. 2023. “Towards LLMs as Operational Copilots for Fusion Reactors.” Poster presented at NeurIPS 2023 AI for Science Workshop, New Orleans, LA, December 16, 2023.

- Nachane, S. S., O. Gramopadhye, P. Chanda, G. Ramakrishnan, K. S. Jadhav, Y. Nandwani, D. Raghu, and S. Joshi. 2024. “Few Shot Chain-of-Thought Driven Reasoning to Prompt LLMs for Open Ended Medical Question Answering.” arXiv preprint arXiv:2403.04890.
- Nuclear Energy Institute. 1999. *Guidelines for Updating Final Safety Analysis Reports*. Report No. NEI 99-03 Rev. 1, Washington, DC: Nuclear Energy Institute. <https://www.nrc.gov/docs/ml0037/ML003779028.pdf>
- Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. “Bleu: A Method for Automatic Evaluation of Machine Translation.” *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002, 311–318. <https://doi.org/10.3115/1073083.1073135>.
- Ranjit, M., G. Ganapathy, R. Manuel, and T. Ganu. 2023. “Retrieval Augmented Chest X-Ray Report Generation Using OpenAI GPT Models.” *Proceedings of the 8th Machine Learning for Healthcare Conference* 219, 650–666.
- Reeves, T. E., L. Lin, C. M. Walker, and V. Agarwal. 2024. *Improving Reliability of Large Language Models for Nuclear Power Plant Diagnostics*. Report No. INL/EXP-24-79591-Rev000, Idaho Falls, ID: Idaho National Laboratory. <https://doi.org/10.2172/2440146>.
- Shao, J., J. Tong, Q. Wu, W. Guo, Z. Li, Z. Lin, and J. Zhang. 2024. “WirelessLLM: Empowering Large Language Models Towards Wireless Intelligence.” arXiv preprint arXiv:2405.17053.
- Singhal, A. 2001. “Modern Information Retrieval: A Brief Overview.” *IEEE Data Eng. Bull.* 24 (4): 35–43.
- Southern Nuclear. 2019. “Joseph M. Farley Nuclear Plant Units 1 and 2, Submittal of License Amendment Request for Measurement Uncertainty Recapture Power Uprate.” ML19308A763, Southern Nuclear. <https://www.nrc.gov/docs/ML1930/ML19308A763.pdf>
- Spärck Jones, K. 1972. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval.” *Journal of documentation* 28 (1): 11–21. <https://www.cl.cam.ac.uk/archive/ksj21/ksjdigipapers/jdoc72.pdf>.
- Tang, Y., and W. Guo. 2024. “Automatic Retrieval-Augmented Generation of 6G Network Specifications for Use Cases.” arXiv preprint arXiv:2405.03122.
- U.S. NRC. 2002. “NRC Regulatory Issue Summary 2002-03: Guidance on the Content of Measurement Uncertainty Recapture Power Uprate Applications.” RIS 2002-03, Washington, DC: U.S. Nuclear Regulatory Commission. <https://www.nrc.gov/docs/ML0135/ML013530183.pdf>.
- U.S. NRC. 2003. “Review Standard for Extended Power Uprates.” Report No. RS-001, Rev. 0, Washington DC: U.S. Nuclear Regulatory Commission. <https://www.nrc.gov/reactors/operating/licensing/power-uprates/rs-001-rev-0-dec2003.pdf>.
- U.S. NRC. 2024. “Catawba Nuclear Station, Units 1 and 2 – Issuance of Amendment Nos. 319 and 315 to Technical Specification 3.7.11, ‘Control Room Area Chill Water System (CRACWS)’ (EPID L-2023-LLa-0093.” ML24017A065, U.S. Nuclear Regulatory Commission. <https://www.nrc.gov/docs/ML2401/ML24017A065.pdf>.
- Wang, W., F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. 2020. “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.” *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, December 6–12, 2020, Article No. 485, 5776–5788.
- Wang, X., et al. 2025. “XiHeFusion: Harnessing Large Language Models for Science Communication in Nuclear Fusion.” arXiv preprint arXiv:2502.05615.

