INS International Nuclear Security
Reducing Risk of Nuclear Terrorism









Guide on Using AI to Accelerate Security Assessments

March 2026

Office of International Nuclear Security

National Nuclear Security Administration

PNNL-39341

Intentionally blank

PNNL-39341

CASE STUDY ON USING AI TO ACCELERATE SECURITY ASSESSMENTS

Pacific Northwest National Laboratory

Date: March 2026

Prepared by

Samuel Clements, Micheal Girard, Kel Krehbiel, Guy Landine, Caleb Thomas

Managed by

Pacific Northwest National Laboratory

for the

U.S. Department of Energy

under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference 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 United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

EXECUTIVE SUMMARY

Case Study on Using AI to Accelerate Security Assessments

This guide is for experienced security assessors who want to use AI—especially large language models (LLMs)—to make security assessments faster, clearer, and more consistent while maintaining strict nuclear security standards.

Used correctly, LLMs can help structure complex assessments, generate assessment checklists and question sets, organize notes, and draft non-sensitive report sections and summaries. They can improve documentation quality and support communication with both technical and non-technical stakeholders.

At the same time, LLMs introduce critical risks: possible exposure of sensitive information, overreliance on an imperfect tool, and potential dual-use concerns. This guide therefore emphasizes three core principles:

1. **Use LLMs as assistants, not decision-makers.**
2. **Protect sensitive information.**
3. **Verify all outputs.**

The guide provides discussion on where LLMs are valuable and where they are not when conducting security assessments. It provides example prompts, takes readers through a sample assessment on a realistic sensitive system, and concludes with next steps that could be taken to further the value of using AI in security assessments.

Intentionally blank

CONTENTS

Executive Summary ii

Abbreviations, Acronyms, and Initialisms ix

1. Background 1
2. Use Case: Model Access Control System 1
    - 2.1. The Model Access Control System Testbed 1
    - 2.2. Large Language Model Evaluation Methodology 2
    - 2.3. Evaluation Across Data Modalities 4
    - 2.4. Large Language Model Performance Comparison Across Data Modalities 6
    - 2.5. Detailed Case Studies: Prompts, Results, and Conclusions 7
3. Guide on Using AI LLMs Within Security Assessments 11
    - 3.1. Why This Report Matters: Accelerating Nuclear Security Assessments with AI 12
    - 3.2. Overview: AI 14
    - 3.3. Security & Privacy Risks When Using Large Language Models 16
4. Where Large Language Models Can Help in Security Assessments 21
    - 4.1. Assessment Planning and Preparation 21
    - 4.2. Document Preparation and Review 22
    - 4.3. Risk Identification & Scenario Brainstorming 24
    - 4.4. Support During the Assessment 24
    - 4.5. Reporting & Follow-up 25
5. Obtaining High-Quality Responses from Large Language Models 26
    - 5.1. Core Principles of Good Prompting 27
    - 5.2. Examples: Weak Versus Strong Prompts 28
6. Verifying and Using Large Language Model Outputs Safely 31
    - 6.1. Always Apply Expert Review 31
    - 6.2. Beware of “Hallucinations” 31
7. Protecting Sensitive Data When Using Large Language Models 32
    - 7.1. Configure Tools Correctly 32
    - 7.2. Establish Guidelines for Staff 33
8. Safety and Abuse Concerns Specific to Nuclear & Critical Infrastructure 33
    - 8.1. Avoid Generating Harmful Content 33
    - 8.2. Be Alert to Dual-Use Risks 34
9. Common Mistakes Users Make (and How to Avoid Them) 35
    - 9.1. Mistake 1: Pasting Sensitive Documents Directly into a Public Large Language Model 35
    - 9.2. Mistake 2: Treating AI Output as Authoritative Fact 36
    - 9.3. Mistake 3: Asking the Wrong Question (Insufficient Prompt Engineering) 36
    - 9.4. Mistake 4: Overloading the Tool with Too Many Tasks at Once 37
    - 9.5. Mistake 5: Not Preserving an Audit Trail 37

10. Practical Checklist for Setting Up Large Language Model Use in Critical Infrastructure Security Assessments 38
11. Improvements & Additional Research 41
    11.1. Enhancing Large Language Model Capabilities for Critical Infrastructure 41
    11.2. Advancing Secure and Autonomous Large Language Model Applications 42
    11.3. Developing Robust Evaluation and Assurance Methodologies 42
12. Conclusions 43

Intentionally blank

LIST OF TABLES

Table A. LLM Performance Comparison Across Data Modalities 6
Table 1. List of Potential Use Cases for LLMs in Security Assessment Contexts 15
Table 2. Limitations of LLMs in Security Assessment Contexts 16
Table 3. Categories of Sensitive Information in Nuclear Security Assessments 17
Table 4. LLM Deployment Models and Associated Risk Profiles 19
Table 5. Appropriate Use Versus Limitations of LLMS in Nuclear Security Assessments 20
Table 6. Key Applications of LLMs in Assessment Planning and Preparation 22
Table 7. Key Applications of LLMs in Document Preparation and Review 22
Table 8. Key Applications of LLMs in Risk Identification and Scenario Brainstorming 24
Table 9. Key Applications of LLMs in Support During the Assessment 25
Table 10. Key Applications of LLMs in Reporting & Follow-up 25
Table 11. CREATE Mnemonic 26
Table 12. Core Principles for Effective Prompt Construction 27
Table 13. Technical Controls for Secure LLM Deployment 32
Table 14. Key Information to Convey to Staff 33
Table 15. Prohibited and Recommended LLM Interactions for Harmful Content 34
Table 16. Managing Dual-Use Risks in LLM Interactions 34
Table 17. Mistake #1: Pasting Sensitive Documents Directly into Public LLMs 35
Table 18. Mistake #2: Treating AI Output as Authoritative Fact 36
Table 19. Mistake #3: Asking the Wrong Question 36
Table 20. Mistake #4: Overloading AI with Too Many Tasks 37
Table 21. Mistake #5: Not Preserving an Audit Trail 37
Table 22. LLM Implementation Checklist for Critical Infrastructure Security Assessments 38

NO TABLE OF FIGURES ENTRIES FOUND.

Intentionally blank

ABBREVIATIONS, ACRONYMS, AND INITIALISMS

ACS	Access Control System
IAEA	International Atomic Energy Agency
INS	International Nuclear Security
LLM	Large Language Model
PNNL	Pacific Northwest National Laboratory

Intentionally blank

1. BACKGROUND

Conducting security assessments can be a time-consuming process. The intent of the work leading up to this document was to evaluate if AI large language models (LLMs) could aid an assessment team in identifying new or novel attack pathways, improve performance, optimize information gathering, and/or shorten the time needed to conduct a security assessment. In the context of this document, a security assessment encompasses anything an assessor might evaluate when trying to discover risk (e.g., process or procedures, cybersecurity, physical security, personnel security, policy weakness, or compliance).

To assess the LLMs, the team did not use a specific assessment framework or methodology; rather, they relied on previous assessment experience and identified three different types of information that an assessor might collect and analyze where an LLM might prove useful.

- Human-focused files in which the intended consumer of the information is a human reader.
- Computer-focus files where a computer is the intended consumer.
- Image files where images are used to illustrate system designs or builds.

This document is organized into two distinct sections. First the document takes the reader through a use case of applying AI to an assessment of a realistic training system with the details of what was done including the tools and systems used. The second section can be used as a guide for those interested in using AI for their own security assessments. It is targeted at novice users and provides background information on large language models (LLMs), security considerations, and recommendations of how LLMs can be used in practice when conducting security assessments in a nuclear system context.

The document concludes with areas of research that could further apply the usage of AI/LLMs in nuclear security assessments.

The recommendations and guidance provided within this document are derived from observations collected during this process.

2. USE CASE: MODEL ACCESS CONTROL SYSTEM

This section details the methodology and environment utilized by Pacific Northwest National Laboratory (PNNL) to evaluate the suitability and applicability of various LLMs for use in conducting security assessments on nuclear systems. The objective was to test LLM capabilities in a controlled, realistic setting that mitigated the risks associated with sensitive data while still providing meaningful insights into their performance.

2.1. THE MODEL ACCESS CONTROL SYSTEM TESTBED

To facilitate safe and effective testing, PNNL employed a representative model of an Access Control Systems (ACS) specifically developed for training purposes. This testbed offered a controlled environment to simulate assessment scenarios without exposing actual critical infrastructure data.

2.1.1 ACCESS CONTROL SYSTEM TESTBED OVERVIEW

Purpose: The ACS training model was designed to simulate a high-security facility's access control system, incorporating commercial components and licensed software typical of such environments.

Intentional Vulnerabilities: The system deliberately included security-related defects or vulnerabilities, mimicking what assessors would be expected to discover in a real-world scenario. This allowed for objective evaluation of the LLMs' ability to identify such issues.

Components: The ACS testbed integrates hardware and software, consisting of both industry-standard ACSs and commercial, of-the-shelf (COTS) components that might be used in real-world security environments. It relies on a mixture of information technology and operational technology to restrict access to vital plant operations and serves as a working collection of subsystems that are used by security assessment teams to learn the process of conducting vulnerability assessments.

Representativeness: The use of commercial components and licensed software ensured the model's relevance to real-world operational systems, making the test results applicable to actual security assessment contexts.

2.1.2 SUPPORTING DOCUMENTATION

To provide targeted LLMs with realistic context for our evaluation of their suitability for accelerating security assessments, the PNNL team relied on a comprehensive suite of fictitious documentation that was created for a hypothetical nuclear facility. This documentation was developed over several years of security assessment course development, achieving a level of depth that mimics real-world conditions. This documentation includes the following:

- National Regulations
- Physical Maps and Plan/Component Diagrams
- Photos of Equipment and Cabling
- Equipment Data Sheets
- Corporate Policies and System Overviews
- Procedures and Vendor Manuals
- Access Logs, Network Diagrams, and Firewall Rules.

All documentation was designed to be representative of real-world conditions while remaining entirely fictitious, ensuring that there were no actual sensitivities that would restrict the information shared, the types of LLMs used, or the questions posed.

2.2. LARGE LANGUAGE MODEL EVALUATION METHODOLOGY

The evaluation process focused on assessing how different LLMs performed with various modalities and complex assessment tasks, consistent with the needs of security assessment teams. Models were selected or excluded based on their suitability for these purposes. The evaluation was scoped with the intention of NOT being a rigorous comparative analysis of specific models, but rather as an exploration of the utility of LLMs for general assessment purposes. This scope was established primarily due to the rapid state of change present in current LLMs.