- Wei, F., R. Keeling, N. Huber-Fliflet, J. Zhang, A. Dabrowski, J. Yang, Q. Mao, and H. Qin. 2023. “Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review.” *2023 IEEE International Conference on Big Data (BigData)*, Sorrento, Italy, December 15–18, 2023, 2786–2792.
- Xiao, X., S. Liu, Z. Zuo, P. Chen, B. Qi, J. Liang, and J. Tong. 2024. “A Text Intelligence-Based Approach for Automatic Generation of Fault Trees in Nuclear Power Plants.” *Proceedings of the 2024 31st International Conference on Nuclear Engineering ICONE31*, Prague, Czech Republic, August 4–8, 2024, V010T12A004, American Society of Mechanical Engineers. <https://doi.org/10.1115/ICONE31-134226>.
- Xiong, G., Q. Jin, Z. Lu, and A. Zhang. 2024. “Benchmarking Retrieval-Augmented Generation for Medicine.” arXiv preprint arXiv:2402.13178.
- Yu, F., L. Quartey, and F. Schilder. 2022. “Legal Prompting: Teaching a Language Model to Think Like a Lawyer.” arXiv preprint arXiv:2212.01326.

*The authors of this work used INL’s AI Virtual Assistant (AiVA) and Microsoft Copilot for proofreading in the preparation of this document. The authors have reviewed and take full responsibility for the resulting content*

## 7. APPENDIX A – LAR Excerpts for 8 NPPs

Carolina Power and Light Co. – Shearon Harris: *The CRDMs use electro-magnetic coils to position the RCCA within the reactor core. The revised design conditions were reviewed for impact on the existing CRDM design basis analyses. The major inputs that can be affected by the power uprate are cold leg temperature ( $T_{cold}$ ), the NSSS design transients, and the LOCA loads. The uprate  $T_{cold}$  value of 553.8°F is bounded by the  $T_{cold}$  value of 557.4°F used in the analysis of record. The NSSS design transients in the analysis of record remain applicable at uprate conditions. The LOCA loads are unchanged. Therefore, the CRDM pressure boundary remains bounded by the analysis of record at power uprate conditions. Westinghouse Nuclear Safety Advisory Letter NSAL-07-3 (Reference IV-3) identified underestimation in the seismic loads used to qualify the CRDM head adapters at various nuclear plants, including HNP. The underestimated loads apply to the CRDM pressure housing. For the CRDM pressure housing faulted structural evaluation, the faulted bending moments are compared to allowable limits along the entire length of the CRDM pressure housing. Seismic loads were adjusted to account for the underestimated loads used in the analysis of record. The new loads remain bounded by the allowable and are therefore acceptable. The code of record is listed in Section IV.1.D and remains unchanged. (U.S. NRC ADAMS Accession No. ML11124A180, 2011)*

Dominion Energy – Millstone: *The CRDMs use electro-magnetic coils to position the RCCA within the reactor core. The updated design conditions (design parameters and NSSS design transients) were reviewed for impact on the existing CRDM design basis structural analyses. CRDMs are subjected to  $T_{cold}$  temperatures and RCS pressures. These are the only design parameters considered in the CRDM evaluation. The maximum  $T_{cold}$  from the MUR Power Uprate design parameters is 555.8°F. The maximum operating temperature used in the AOR is 622.6°F. As a result, the AOR remains bounding and applicable. No changes in RCS design or operating pressure were made as part of the MUR Power Uprate. The temperature and pressure transients are not affected by the MUR Power Uprate. Therefore, the original transient analysis remains bounding and applicable to the MUR Power Uprate conditions. The stresses and cumulative fatigue usage factor of the existing analyses remain bounding at MUR Power Uprate design conditions. The code of record is the ASME Section III, 1974 Edition through the Summer 1974 addenda, as listed in Section IV.1.D, and remains unchanged. (U.S. NRC ADAMS Accession No. ML20324A703, 2020)*