Assessing LLM Performance

The evaluation process systematically assessed the utility of each LLM's output against defined criteria. This included verifying the factual accuracy of responses, specifically identifying instances of hallucination or inaccurate information. Furthermore, contextual accuracy was rigorously assessed, ensuring the LLM's understanding aligned with the specifics of the system under assessment. Responses were also scrutinized for completeness, noting any significant gaps or omissions. A critical factor in ranking utility was the depth of detail provided; outputs were evaluated on their capacity to offer analytical insights beyond superficial observations (e.g., distinguishing actionable findings from mere descriptions of obvious system components). Additionally, the evaluation considered the range of input formats each LLM could effectively process.

Ultimately, two overarching questions guided the final assessment of each model:

1. Does the level of accuracy and reasoning demonstrated by this LLM exhibit practical utility for an assessor conducting a security assessment?
2. Are there any identifiable deficiencies within the model that would significantly limit its effective application in assisting in conducting security assessments?

2.2.1 MODEL SELECTION

NOTE: This document does not endorse, recommend, or support the use of specific LLMs for the purposes of conducting critical infrastructure security assessments.

Testing was conducted between October 2025 and February 2026, and was based on the following LLMs:

- **ChatGPT 4.0** – A now deprecated multimodal LLM that can process and generate text, images, and audio. It is not optimized for deep analytical reasoning.
- **ChatGPT 5.0** – Then OpenAI's most advanced AI model, offering state-of-the-art performance across coding, math, writing, health, and visual perception.
- **Claude Sonnet 4** – Then Anthropic's hybrid reasoning model designed for complex tasks and multi-stage operations. Reported to excel in code generation, reasoning, and instruction-following and is focused on programming.
- **Gemini 2.5 Pro** – Then Google's most advanced reasoning AI model, designed to handle complex problem solving across datatypes, including text, audio, images, and video. It reportedly excels in coding, math, and science, utilizing a 1-million token context window to process large datasets and perform advanced reasoning tasks.

Excluded Models

- **Microsoft Copilot** – Initially considered but was ultimately excluded due to lack of clarity on the underlying LLM version used, which was modified several times during the assessment window. Copilot also had restrictive limitations on the number of files that could be evaluated per query.

Disclaimer

Our use of these models to perform this assessment occurred during a very specific period, during which substantial changes were made to LLMs across the board. Our assessment of LLM suitability for critical infrastructure security assessments should not be viewed as an endorsement of specific models.

2.3. EVALUATION ACROSS DATA MODALITIES

The analysis focused on three primary data modalities to test varied LLM capabilities: human-focused files, computer-focused files, and image files. Quality prompts were engineered specifically for each modality and task, demonstrating that time invested in learning and applying prompt engineering techniques (see Section 5) significantly enhanced the quality and relevance of the responses received across all tested LLMs.

2.3.1 HUMAN-FOCUSED FILES

Assessors frequently require context and the ability to reason across multiple documents to identify vulnerabilities not immediately apparent from a single source. For this modality, each LLM was provided with the same 13 files, representative of what might be evaluated during an assessment. These included a system description, system layout drawings, wiring diagrams, corporate policies, national regulations, a previous vulnerability scan report, and a security zone map. The goal was to assess the LLMs' ability to summarize data and identify vulnerabilities by reasoning across these diverse documents.

- **ChatGPT 4.0:** Provided general attack pathways applicable to any facility, highlighting specific hardware and software identified in the documentation. The conclusions were largely generic, similar to what might be expected from a security methodology or a beginner in security best practices without extensive experience.
- **ChatGPT 5.0:** Demonstrated strong performance, delivering useful and prioritized information. Its output included an executive summary with key findings, highlighted specific actionable tasks for risk reduction, and clearly identified the assumptions made by the model, indicating areas needing further information. This model successfully identified specific problematic models, protocols, and configurations, suggesting further investigation and mitigating actions.
- **Claude Sonnet 4:** Provided the most organized response, featuring an executive summary, a system overview and analysis, and identified primary risks, along with time-based mitigating actions. While its information was not as detailed or actionable as ChatGPT 5.0, its structured output was noted as being highly beneficial for accelerating assessment report writing.
- **Gemini 2.5 Pro:** Similar to ChatGPT 4.0, Gemini 2.5 Pro offered general attack pathways. It summarized findings and provided somewhat more detailed information than ChatGPT 4.0, but its overall output remained generic compared to the more advanced models.

2.3.2 COMPUTER-FOCUSED FILES

Security assessments often necessitate the analysis of specialized files such as network data captures (.pcap), firewall configurations (.conf), or system logs. While the LLMs could not directly interpret binary files like .pcap or system logs, they consistently provided quality recommendations

for the application software required to evaluate such files, along with instructions on effective tool usage. For a firewall configuration stored in a text file alongside an image file of the network architecture, the LLMs varied significantly in their analytical capabilities.

- **ChatGPT 4.0:** Performed well in parsing the firewall configuration file, answering questions about existing vulnerabilities, rules assigned to interfaces, and enumerating users with their access levels. However, it encountered a functional freeze when attempting to analyze a .png network diagram, limiting its utility for combined text and image analysis in this modality.
- **ChatGPT 5.0:** Proved to be one of the most well-rounded models. It was highly adept at interpreting the .png network diagram and produced the most thorough analysis. Notably, it was the only model to organize vulnerability findings by severity level and to generate a table detailing allowed traffic and protocols across each subnet, which was considered an invaluable aid for human analysts.
- **Claude Sonnet 4:** Exhibited performance similar to ChatGPT 4.0. Its output was often presented in more technical language but occasionally prioritized length and jargon over accuracy. A notable issue was its failure to recognize that ANY-ANY rules were disabled in the configuration, and thus not an active security concern. Claude's strength lay in its ability to write valid firewall rules when prompted and was unique in correctly identifying the dual-port requirement for active mode FTP sessions (TCP ports 20 and 21). Its primary value was for code generation, making it less suitable as a general analytical tool compared to others.
- **Gemini 2.5 Pro:** While designed for image analysis, Gemini also demonstrated decent analytical capabilities for parsing and explaining the firewall configuration. However, its suggestions for improvement were often overly burdensome or impractical, such as recommending individually created rules for every client-to-server connection on a subnet. Its advice in this domain required significant skepticism.

2.3.3 IMAGE FILES

Assessors often compare as-designed documentation with as-built conditions to identify discrepancies. To test this, LLMs were provided with photos of an ACS cabinet (from multiple angles and lighting) and corresponding wiring diagrams. The cabinet intentionally contained wiring errors and undocumented hardware.

- **ChatGPT 4.0:** Offered generic observations, highlighted the absence of tamper detection, and suggested improved wire management. Its insights were largely superficial and readily apparent to a seasoned assessor.
- **ChatGPT 5.0:** Accurately identified aspects of the image it could not fully discern, correctly pinpointed undocumented hardware, and recognized associated security concerns. It recommended additional improvements and provided a prioritized action list, proving valuable for documenting findings.
- **Claude Sonnet 4:** Focused primarily on suspected wiring discrepancies and the lack of tamper protection. It failed to identify the undocumented hardware, offering minimal aid in comparing as-designed with as-built states.

- **Gemini 2.5 Pro:** Provided a useful reporting format with an executive summary, key findings, and reasonable prioritized responses. It successfully identified the undocumented hardware and suggested recommended mitigations, demonstrating strength in visual analysis.

2.3.4 OVERALL OBSERVATIONS ON PROMPT ENGINEERING

Across all models and modalities, the quality of the prompt entered significantly influenced the quality of the response received. Time spent learning and applying prompt engineering techniques (as detailed in Section 5) paid off significantly, leading to more relevant, accurate, and actionable outputs from the LLMs.

2.4. LARGE LANGUAGE MODEL PERFORMANCE COMPARISON ACROSS DATA MODALITIES

The evaluation of selected LLMs against various data modalities provided crucial insights into their comparative strengths and weaknesses within the context of security assessments. Table **Error! No text of specified style in document.** synthesizes the performance of each model—ChatGPT 4.0, ChatGPT 5.0, Claude Sonnet 4, and Gemini 2.5 Pro—across human-focused, computer-focused, and image file analyses, highlighting their respective capabilities and deficiencies in identifying vulnerabilities, structuring information, and providing actionable intelligence.

Table Error! No text of specified style in document. **LLM Performance Comparison Across Data Modalities**

Attribute	Human-Focused Files (PDFs, Policies)	Computer-Focused Files (Configs, PCAP)	Image Files (Photos, Diagrams)
ChatGPT 4.0			
Strengths	Provided general attack pathways; highlighted specific hardware/software.	Good at parsing firewall configs, identifying vulnerabilities, rules, and enumerating users.	Offered generic observations regarding tamper detection and wire management.
Weaknesses	Generic conclusions, similar to beginner-level analysis; lacked depth.	Froze when attempting to analyze .png network diagram; limited combined text/image analysis.	Insights were superficial and readily apparent to an experienced assessor; lacked specific detail.
ChatGPT 5.0			
Strengths	Strong performance; useful, prioritized, and actionable info; executive summary with key findings; identified assumptions; specific problematic components/protocols.	Most well-rounded; highly adept at .png network diagram interpretation; most thorough analysis; organized vulnerabilities by severity; generated traffic/protocol table (invaluable for human analysts).	Accurately identified indiscernible aspects; correctly pinpointed undocumented hardware and security concerns; recommended improvements; provided prioritized action list.
Weaknesses	*None noted.*	*None noted.*	*None noted.*

Attribute	Human-Focused Files (PDFs, Policies)	Computer-Focused Files (Configs, PCAP)	Image Files (Photos, Diagrams)
Claude Sonnet 4			
Strengths	Most organized response; executive summary, system overview/analysis, primary risks, time-based mitigations; beneficial for report writing acceleration.	Good at writing firewall rules; correctly identified dual-port FTP requirement.	Focused on suspected wiring discrepancies and lack of tamper protection.
Weaknesses	Less detailed or actionable than ChatGPT 5.0.	Occasionally prioritized jargon over accuracy; failed to identify disabled rules; less suitable as a general analytical tool.	Failed to identify undocumented hardware; offered minimal aid for as-designed vs. as-built comparison.
Gemini 2.5 Pro			
Strengths	General attack pathways; summarized findings with some detail.	Decent analytical capabilities for parsing/explaining firewall configs.	Useful reporting format (summary, key findings, prioritized responses); successfully identified undocumented hardware and suggested mitigations.
Weaknesses	Overall output remained generic compared to more advanced models.	Suggestions for improvement were often burdensome/impractical (e.g., individual rules for each client-to-server connection); required significant skepticism for its advice.	*None noted.*

The evaluation demonstrated that advanced multimodal LLMs, particularly ChatGPT 5.0, exhibited a superior capability for comprehensive analysis, deep reasoning, and structured output generation across diverse data types, proving highly beneficial for complex assessment tasks. Other models showed specialized strengths, such as Claude Sonnet 4 for code generation or Gemini 2.5 Pro for specific visual analysis, but often presented limitations in overall analytical depth or practicality of recommendations. This reinforces that the choice of LLM should be based on the specific assessment task and data modality, and that constant human oversight is essential to validate outputs and filter out generic or impractical advice.

2.5. DETAILED CASE STUDIES: PROMPTS, RESULTS, AND CONCLUSIONS

To further illustrate the practical application and performance nuances of LLMs in security assessments, this section presents detailed case studies from each primary data modality. These examples demonstrate the types of prompts used, the specific results obtained from the various LLMs, and the conclusions drawn regarding their suitability for particular assessment tasks.

2.5.1 CASE STUDY: HUMAN-FOCUSED FILES – COMPREHENSIVE SECURITY ASSESSMENT

This case study demonstrates the LLMs' ability to synthesize information from multiple text-based documents to identify vulnerabilities and propose mitigations for a fictitious ACS.

Prompts Used

Prompt 1 Used:

"You are a nuclear security professional with twenty years of experience conducting holistic security assessments that include both physical security and cyber security. You have been asked to conduct a security assessment on the access control system at Vitricity Nuclear Corporation. The attached materials provide background information to enable the security assessment. Please provide a report that defines the pathways that could lead to a compromise, highlight the areas and pathways most susceptible to compromise, and recommend mitigations that would fix the problems identified. Feel free to ask questions that will help clarify what is desired. If there are better AI engines for assisting with this task please recommend which to use for specified tasks. For example image analysis."

Prompt 2 Used:

"You are a nuclear security professional with over twenty years of experience conducting comprehensive security assessments, specializing in both physical security and cybersecurity. You have been tasked with conducting a detailed security assessment of Vitricity's access control system to identify vulnerabilities and recommend mitigations. Using the attached background materials, analyze the access control system to:

- *Identify and define all potential attack pathways that could lead to a compromise, including insider threats, external attacks, and technical failures.*
- *Pinpoint the areas and pathways that are most susceptible to compromise, supported by a risk assessment and specific examples, where possible.*
- *Provide actionable, prioritized recommendations to mitigate the identified vulnerabilities, ensuring consistency with industry best practices in nuclear security. Additionally, include a brief summary defining the overall risk level of the system, highlighting key findings and providing practical guidance for implementation and ongoing monitoring. Please ask precise and targeted questions if additional information or clarification is needed to ensure the assessment is thorough and accurate. Where appropriate, articulate assumptions based on the limitations of the provided materials and emphasize areas where further investigation may be required."*

Results Obtained

- **ChatGPT 4.0:** Generated general attack pathways applicable to any facility, identifying specific hardware and software from the documentation. The findings were largely superficial and aligned with a generic assessment methodology.

- **ChatGPT 5.0:** Provided highly useful and prioritized information, including an executive summary, key findings, and actionable tasks for risk reduction. It distinguished itself by identifying specific problematic models, protocols, and configurations, along with assumptions and areas needing further information.
- **Claude Sonnet 4:** Produced a well-organized response with an executive summary, system overview, analysis, primary risks, and time-based mitigations. While structured, its content was less detailed and actionable compared to ChatGPT 5.0.
- **Gemini 2.5 Pro:** Offered general attack pathways and a summary of findings but remained generic in its insights, providing less practical value than more advanced models.

Conclusions:

The more detailed and carefully structured "Prompt 2" yielded significantly better results, especially from advanced models like ChatGPT 5.0. This demonstrated that LLMs can provide substantial support for synthesizing complex human-focused documents and identifying vulnerabilities, provided the prompt guides them effectively. ChatGPT 5.0 consistently provided the most actionable and comprehensive analysis for this task.

2.5.2 CASE STUDY: COMPUTER-FOCUSED FILES – FIREWALL CONFIGURATION ANALYSIS

This case study evaluated the LLMs' ability to interpret technical configuration files and network diagrams to identify security concerns and communication flows.

Prompts Used

Prompt 1 Used:

"I am an experienced auditor conducting a cyber security assessment of a nuclear facility with a complex network infrastructure. Please analyze this file for security concerns and summarize network communication flows."

Prompt 2 Used:

"Continuing my role as a cyber security expert. Please review this network diagram and highlight any inconsistencies with the pfsense firewall configuration. Ask any clarifying questions needed."

Results Obtained

- **ChatGPT 4.0:** Successfully analyzed the firewall configuration text file to identify vulnerabilities, rule assignments, and user access levels. However, it failed to process an attached .png network diagram, limiting its utility for visual-text correlation.
- **ChatGPT 5.0:** Excelled in this modality, adeptly processing both the firewall configuration and the .png network diagram. It provided the most thorough analysis, organizing vulnerability findings by severity and generating a table of allowed traffic and protocols across subnets, which was highly beneficial for human analysts.

- **Claude Sonnet 4:** Performed similarly to ChatGPT 4.0 in text analysis but struggled with prompt interpretation, sometimes prioritizing jargon over accuracy. It uniquely identified the dual-port requirement for active mode FTP and was proficient at generating new firewall rules but missed identifying disabled rules.
- **Gemini 2.5 Pro:** Showed decent analytical capabilities for the firewall configuration. While designed for image analysis, its output was often accompanied by impractical suggestions for improvement (e.g., individual rules for every client-to-server connection), requiring significant human skepticism.

Conclusions

LLMs could effectively analyze text-based configuration files, but the ability to handle supplementary visual data (like network diagrams) critically influenced the depth of analysis. ChatGPT 5.0 demonstrated superior multimodal capabilities, providing highly actionable insights. While Claude Sonnet 4 showed promise for code generation, its overall analytical depth for identifying inconsistencies was limited.

2.5.3 CASE STUDY: IMAGE FILES – AS-DESIGNED VERSUS AS-BUILT DISCREPANCY

This case study explored the LLMs' capability to identify discrepancies between documentation and physical reality using images, a common task in security assessments.

Prompts Used

Prompt 1 Used:

"I am an experienced cyber security professional conducting a cyber security assessment of a nuclear facility. I have attached a photo of an industrial controls panel from the Access Control System inside the facility, as well as the wiring diagram showing the layout of the terminal blocks. Analyze the image according to the diagrams and identify any discrepancies."

Prompt 2 Used:

"As an experienced security engineer conducting an assessment. I need to compare as-built with as-designed. Two documents attached are the design documents and the three images are the cabinet where the system is installed. Your tasks are to identify differences, for each difference explain any security impacts, highlight potential vulnerabilities, and prioritize any recommendations. Explicitly state any assumptions made due to limited visibility or missing information in the images. Do not fabricate non-visible elements only infer cautiously and label inferences clearly."

Results Obtained

- **ChatGPT 4.0:** Delivered generic observations such as suggesting improved wire management or noting the lack of tamper detection. Its analysis was largely superficial and self-evident to an experienced human assessor.

- **ChatGPT 5.0:** Demonstrated strong visual analysis, accurately identifying elements it could not clearly identify. It correctly pinpointed undocumented hardware and associated security concerns and provided both recommended improvements and a prioritized action list, proving highly useful for documentation.
- **Claude Sonnet 4:** Focused mainly on potential wiring discrepancies and the absence of tamper switches/seals. Critically, it overlooked the undocumented hardware, limiting its value for detailed as-designed versus as-built comparisons.
- **Gemini 2.5 Pro:** Provided well-formatted output, with an executive summary and key findings. It successfully identified the undocumented hardware and offered reasonable mitigations, showcasing its strengths in image recognition and analysis.

Conclusions

The ability of LLMs to effectively compare visual data with documentation varied significantly. ChatGPT 5.0 and Gemini 2.5 Pro demonstrated superior capabilities in identifying physical discrepancies and security implications from images, particularly when prompts were specific and guided the model to look for differences and their impacts.

2.5.4 SUMMARY OF OBSERVATIONS ON DETAILED CASE STUDIES

The detailed case studies across human-focused, computer-focused, and image data modalities consistently revealed several overarching themes. First, the quality of prompt engineering emerged as a decisive factor in eliciting actionable and relevant results, underscoring the principles detailed in Section 5. Second, the evaluation highlighted a clear hierarchy of LLM performance, with advanced multimodal models like ChatGPT 5.0 generally demonstrating superior analytical capabilities, depth of insight, and structured output generation across diverse data types. While other models exhibited specialized strengths, such as Claude Sonnet 4's proficiency in code-related tasks or Gemini 2.5 Pro's visual processing, their overall utility and the practicality of their recommendations varied. This underscores that while LLMs offer significant potential to enhance security assessments, their optimal application requires careful model selection aligned with specific task requirements, a nuanced understanding of their limitations, and continuous human validation of all outputs to filter out generic, impractical, or erroneous information.

3. GUIDE ON USING AI LLMS WITHIN SECURITY ASSESSMENTS

The integration of Large Language Models (LLMs) into nuclear security assessments presents both significant opportunities and inherent risks. To maximize their benefits while rigorously safeguarding critical assets and information, the following core principles must strictly govern their application:

1. **LLMs as Assistants, Not Decision-Makers:** LLMs are advanced computational tools designed to augment human capabilities by processing information and generating text. They are powerful assistants for analysis, organization, and drafting, but they are not infallible authorities. All substantive security judgments, critical decisions, and final validations must remain with qualified human experts who possess the necessary domain knowledge,

situational awareness, and accountability. LLMs do not "understand" in a human sense and cannot replace expert human oversight.

2. **Protect Sensitive Information:** The confidentiality, integrity, and availability of sensitive information are paramount in nuclear security. This includes, but is not limited to, facility-specific details, operational procedures, personnel data, and classified material. No sensitive or classified information, nor any data that could inadvertently reveal such details, should ever be entered into an LLM without explicit, prior organizational authorization and only on approved, properly configured, and protected platforms. Rigorous redaction and anonymization protocols are essential before any sensitive data interaction.
3. **Verify All Outputs:** LLM-generated content must always be treated as preliminary draft input or suggestions. Users must independently and meticulously verify every piece of information, claim, or recommendation against authoritative sources, local requirements, and real-world conditions. A critical security conclusion or operational recommendation must never be accepted solely because an LLM generated it. This vigilance is crucial to mitigate "hallucinations" (plausible-sounding but false statements) and ensure accuracy and reliability.