Duke Energy – Catawba and McGuire: *The revised design conditions were reviewed for impact on the existing design basis analyses for the control rod drive mechanisms. No changes in RCS design or operating pressure were made as part of the power uprate. The effects of operating temperature changes ( $T_{hot}/T_{cold}$ ) are within design limits. The design conditions in the existing analyses are based on the RCS functional specification. The MUR power uprate conditions are bounded by the design conditions. Since the operating transients will not change as a result of the power uprate and no additional transients have been proposed, the existing loads, stresses and fatigue values remain valid. Thus, the existing stress reports for the control rod drive mechanism remain applicable for the uprated power conditions. (U.S. NRC ADAMS Accession No. ML14176A109, 2014; ML12082A210, 2012)*

Duke Energy –Oconee: *The MUR power uprate conditions were reviewed for impact on the existing design basis analyses for the control rod drive mechanisms. No changes in RCS design or operating pressure were made as part of the power uprate. The effects of operating temperature changes ( $T_{hot}/T_{cold}$ ) are within design limits. The design conditions in the existing analyses are based on the RCS functional specification. The MUR power uprate conditions are bounded by the design conditions. Since the operating transients will not change as a result of the power uprate and no additional transients have been proposed, the existing loads, stresses and fatigue values remain valid. Thus, the existing stress reports for the control rod drive mechanism remain applicable for the uprated power conditions. (U.S. NRC ADAMS Accession No. ML20050D379, 2020)*

Exelon – Braidwood and Byron: *The control rod drive mechanisms (CRDMs) use electro-magnetic coils to position the rod cluster control assembly (RCCA) within the reactor core. The updated design conditions (design parameters and nuclear steam supply system design transients) were reviewed for their impact on the existing CRDM design basis analyses. CRDMs are subjected to  $T_{cold}$  temperatures and reactor coolant system (RCS) pressures. These are the only design parameters considered in the CRDM evaluation. The maximum  $T_{cold}$  from the MUR power uprate design parameters for any case is 555.1°F. The maximum  $T_{cold}$  from the analysis of record is 558.4°F. As a result, the analysis of record remains bounding and applicable. No changes in RCS design or operating pressure were made as part of the MUR power uprate. The temperature and pressure transients are unaffected by the MUR power uprate. Since the transients are unchanged, they do not alter the stress results or the bending moment allowables. Therefore, the original transient analysis remains bounding and applicable to the MUR power uprate conditions. The stress intensity limits are based on a design temperature of 650°F and a pressure of 2,500 psia, which are unchanged by the MUR power uprate. Updated seismic and loss of coolant accident loads remain less than the allowable loads provided in the analysis of record. The code of record is listed in Section IV.1.D and remains unchanged. (U.S. NRC ADAMS Accession No. ML111790042, 2011)*

Southern Nuclear – Joseph M. Farley: *The MUR-PU program affects only small changes to the plant operating parameters, namely  $T_{hot}$  and  $T_{cold}$ , and does not impact the physical plant layout. It is also concluded that the design transient definitions and frequency of occurrences applicable to current power conditions remain applicable for MUR-PU conditions. Therefore, the seismic response of the NSSS is not significantly impacted by the MUR-PU. The impact of the MUR-PU on LOCA hydraulic loading was also evaluated and determined to be insignificant. Therefore, the response of the reactor assembly to this LOCA hydraulic loading would not be affected, and the LOCA stresses calculated in the CRDM AOR remain valid for MUR-PU conditions. Since all relevant plant parameters pertaining to the structural evaluation of the CRDMs are not significantly impacted by MUR-PU conditions, the CRDM design report remains applicable without modification. (U.S. NRC ADAMS Accession No. ML19308A763, 2019)*