These principles form the bedrock of responsible LLM integration, enabling enhanced efficiency, consistency, and depth in security assessments while upholding the highest standards of nuclear security.

3.1. WHY THIS REPORT MATTERS: ACCELERATING NUCLEAR SECURITY ASSESSMENTS WITH AI

This report addresses the timely and critical need to strategically leverage advanced technologies like AI/LLMs in enhancing global nuclear security. In an era characterized by evolving threats and increasing complexity in protecting nuclear facilities and material, the ability to conduct security assessments with greater efficiency, consistency, and depth is paramount.

The insights and recommendations presented herein are highly relevant for several key reasons:

- **Responding to Evolving Threats:** The landscape of nuclear security threats is dynamic, encompassing advanced cyber-physical attacks, insider threats, and sophisticated adversarial tactics. Traditional assessment methodologies, while robust, can be resource-intensive and may struggle to keep pace with these rapidly changing challenges. This report demonstrates how AI/LLMs can act as force multipliers, enabling more proactive and comprehensive identification of vulnerabilities.
- **Optimizing Resource Allocation**: Every nuclear security program has finite resources. By illustrating how LLMs can streamline various stages of security assessments—from planning and document review to risk identification and report generation—this guide offers pathways to optimize human expert time and bandwidth. This efficiency gain allows experts to focus on the most complex analytical tasks and strategic decision-making, thereby enhancing overall mission effectiveness.
- **Fostering Consistency and Quality:** Security assessments often involve diverse teams and methodologies, which can sometimes lead to variability in findings and reporting. This report highlights LLMs' capacity to generate standardized checklists, assist in structured

data analysis, and promote consistent language in documentation. This capability can lead to more uniform, high-quality, and defensible assessment outcomes across international partnerships.

- **Informing Strategic Technology Adoption:** As AI rapidly integrates into various sectors, a clear understanding of its appropriate and secure application within nuclear security is essential. This report provides a foundational framework, based on Pacific Northwest National Laboratory (PNNL)'s comprehensive testing, for readers to make informed decisions regarding AI tool adoption, policy development, and training initiatives. It emphasizes critical safeguards, including data protection protocols, expert human oversight, and the mitigation of "hallucinations" and dual-use risks, ensuring that AI deployment aligns with the nuclear security requirements.
- **Building a Foundation for Future Innovation:** This guide not only addresses current applications but also identifies areas for future research and development, particularly in advancing secure and autonomous LLM applications and robust evaluation methodologies. For INS leadership, this lays the groundwork for continuous improvement and sustained technological leadership in international nuclear security cooperation.

In essence, this report serves as a valuable resource, providing actionable insights into how to harness the transformative potential of AI/LLMs to strengthen nuclear security global efforts, ensuring that assessment capabilities remain cutting-edge, efficient, and resilient against future threats.

3.1.1 CONTEXUTALIZING LARGE LANGUAGE MODEL TESTING FOR SECURITY ASSESSMENTS

The integration of LLMs into security assessment methodologies presents both significant opportunities for efficiency gains and complex challenges, particularly concerning the handling of sensitive information within critical infrastructure. This section provides a high-level overview of the foundational approach and rationale that guided the testing of LLMs in realistic security assessment scenarios. The primary objective of this work was to establish a comprehensive understanding of how LLMs can augment human expertise while rigorously identifying their limitations and inherent risks.

Our testing of LLMs for applications in nuclear security assessments was conducted within a controlled environment designed to accurately simulate real-world conditions without compromising actual sensitive data. This systematic approach involved:

1. **Establishing a Representative Testbed:** We utilized an existing model system, specifically an access control system (ACS) testbed, which closely mimics the complexity and components of critical infrastructure. This testbed was configured with intentional vulnerabilities and realistic setups to rigorously challenge the LLMs' analytical capabilities.
2. **Developing Fictitious but Realistic Documentation:** To provide the LLMs with the necessary context, a comprehensive suite of fabricated documentation was developed and employed. This documentation covered a broad spectrum, including system designs, operational manuals, policies, and network diagrams, achieving high fidelity to real-world counterparts to ensure meaningful test results.

3. **Defining Evaluation Modalities:** Recognizing that LLMs' capabilities vary significantly across different data types, our testing encompassed multiple modalities. These included human-focused textual files (e.g., policies, reports), computer-focused technical files (e.g., configurations, logs), and image files (e.g., photos, diagrams). This multimodal approach ensured a holistic assessment of LLM performance.
4. **Assessing LLM Performance:** The evaluation focused on several key aspects, gauging the LLMs' ability to
   - **Identify Vulnerabilities:** Determine how effectively LLMs could pinpoint known security weaknesses within the provided documentation and testbed context.
   - **Synthesize Information:** Evaluate the LLM's capacity to reason across disparate documents and data types to form cohesive assessments.
   - **Provide Actionable Insights:** Assess whether the LLMs generated recommendations that were practical, prioritized, and relevant for security assessors.
   - **Address Prompt Sensitivity:** Analyze how variations in prompt engineering impacted the quality and utility of LLM outputs.

The insights gleaned from this systematic testing are instrumental in developing best practices for LLM deployment in sensitive environments. These findings inform decisions regarding model selection, necessary security safeguards, and the development of robust human-AI collaborative workflows. The ultimate goal of this research was to demonstrate how LLMs can enhance the speed, consistency, and depth of security assessments while always ensuring that human experts retain oversight and final decision-making authority, particularly in contexts where national security and public safety are paramount.

For a detailed breakdown of the specific processes, models evaluated, and comprehensive results, please refer to Section 2's Use Case.

3.2. OVERVIEW: AI

This guide addresses the application of AI, specifically LLMs, within the domain of nuclear security assessments. AI, as a broad field, encompasses various technologies designed to simulate human intelligence, with LLMs being a significant subset capable of processing and generating human-like text. The intent of this document is to clarify how these advanced tools can augment the capabilities of security assessors by improving efficiency and consistency in their work while strictly adhering to established nuclear security protocols. It will delineate the capabilities and limitations of LLMs, emphasizing their role as support tools rather than autonomous decision-makers in sensitive environments. The guide also covers crucial considerations such as data protection, risk mitigation, and best practices for integrating AI into assessment workflows for critical infrastructure.

3.2.1 WHAT LARGE LANGUAGE MODELS ARE (AND ARE NOT)

LLMs represent a significant advancement in AI, offering capabilities that can greatly augment various tasks, including those within security assessments. However, to effectively and safely integrate these tools, it is imperative to clearly understand their fundamental nature, their strengths, and critically, their inherent limitations. This section clarifies what LLMs are designed to

do and, equally important, what they are not, particularly in the sensitive context of nuclear security.

3.2.2 WHAT LARGE LANGUAGE MODELS ARE

LLMs are advanced computational tools trained to generate probabilistically realistic responses to text-based human language input prompts. They can be used in an extensive range of tasks, including data processing, code generation, orchestration of other LLMs for multi-step problem solving, and text summary or distillation.

In security assessment contexts, LLMs can contribute across various stages, from initial planning to documentation. Their utility spans several key functional areas, leveraging their ability to rapidly synthesize and organize information.

LLMs, when deployed with proper oversight, have the potential to improve the efficiency, systematization, and accessibility of security assessment efforts. Table 1 highlights some specific capabilities and their practical applications.

Table 1. List of Potential Use Cases for LLMs in Security Assessment Contexts

Category of LLM Capability	Specific Applications in Security Assessments
Drafting and Summarizing Text	Producing first drafts of assessment plan sections and reports.
	Summarizing lengthy background documents.
	Creating plain language explanations of technical concepts for non-technical audiences.
	Translating technical information into business-friendly language.
Organizing and Structuring Information	Turning bullet-point notes into coherent outlines.
	Grouping observations, findings, and recommendations into logical categories.
	Suggesting structures for assessment reports or briefing materials.
Suggesting Generic Checklists, Frameworks, and Questions	Proposing question sets for interviews (e.g., for access control, cyber-physical interfaces, maintenance practices).
	Drafting assessment frameworks or evaluation criteria based on widely recognized or open standards.
	Brainstorming potential lines of inquiry or topics to cover during planning.
Supporting "First Pass" Review and Brainstorming	Providing alternative formulations of observations or recommendations.
	Highlighting possible gaps or issues in a generic scenario for human review.

Category of LLM Capability	Specific Applications in Security Assessments
	Offering examples or analogies to support training and exercises.

3.2.3 WHAT LARGE LANGUAGE MODELS ARE NOT

While LLMs offer significant capabilities, it is equally crucial to understand their inherent limitations, particularly in the sensitive domain of nuclear security. Table 2 outlines some key areas where LLMs fall short and clarifies their boundaries.

Table 2. Limitations of LLMs in Security Assessment Contexts

Area of Limitation	LLM Specific Boundaries and Risks
Not Subject Matter Authorities	Can produce confident, detailed answers that are incomplete, outdated, or incorrect.
	Lacks authoritative knowledge of national regulations, site-specific designs, or threat environments.
	May "hallucinate" plausible but false references, requirements, or technical details.
Not Replacements for Security Experts	Cannot conduct site visits, interpret real-world conditions, or weigh trade-offs between safety, security, and operations.
	Cannot determine if a facility is adequately protected or meets national/international requirements.
	LLMs cannot be held accountable for security assessments; therefore, all substantive security judgments must remain with qualified human experts.
Do Not "Understand" in a Human Way	Operate by predicting patterns in training data; do not exhibit situational awareness, intent, or conceptual reasoning.
	Cannot reliably distinguish between safe and unsafe uses of generated information; require human guidance and supervision.
	Does not independently recognize information sensitivity, classification levels, or legal obligations.

Because of these fundamental limitations, LLM outputs must always be treated as draft input to human review and judgment—not as final conclusions. This is especially important when LLMs are used for security assessments of nuclear and other high-risk facilities.

3.3. SECURITY & PRIVACY RISKS WHEN USING LARGE LANGUAGE MODELS

When integrating LLMs into security assessment processes, particularly for critical infrastructure, such as nuclear facilities, the paramount concern is the robust protection of information. The

unauthorized exposure, modification, or destruction of sensitive data carries direct implications for national security, public safety, and international non-proliferation efforts.

While LLMs offer powerful analytical capabilities, their use introduces new vectors for potential data compromise, necessitating a comprehensive understanding of what constitutes sensitive information and how its integrity must be preserved across diverse operational and regulatory landscapes. This section delineates the critical security and privacy considerations relevant to all stakeholders involved in international nuclear security, ensuring a common understanding of risk profiles and protective measures.

What Is Information Security?

Information security, in its broadest sense, encompasses the measures and practices designed to protect information from unauthorized access, use, disclosure, disruption, modification, or destruction. Within the context of nuclear security and critical infrastructure, this discipline is crucial for safeguarding the confidentiality, integrity, and availability of data. It involves deploying technical controls, establishing robust policies, and fostering a deep understanding among personnel regarding the value and vulnerability of information assets. Effective information security ensures that critical operational, design, and procedural details remain protected against threats, whether deliberate or accidental. For nuclear security, maintaining strict control over information is not merely a regulatory compliance issue but a fundamental imperative to prevent malicious acts and maintain public trust.

The International Atomic Energy Agency (IAEA) defines sensitive information as, "*Information, in whatever form, including software, the unauthorized disclosure, modification, alteration, destruction, or denial of use of which could compromise nuclear security.*"[1]

3.3.1 WHAT COUNTS AS SENSITIVE?

In the context of protecting nuclear facilities and critical infrastructure, certain types of information, if compromised, could pose significant security risks. Understanding and categorizing these sensitive data points is paramount to preventing unauthorized disclosure when utilizing LLMs. Table 3, while not exhaustive, broadly defines what is typically considered sensitive:

Table 3. Categories of Sensitive Information in Nuclear Security Assessments

Category of Sensitive Information	Examples of Specific Sensitive Data
Facility-Specific Details Information that reveals physical layout, internal workings, or security system configurations of a facility.	Site-specific drawings, diagrams, sketches, or maps of physical security system features.
	Alarm system layouts (e.g., intrusion detection, assessment equipment, wiring, power sources, duress alarms).
	Site-specific design features of plant security communications systems.

[1] IAEA Nuclear Safety and Security Glossary 2022 (Interim) Edition

Category of Sensitive Information	Examples of Specific Sensitive Data
	Details of cyber architecture, IP ranges, and system names.
Operational Details Information pertaining to daily functions, security posture, and identified vulnerabilities within a facility.	Physical security orders and procedures (e.g., duress codes, patrol routes, contingency responses).
	Physical or cybersecurity assessments detailing specific vulnerabilities, weaknesses, or gaps.
	Failsafe/fail-secure behaviors of systems.
	Inspection reports revealing uncorrected defects or vulnerabilities in systems.
Personnel and Personal Information Data related to individuals, their roles, and security permissions/ clearances, which could be exploited in security breaches.	Names, contact details, roles, and clearance levels of personnel.
	Disciplinary records and behavioral concerns.
Security Procedures and Contingency Plans Information outlining responses to security incidents, emergency protocols, and the deployment of protective forces.	The composite physical security plan for the facility or site.
	The composite safeguards contingency plan/measures for the facility or site.
	Information on on-site or off-site response forces (e.g., size, armament, arrival times).
	Detailed intrusion response procedures.
	Cybersecurity incident playbooks with specific timings and steps.
	Workarounds, emergency overrides.
Regulatory and Sensitive/Classified Content Any information subject to national laws or specific regulatory frameworks governing its protection.	Any information classified or restricted under national law.
	Safeguards and nuclear material accounting details.

In many jurisdictions, sharing such information with external services may be illegal or a serious policy violation.

3.3.2 TYPES OF LARGE LANGUAGE MODELS AND THEIR RISK PROFILES

The security posture and suitability of an LLM for processing sensitive information vary significantly based on its deployment model. Understanding these differences is critical for risk management when integrating AI into nuclear security assessments.

The following matrix details three generic deployment types, outlining their characteristics, typical risk profiles, and implications for sensitive data handling. It is essential to consult with IT and security teams to determine organizational approvals for each use case.

Table 4. LLM Deployment Models and Associated Risk Profiles

Always consult with IT/security teams to understand which type is approved for your use.

Deployment Model	Key Characteristics	Typical Risk Profile & Data Handling Implications
HIGHEST RISK Publicly accessible cloud services with general terms of service.		
Free Public Cloud LLMs	*Often consumer-facing web interfaces.*	**Risk Profile:** Inputs may be stored, logged, and potentially used to train and improve models, sometimes without straightforward consent or clear data retention policies.
	Minimal user control over data retention or privacy settings.	**Data Exposure:** Minimal to no control over data residency or processing locations. Automatically results in inadvertent disclosure.
Recommendation: Strictly unsuitable for sensitive, classified, or controlled information. Should only be used for openly available, non-sensitive information, if permitted by organizational policy.		
MEDIUM RISK Cloud services that are specifically designed for enterprise use, often with dedicated contracts and service level agreements.		
Enterprise or "Private" Cloud LLMs	*Vendor-managed infrastructure with enhanced security features and compliance certifications.*	**Risk Profile**: Offers improved controls and contractual guarantees regarding data handling, retention, and non-use for model training. Often allows users to disable logging and training on their inputs.
	Typically offers more granular access controls and audit capabilities.	**Controlled Sensitivity:** Can securely process some sensitive information when properly configured, with data retention policies aligned to organizational needs. However, specific authorization and rigorous vetting of vendor security practices are mandatory.
Recommendation: Unlikely to be suitable for classified information unless expressly authorized and tailored to meet specific national security requirements. Requires thorough assessment of vendor's security architecture and contractual terms.		
LOWEST RISK LLM models deployed and run entirely within an organization's isolated and controlled network environment.		
On-Premise "Air-Gapped" LLMs	*Not connected to external networks (air-gapped) or operates within highly restricted internal networks.*	**Risk Profile:** Provides the highest degree of control over data. No data leaves the organization's infrastructure, significantly reducing risks of external exposure or unauthorized third-party access.

Deployment Model	Key Characteristics	Typical Risk Profile & Data Handling Implications
	Requires significant internal investment in infrastructure, expertise, and ongoing maintenance.	**Maximum Control**: Organizational IT and security teams have full oversight of infrastructure, data residency, access controls, and security policies. Can be configured to meet the most stringent security and compliance requirements, including for sensitive information.

Recommendation: Ideal for processing sensitive or proprietary information where data sovereignty and maximum security are paramount. Requires substantial resource allocation for configuration, maintenance, and expert staffing.

The selection of an LLM deployment model must be directly commensurate with the sensitivity of the information handled and strictly adhere to organizational security policies and national regulations.

CAUTION: *Always consult with your IT and Security Teams before selecting and using an LLM.*

3.3.3 IMPLICATIONS FOR NUCLEAR SECURITY ASSESSMENTS

Integrating LLMs into nuclear security assessments necessitates a critical and nuanced understanding of their capabilities and, more importantly, their inherent limitations and associated risks. LLMs serve as powerful assistants for text processing and organization but are fundamentally not subject matter authorities or replacements for human experts.

The paramount concern remains the protection of sensitive information. Any unauthorized disclosure of facility-specific, operational, personnel, procedural, or regulatory data can directly compromise nuclear security, national interests, and international non-proliferation efforts. Therefore, the choice of LLM deployment model is a critical risk management decision. Public cloud LLMs pose unacceptable risks for sensitive data, while enterprise and on-premise solutions offer escalating levels of control necessary for safeguarded information.

Table 5. Appropriate Use Versus Limitations of LLMS in Nuclear Security Assessments

LLMS Can (Appropriate Uses)	LLMS Cannot (Critical Limitations)
✓ Help you think more systematically and communicate more clearly.	× Replace your national framework for nuclear security regulation and oversight.
✓ Accelerate preparation of generic materials, training content, and non-sensitive documentation.	× Serve as an authority on classification, safeguards, or physical protection requirements.
✓ Support internal capacity-building by explaining concepts at different technical levels.	× Make operational decisions about security measures, configurations, or posture.

In essence, while LLMs can enhance assessment efficiency for non-sensitive tasks and documentation, their application in nuclear security demands unwavering adherence to strict data

governance, continuous human oversight, and a clear recognition that all final security judgments must rest with qualified human experts operating within approved, secure environments.

4. WHERE LARGE LANGUAGE MODELS CAN HELP IN SECURITY ASSESSMENTS

Building upon the established understanding of LLM capabilities and limitations and the critical security considerations for sensitive information, this section details concrete and appropriate applications of LLMs across various stages of a security assessment life cycle. While LLMs must always function as adaptable assistants under expert human oversight, they can significantly enhance efficiency and systematic thinking. The following subsections describe specific use cases, offering example prompts designed to integrate LLMs effectively while rigorously maintaining security and confidentiality.

LLMs can add value at every stage of a security assessment if they are configured properly and used deliberately, on carefully controlled information, under expert oversight. They are best treated as adaptable assistants that help you work faster and more systematically—not as authorities on nuclear security or your national regulations.

In practice, this means using LLMs to draft and refine text, organize complex information, and support structured thinking, while keeping sensitive facility, operational, and personnel details out of unapproved tools. When combined with sound security judgment, they can

- Streamline preparation and planning by generating or refining assessment plans, checklists, and question sets based on general good practices.
- Help digest and review non-sensitive documents, such as redacted policies or public standards, to surface key points and obvious inconsistencies.
- Support high-level risk identification and scenario brainstorming without providing or exposing exploitable details.
- Assist during assessments by structuring sanitized—information that has been scrubbed of sensitive details—notes into clear findings using consistent language.
- Improve reporting and follow-up by drafting summaries, grouping issues thematically, and tailoring messages to different audiences.

The subsections that follow describe concrete, appropriate uses of LLMs across the assessment life cycle, with example prompts that illustrate how to keep both security and confidentiality at the forefront.

4.1. ASSESSMENT PLANNING AND PREPARATION

The initial stages of a security assessment, encompassing planning and preparation, are critical for defining scope, objectives, and methodologies. LLMs can serve as valuable tools during this phase, primarily by assisting in the creation of structured frameworks, adapting standard guidelines, and clarifying complex concepts (see Table 6). Their utility lies in generating foundational content and organizational structures that human experts can then refine, tailor, and approve, thereby streamlining the preparatory workload.

Table 6. Key Applications of LLMs in Assessment Planning and Preparation

Functional Area	Description of LLM Contribution
Clarifying Scope and Objectives	Translating high-level mandates into structured assessment objectives; drafting generic statements of work or preliminary plans for human refinement.
Developing Generic Frameworks and Checklists	Generating assessment domains, objectives, and example questions; mapping to recognized standards; brainstorming lines of inquiry for exercises.
Preparing Interview and Visit Materials	Drafting non-sensitive interview guides and question sets based on role-based descriptions; structuring planning notes into visit agendas.
Supporting Training and Pre-briefs	Producing plain language explanations of technical concepts for stakeholders; creating draft training materials on safe AI usage in assessment.

Preparation & Planning: Sample Prompt

"I'm preparing a security assessment for an new digital feedwater control system. Using general best practices only (no site-specific details), outline key categories I should assess for physical security, cybersecurity, personnel reliability, and incident response. Present the output as a checklist."

4.2. DOCUMENT PREPARATION AND REVIEW

Effective security assessments often entail extensive review and preparation of documentation, ranging from policies and procedures to system descriptions. LLMs can significantly streamline this phase by assisting in the analysis and synthesis of information, identifying potential inconsistencies, and providing clear summaries of complex textual data (see Table 7). This ultimately supports assessors in efficiently extracting key insights and preparing supporting materials.

Table 7. Key Applications of LLMs in Document Preparation and Review

Functional Area	Description of LLM Contribution
Summarizing Textual Data	Condensing lengthy policies, procedures, or regulations into concise overviews; assisting in drafting policy updates, training materials, or briefing notes.
Analyzing System Documentation (Non-Sensitive)	Reviewing and interpreting system documentation and descriptions to extract key features, functionalities, or potential interactions.
Identifying Inconsistencies (Non-Sensitive)	Pinpointing discrepancies or ambiguities in non-sensitive documents (e.g., contrasting policy statements with guideline provisions).

Document Preparation and Review: Sample Prompt

When utilizing LLMs for document review, it is paramount to ensure all sensitive details have been thoroughly removed (redacted) from any input provided. This prevents inadvertent exposure of critical facility, operational, or personnel information.

"As a senior security analyst focused on continuous improvement and high reliability of critical infrastructure, review the attached redacted version of our site access policy with sensitive details removed. Summarize the key controls and identify any obvious gaps or ambiguities from a security perspective."

4.3. RISK IDENTIFICATION & SCENARIO BRAINSTORMING

Risk identification and scenario brainstorming are crucial steps in understanding potential threats and vulnerabilities to a nuclear facility or critical infrastructure. LLMs can serve as effective catalysts in this creative and analytical process, stimulating divergent thinking and structuring complex risk data (see Table 8). They can assist human experts in exploring a broader range of potential risks and attack scenarios more efficiently, provided that strict security protocols are observed regarding the input of sensitive information.

Table 8. Key Applications of LLMs in Risk Identification and Scenario Brainstorming

Functional Area	Description of LLM Contribution
Brainstorming High-Level Threat Scenarios	Generating diverse and plausible threat scenarios, including various attack vectors and actors (e.g., criminal, state-sponsored, insider), *without* incorporating sensitive facility details.
Exploring "What If" Situations	Probing complex hypothetical situations such as cascading failures, dual failures (e.g., combined cyber/physical events), or unanticipated interactions between systems.
Structuring Risk Registers and Taxonomies	Assisting in the organization of risk data by suggesting categories (e.g., likelihood/impact, consequence types), frameworks (e.g., MITRE ATT&CK), and relationships between risks.

Risk Identification & Scenario Brainstorming: Sample Prompt

When employing LLMs for brainstorming risks, it is imperative to avoid providing any site-specific sensitive details that could inadvertently create exploitable information. Focus on generic descriptions of assets and threats.

"For a generic nuclear power plant (no specific site) operating in a country with a high criminal but minimal terrorist threat, list potential combined cyber and physical attack scenarios that security planners might consider. Do not provide step-by-step instructions, only high-level description of risks and required control types."

4.4. SUPPORT DURING THE ASSESSMENT

During the active phase of a security assessment, LLMs can act as intelligent assistants, helping to process and organize information collected in real-time (see Table 9). Their utility lies in transforming raw, often unstructured, assessment notes into coherent findings and observations. This support, however, is strictly confined to assisting with data structuring and language refinement, always under the direct supervision and expert judgment of human assessors, and always with sanitized or non-sensitive data inputs.

Table 9. Key Applications of LLMs in Support During the Assessment

Functional Area	Description of LLM Contribution
Structuring Raw Notes into Findings	Converting unstructured data gathered from walkthroughs, interviews, or observations (appropriately sanitized) into structured findings or identified issues.
Suggesting Standardized Wording	Proposing consistent and clear language for observations, findings, and recommendations, which helps ensure uniformity across assessment reports.
Formulating Non-Leading Interview Questions	Assisting in the development of neutral and unbiased follow-up questions for personnel, based on generic observations, to elicit detailed information without bias.

Assessment Support: Sample Prompt

It is critical that any notes or observations provided to the LLM for structuring or analysis during an assessment are thoroughly sanitized to remove site names, locations, detailed layouts, or any information that could compromise security.

"Given these generic observations from a security walkthrough (no site names, locations, or detailed layouts), draft three neutral interview questions to clarify how access control is actually implemented."

4.5. REPORTING & FOLLOW-UP

The final phase of a security assessment involves effectively communicating findings, recommendations, and follow-up actions to various stakeholders. LLMs can significantly enhance this process by aiding in the drafting of reports, summarizing complex information, and tailoring messages to different audiences (see Table 10). Their role is to refine the presentation of assessment results, ensuring clarity, consistency, and impact, while always operating under strict human review and with due regard for information sensitivity.

Table 10. Key Applications of LLMs in Reporting & Follow-up

Functional Area	Description of LLM Contribution
Drafting Executive Summaries	Condensing comprehensive assessment findings into high-level overviews suitable for senior management and non-technical leadership.
Reformatting Technical Findings	Translating complex technical details and security jargon into plain language explanations suitable for diverse audiences.
Grouping Issues Thematically	Organizing observations, findings, and recommendations into logical themes (e.g., governance, technology, human factors) for improved report clarity and actionability.

***CAUTION**: Security assessment reports will typically include highly sensitive information relating to vulnerabilities, threats, and potential mitigations—information that may not be suitable for processing or consumption by LLMs.*

Reporting & Follow-up: Sample Prompt

When using LLMs for reporting, ensure that all source material provided has been thoroughly anonymized and stripped of any proprietary, classified, or sensitive data. LLM outputs should always be reviewed by a human expert before dissemination.

"Using this anonymized list of findings, group them into 4–6 themes and suggest a one-paragraph high-level summary for senior non-technical management."

5. OBTAINING HIGH-QUALITY RESPONSES FROM LARGE LANGUAGE MODELS

LLMs are trained on vast datasets, which means that the quality of their responses can be highly sensitive to the specificity and clarity of user input. To consistently elicit accurate, useful, and relevant outputs for security assessments, effective prompting techniques are essential. Poorly formulated prompts can lead to vague, irrelevant, or even incorrect information, diminishing the utility of these powerful tools.

To guide the construction of effective prompts, Dave Birss's *How to Write an Effective Prompt for AI*[2] course recommends the "CREATE" mnemonic, shown in Table 11, which offers a comprehensive framework.

Table 11. CREATE Mnemonic

ELEMENT					
C	**R**	**E**	**A**	**T**	**E**
Create	Request	Examples	Adjustments	Type of Output	Extras
DESCRIPTION					
Define the role or persona for the AI to adopt.	Clearly state what you want the AI to do.	Provide examples of the desired output or relevant context.	Indicate how to refine the prompt or response based on initial results.	Specify the exact format of the AI's response.	Include any additional instructions, constraints, or behavioral requests.

[2] How to write and effective prompt for AI – How to write an effective prompt for AI | LinkedIn Learning

ELEMENT					
C	**R**	**E**	**A**	**T**	**E**
EXAMPLE					
Include any additional instructions, constraints, or behavioral requests.	"...summarize the findings." "...list key vulnerabilities."	"Provide a response like this: [Paragraph 1: Summary. Paragraph 2: Recommendation]" "Here is a good example of a vulnerability description: [Text]"	"If the answer is too technical, simplify it." "If it misses X, elaborate on X."	"...as a bulleted list." "...in markdown table format with columns 'Risk' and 'Mitigation.'"	"Do not include any site-specific details." "Ask clarifying questions if unsure." "Explain your reasoning."

Example Prompts Using the CREATE Mnemonic:

To further illustrate the practical application of the CREATE mnemonic in generating high-quality responses for security-related tasks, consider the following examples.

Example 1: High-Level Assessment Framework

"You are assisting a security professional as they prepare a high-level assessment framework for a generic nuclear power plant. The audience is composed of experienced security assessors who are not AI experts. The goal is to outline key assessment domains and provide example questions within each domain. The output should contain no site-specific details, no operational secrets, and no instructions on how to defeat security measures. Provide the output as a table with the following columns: 'Domain,' 'Objective,' and 'Example Assessment Questions.' Focus on physical security, cybersecurity, personnel reliability, and incident response."

Example 2: Policy Gap Identification

"As a senior security analyst focused on continuous improvement and high reliability of a critical infrastructure, review the attached redacted version of our site access policy, with sensitive details removed. Your task is to summarize the key controls present in the policy and then identify any obvious gaps or ambiguities from a security perspective based on general industry best practices (do not refer to specific national regulations). Present your findings in two distinct sections: 'Policy Summary' and 'Identified Gaps and Ambiguities,' using bullet points for each."

5.1. CORE PRINCIPLES OF GOOD PROMPTING

Beyond the foundational CREATE mnemonic, adhering to specific core principles when constructing prompts is vital for maximizing the quality and utility of LLM responses, particularly in sensitive security contexts (see Table 12). These principles ensure clarity, focus, and adherence to security requirements.

Table 12. Core Principles for Effective Prompt Construction

	Principle	Description	Why It Is Important for Security Assessments
1	Be specific about your goal.	State precisely what you want the LLM to achieve, defining the scope (e.g., system boundaries, assumptions, exclusions), intended audience, and desired success criteria.	Ensures the LLM's output is directly relevant and actionable, preventing vague advice that cannot be applied to critical security decisions.
2	Provide context (within security limits).	Furnish just enough relevant system context, architecture overview, asset roles, and threat model assumptions. Use sanitized placeholders and label hypothetical aspects clearly. **Crucially, avoid sharing sensitive details (e.g., live IP addresses, credentials, proprietary diagrams, site-specific vulnerabilities).**	Allows the LLM to tailor recommendations to cyber-physical realities without exposing Confidential, Proprietary, or Classified Information. This is fundamental to safeguarding critical infrastructure.
3	Ask for structure.	Request a specific output format, such as step-by-step procedures, tables, checklists, or sections aligned with recognized frameworks (e.g., MITRE ATT&CK/Industrial Control Systems, National Institute of Standards and Technology).	Encourages the LLM to organize information logically, making outputs easier to verify; integrate into workflows; and validate against security standards and local requirements.
4	Iterate and refine.	Treat the initial LLM response as a draft. Be prepared to ask follow-up questions to refine, clarify uncertainties, add constraints, or correct assumptions.	Improves accuracy and usefulness over successive turns, allowing human experts to guide the LLM toward precise and reliable outcomes, mitigating risks of "hallucination" and irrelevance.
5	Validate all outputs.	Independently verify every piece of information, claim, or recommendation generated by the LLM using authoritative sources, subject matter expertise, and real-world data.	**Crucial for mitigating "hallucinations" and ensuring accuracy.** Prevents the incorporation of incorrect or misleading information into critical security assessments, protecting against misinformed decisions that could compromise safety or security.

Applying these principles effectively transforms LLMs into more reliable and precise tools. The stark difference between a poorly conceived prompt and a well-structured one is best illustrated through direct comparison, as demonstrated in the following section.

5.2. EXAMPLES: WEAK VERSUS STRONG PROMPTS

Applying these principles effectively transforms LLMs into more reliable and precise tools. The stark difference between a poorly conceived prompt and a well-structured one is best illustrated through direct comparison, as demonstrated in the following section.

Example 1: Generating an Assessment Framework

Weak Prompt: "Give me a nuclear plant security assessment."

Problem: This prompt contains no defined *Character*, *Request* is vague, no *Examples*, no *Adjustments* guidance, no *Type of Output* specified, and no *Extras* for constraints or security. Without this information, the response is likely to be vague, generic, and potentially include inappropriate or insecure details.

Strong Prompt: "You are assisting a security professional as they prepare a high-level assessment framework for a generic nuclear power plant. The audience is composed of experienced security assessors who are not AI experts. The goal is to outline key assessment domains and provide example questions within each domain. The output should contain no site-specific details, no operational secrets, and no instructions on how to defeat security measures. Provide the output as a table with the following columns: 'Domain,' 'Objective,' and 'Example Assessment Questions.' Focus on physical security, cybersecurity, personnel reliability, and incident response."

Analysis of Strong Prompt:

What makes this prompt strong is the application of the CREATE Mnemonic, along with the inclusion of core principles outlined in Section 5.1:

CREATE Mnemonic Applied:

- *Character:* "You are assisting a security professional ..."
- *Request:* "... outline key assessment domains and provide example questions within each domain."
- *Examples:* Implied by the clear context and structure request.
- *Adjustments:* Not explicitly stated, but the clarity of the request makes initial adjustments less likely to be needed.
- *Type of Output:* "... as a table with the following columns: 'Domain,' 'Objective,' and 'Example Assessment Questions.'"
- *Extras:* "The audience is experienced security assessors who are not AI experts." "Goal is to outline ..." "No site-specific details, no operational secrets, no instructions on how to defeat security measures." "Focus on physical security, cybersecurity, personnel reliability, and incident response."

Core Principles (Section 5.1) Applied:

- *Be Specific About Your Goal:* Clearly states the goal (outline domains/questions) and audience.
- *Provide Context (Within Security Limits):* Specifies "generic nuclear power plant" and explicitly excludes "site-specific details, operational secrets, instructions on how to defeat security measures," directly addressing the critical need to protect sensitive information.
- *Ask for Structure:* Requests a very specific table format with defined columns.
- *Iterate:* While not explicitly mentioned in the prompt, the clarity provided by this prompt drastically reduces the need for extensive iteration, making the first response far more useful.

- *Validate All Outputs:* The prompt's constraints ensure the output is suitable for expert validation, reducing the likelihood of critical errors.

Example 2: Analyzing a Policy Document for Gaps

Weak Prompt: "Review this access policy for problems."

Problem: This prompt lacks a *Character*, the *Request* is extremely vague ("problems"), no *Examples* are given, no *Adjustments* guidance, no *Type of Output* is specified, and crucial *Extras* regarding security and context are missing. It assumes the LLM understands "problems" from a security perspective without guidance.

Strong Prompt: "As a senior security analyst focused on continuous improvement and high reliability of a critical infrastructure, review the attached redacted version of our site access policy, with sensitive details removed. Your task is to summarize the key controls present in the policy and then identify any obvious gaps or ambiguities from a security perspective based on general industry best practices (do not refer to specific national regulations). Present your findings in two distinct sections: 'Policy Summary' and 'Identified Gaps and Ambiguities,' using bullet points for each."

Analysis of Strong Prompt:

CREATE Mnemonic Applied:

- *Character:* "As a senior security analyst focused on continuous improvement and high reliability of a critical infrastructure ..."
- *Request:* "... summarize the key controls ... and then identify any obvious gaps or ambiguities from a security perspective ..."
- *Examples:* Implied by the request for a summary and gap identification from a policy document (which would be provided as an attachment).
- *Adjustments:* Not directly stated, but the structured request makes the output easier to review and refine.
- *Type of Output:* "... in two distinct sections: 'Policy Summary' and 'Identified Gaps and Ambiguities,' using bullet points for each."
- *Extras:* "... redacted version ... with sensitive details removed." "Based on general industry best practices (do not refer to specific national regulations)."

Core Principles (Section 5.1) Applied:

- *Be Specific About Your Goal:* Clearly defines both a summary task and a gap identification task, specifying the perspective ("security perspective based on general industry best practices").
- *Provide Context (Within Security Limits):* Explicitly states "redacted version" and "sensitive details removed," and "do not refer to specific national regulations" to avoid sensitive or inaccurate domain knowledge. This is critical for data protection.
- *Ask for Structure:* Requests presentation in "two distinct sections" with "bullet points," ensuring organized output.

- *Iterate:* A structured output makes it easier to iterate on specific findings or request elaboration on certain gaps.
- *Validate All Outputs:* The request for "obvious gaps or ambiguities" implies human expert review to confirm the LLM's findings against actual policy and operational reality.

6. VERIFYING AND USING LARGE LANGUAGE MODEL OUTPUTS SAFELY

Building on the principle of "Validate All Outputs", this section delves deeper into the critical practices required to ensure the reliability and safety of LLM-generated content in nuclear security assessments. While LLMs are powerful tools, their outputs are not always reliable. Therefore, rigorous verification and mindful application are nonnegotiable.

6.1. ALWAYS APPLY EXPERT REVIEW

All outputs from LLMs, regardless of the prompt quality, must be treated as preliminary drafts or suggestions, not as definitive truths. This necessitates comprehensive review by qualified human subject matter experts.

- **Fact Checking:** Meticulously verify all technical claims, references to standards, regulatory requirements, and any quantitative data (e.g., probabilities, distances, timings) generated by the LLM.
- **Contextual Validation:** Have subject matter experts review all key outputs, especially those related to nuclear safety, security, or emergency response functions, against real-world conditions, site-specific knowledge, and operational realities.
- **Never Blindly Accept:** A critical security conclusion or operational recommendation must never be accepted solely because "the AI said so."

6.2. BEWARE OF "HALLUCINATIONS"

LLMs can confidently generate plausible-sounding but entirely false statements, known as "hallucinations." These are a significant risk because they can mislead users into taking the wrong actions.

Examples: Hallucinations in a security context might include citing non-existent regulatory requirements, inventing standards or guidance documents, making generalized claims about "all nuclear plants" that are untrue, or confidently asserting unproven security measures.

Mitigation Strategies

- **Explicitly Prompt for References:** Confirm that the LLM model in use is capable of web search to access references and include instructions like the following: "Provide references to widely recognized standards or guidance (e.g., National Institute of Standards and Technology, International Organization for Standardization, IAEA). If uncertain, state so."
- **Prompt for Uncertainty:** Ask the LLM to identify potential areas of uncertainty by requesting facts to verify: "List any parts of your answer that may be uncertain, outdated, or differ between jurisdictions and cite references accordingly."

- **Cross-Verification:** Always cross-verify LLM outputs with established, authoritative, human-curated databases; official documents; and expert knowledge.

7. PROTECTING SENSITIVE DATA WHEN USING LARGE LANGUAGE MODELS

As emphasized by the Protect Sensitive Information principle, the integrity and confidentiality of sensitive information are paramount in nuclear security. While LLMs offer powerful capabilities, their use introduces inherent risks of data exposure if not managed with extreme diligence. Sensitive information (including facility details, operational data, and personnel information) must be rigorously protected. Therefore, no sensitive information should ever be transmitted to an LLM without explicit organizational policy authorization. If you are using an LLM for sensitive information, the LLM system itself must be demonstrably secured and controlled to meet all relevant security, regulatory, and classification requirements.

To actively protect sensitive data when utilizing LLMs, organizations must implement comprehensive technical controls and establish clear, enforceable guidelines for personnel.

7.1. CONFIGURE TOOLS CORRECTLY

Implementing robust technical controls is the first line of defense for protecting sensitive data processed by LLMs. This requires close coordination between IT, security, legal, and compliance teams. Table 13 outlines key configuration requirements:

Table 13. Technical Controls for Secure LLM Deployment

Control Area	Key Actions and Considerations
Data Usage and Retention	Confirm that input data is *not* used for model training.
	Ensure the ability to opt out of logging where regulations or policy require.
	Understand data residency (where data is stored) and retention periods per legal and organizational requirements.
Access Management	Implement single sign-on and robust access control mechanisms.
	Ensure only authorized staff with appropriate roles can access and use approved LLM tools.
Feature Control	Disable or restrict features that automatically send data to third parties (e.g., plugins, external connectors) unless explicitly vetted and approved.
Policy Integration	Incorporate LLM use into existing information security policies and acceptable use guidelines.
Coordination	Maintain close coordination with Chief Information Security Officer/IT Security, Legal, and Data Protection Officers (if applicable) to ensure compliance and risk mitigation.

Note: The specific technical controls and security requirements for an LLM deployment will vary significantly based on the chosen deployment model. Refer to Section 3.3.2 for a detailed overview of the different models and their inherent security considerations.

7.2. ESTABLISH GUIDELINES FOR STAFF

Even with robust technical controls, human error remains a significant risk. Clear, concise, and mandatory guidelines for staff are essential to ensure the safe and responsible use of LLMs (see Table 14).

Table 14. Key Information to Convey to Staff

Guideline Area	Key Actions and Considerations
Prohibited Information	Clearly define and communicate what types of information (e.g., classified, site-specific operational details, unredacted personal data) must *never* be entered into any LLM, regardless of deployment type.
Approved Platforms	Specify which LLM platforms or tools are officially approved for use and for what purposes, and explicitly list any disallowed tools (e.g., free public cloud LLMs for sensitive tasks).
Redaction Protocols	Provide detailed instructions on how documents or data must be thoroughly redacted or anonymized before being used with any LLM. *Note: Recognize that LLMs can sometimes infer connections from seemingly disparate data, making complete anonymization challenging. Human review remains paramount.*
Authorization for Use Cases	Define the process and criteria for authorizing new LLM use cases or integrations and identify who has the authority to grant such approvals.
Incident Response	Establish a clear process for reporting if sensitive information is accidentally, or intentionally, shared with an LLM, including steps for immediate containment and remediation.

Ensure appropriate use guidelines are an integral part of ongoing security awareness and AI training for all relevant personnel.

8. SAFETY AND ABUSE CONCERNS SPECIFIC TO NUCLEAR & CRITICAL INFRASTRUCTURE

The application of LLMs within the nuclear security and critical infrastructure sectors demands an exceptionally high degree of caution and ethical consideration. Given that these targets are of high value and potential compromise could lead to catastrophic consequences (including national security threats, environmental damage, and loss of life), the misuse or accidental revelation of information through LLMs carries unparalleled risks compared to other domains. This section outlines critical safety and abuse concerns unique to this sensitive environment.

8.1. AVOID GENERATING HARMFUL CONTENT

LLMs can process and generate information on a vast array of topics. In the context of nuclear and critical infrastructure security, it is imperative to prevent their use in ways that could facilitate

malevolent actors (see Table 15). Deliberately or inadvertently prompting an LLM to generate content that could be exploited for harmful purposes poses an unacceptable risk.

Table 15. Prohibited and Recommended LLM Interactions for Harmful Content

Prohibited Interaction (NEVER ask for)	Recommended Interaction (Frame question around defense & resilience)
× Step-by-step guidance on how to defeat specific security systems.	✓ "Describe general defensive measures for protecting control systems in power plants from cyberattacks."
× Methods to exploit specific vulnerabilities (either generic or real).	✓ "List indicators that a physical security system might be degraded or compromised."
× Instructions for sabotage, bypassing controls, or harming operations.	✓ "What are best practices for maintaining security system integrity?"
× Information that could directly contribute to the creation or deployment of weapons of mass destruction.	✓ "How can security awareness training be improved to mitigate insider threats?"

The generation of such harmful content, even if hypothetical, risks contributing to a body of knowledge that could be misused, potentially facilitating malicious activities against critical infrastructure.

8.2. BE ALERT TO DUAL-USE RISKS

Certain information or insights derived from LLMs can be inherently "dual-use"—meaning they possess the potential for both beneficial (defense, assessment) and malevolent (offense, exploitation) applications. Understanding and managing this dual-use nature is paramount in preventing unintended security compromises (see Table 16).

Table 16. Managing Dual-Use Risks in LLM Interactions

Risk Management Strategy	Description and Why It Is Important
Keep Discussions High-Level and Conceptual	Focus prompts on abstract principles, general methodologies, and widely recognized threats rather than specific details. This reduces the chance that an LLM's output could inadvertently form part of an exploitable knowledge chain.
Do Not Connect to Specific Weaknesses	Never link LLM queries to a specific facility's known vulnerabilities, incident reports, or audit findings. This prevents the LLM from synthesizing or drawing connections that could reveal exploitable points in your infrastructure.
Avoid Combining Detailed Pieces	Be cautious about feeding multiple, seemingly innocuous, detailed pieces of information into an LLM during a single conversation or over iterative prompts, especially if those pieces, when combined, could logically reveal a viable attack path against a system.

"Attacker Mindset" Check	Before submitting a prompt, consider the following: "Does this question 'feel' like something an attacker might ask to gain advantage or information?" If the answer is yes, it is likely not appropriate to pose directly to an LLM, particularly if it contains any context about your own facility or specific systems. Opt for defensive framing instead.

By adhering to these principles, assessors can leverage LLMs for beneficial analysis while vigilantly guarding against the inadvertent creation or dissemination of information that could be weaponized.

9. COMMON MISTAKES USERS MAKE (AND HOW TO AVOID THEM)

Even with clearly defined policies and robust systems, users can inadvertently make mistakes that compromise security or diminish the utility of LLMs. This section highlights prevalent errors and provides actionable strategies for prevention, building upon the principles and guidance established throughout this document. Adherence to these recommendations is critical to mitigate risks and maximize the benefits of LLM integration into security assessments.

9.1. MISTAKE 1: PASTING SENSITIVE DOCUMENTS DIRECTLY INTO A PUBLIC LARGE LANGUAGE MODEL

This mistake involves uploading unredacted site security plans, operational details, personnel data, or any other sensitive/classified material directly into a public, consumer-grade LLM interface (see Table 17).

Table 17. Mistake #1: Pasting Sensitive Documents Directly into Public LLMs

Problems and Risks Introduced	Prevention Strategies
× **Data Exposure and Loss of Control** Inputs are highly likely to be stored, logged, and potentially utilized by the LLM provider for model training, leading to unauthorized disclosure of critical information.	✓ **Approved Platforms Only** Strictly adhere to using only LLM platforms explicitly approved and configured (*refer to Section 3.3.2 on deployment models and Section 7.1 on technical controls*).
× **Legal/Policy Breach** Violates national laws, regulatory requirements (e.g., IAEA safeguards), and internal information security policies, leading to severe legal, financial, and professional consequences.	✓ **Rigorous Redaction** Thoroughly redact or anonymize all documents or data before any interaction with an LLM, even approved ones *(refer to Section 3.3.1 on sensitive data and Section 7.2 on staff guidelines*). *Note: Staff should ensure that redacted materials are vetted for LLM use.*
× **Irreversible Compromise** Once sensitive data enters unapproved external systems, it is practically impossible to regain control, with potentially catastrophic implications for security and non-proliferation.	✓ **"Assume Untrusted" Rule** When in doubt about a platform's security posture or an information's sensitivity, assume any external or unapproved tool is unsuitable for sensitive content.

9.2. MISTAKE 2: TREATING AI OUTPUT AS AUTHORITATIVE FACT

This mistake involves accepting LLM-generated content—such as security conclusions, vulnerability assessments, recommended mitigations, or factual statements—as definitive and accurate without independent human verification (see Table 18).

Table 18. Mistake #2: Treating AI Output as Authoritative Fact

Problems and Risks Introduced	Prevention Strategies
× **False Confidence & Misinformed Decisions** LLMs can present plausible-sounding but incorrect, outdated, or contextually irrelevant information with high confidence. Blindly trusting these outputs can lead to critical security decisions that compromise safety, security, or regulatory compliance.	✓ **AI as Assistant, Not Authority** Reinforce the principle that LLMs are assistive tools, not decision-makers or validators (*refer to Section 3.2.3 and the strong caution statement in Section 3.3.3*).
× **Lack of Real-World Context** LLMs do not possess facility-specific, regulatory, or operational understanding, making their unverified conclusions unreliable.	✓ **Mandatory Expert Review** Enforce human subject matter expert review for all substantive LLM outputs (*refer to Section 6.1 and the "Validate All Outputs" principle in Section 5.1*).
× **Hallucinations** Risk of incorporating invented facts or non-existent requirements into critical assessments, which can be particularly dangerous in a nuclear security context (*refer to Section 6.2*).	✓ **Cross-Verification** Always cross-verify LLM outputs against known standards, local regulations, site-specific reality, and authoritative sources before acting on any information.

9.3. MISTAKE 3: ASKING THE WRONG QUESTION (INSUFFICIENT PROMPT ENGINEERING)

This mistake occurs when prompts are vague, unconstrained, or incomplete, failing to specify clear goals, sufficient context (within security limits), desired output structure, or essential security constraints.

Table 19. Mistake #3: Asking the Wrong Question

Problems and Risks Introduced	Prevention Strategies
× **Vague/Generic Advice** Unclear prompts yield unactionable or irrelevant responses, wasting time and resources.	✓ **Apply Prompt Engineering Principles** Rigorously follow the CREATE mnemonic (*Section 5*) and the "Core Principles of Good Prompting" (*Section 5.1*) for every interaction to ensure specificity, context, and structure.
× **Unrealistic or Insecure Assumptions** The model may make incorrect assumptions about the environment or task, potentially leading to irrelevant, misleading, or even unsafe outputs if not properly contained.	✓ **Define Constraints Explicitly** Always clearly state what information should be excluded (e.g., "no site-specific details," "no classified data," "avoid prescriptive advice") to align with

	Section 7.2 on staff guidelines and Section 8 on safety concerns.
× **Inefficient Iteration** Poor initial prompts necessitate extensive iterative refinement, which is time-consuming and can be frustrating.	✓ **Provide Clear, Specific Context** Focus on crafting prompts that are clear to increase the likelihood of obtaining a useful response.

9.4. MISTAKE 4: OVERLOADING THE TOOL WITH TOO MANY TASKS AT ONCE

This mistake involves attempting to perform multiple, complex, distinct analyses or generation tasks within a single, lengthy prompt (see Table 20).

Table 20. Mistake #4: Overloading AI with Too Many Tasks

Problems and Risks Introduced	Prevention Strategies
× **Inconsistent or Shallow Output** LLMs may struggle to maintain focus and context across disparate tasks in a single turn, leading to fragmented, incomplete, or superficial results for each sub-task.	✓ **Break Down Complex Tasks** Decompose large, multi-faceted requests into smaller, sequential, and manageable steps (e.g., first, summarize the document; second, analyze the summary for gaps; third, draft improvements based on identified gaps).
× **Difficulty in Verification** It becomes challenging to isolate and verify the quality and accuracy of each individual component of the LLM's response, making the overall output less trustworthy for critical applications.	✓ **Iterate Sequentially** Use the output from one LLM interaction as a refined input for the next step, allowing for focused refinement and validation at each stage, improving overall accuracy and control.
× **Reduced Control and Guidance** Combining too many steps in one prompt diminishes the user's ability to guide the LLM effectively and iterate on specific aspects of the task.	✓ **Use AI Agents** Build and use multiple AI agents for repetitive and simple tasks to automate workflows and reduce tedious labor-intensive work.

9.5. MISTAKE 5: NOT PRESERVING AN AUDIT TRAIL

This mistake involves generating important assessment content (e.g., policy drafts, risk analyses, finding summaries) using an LLM without maintaining a comprehensive record of the original prompt, the specific LLM model and version used, and the subsequent human edits, reviews, and approvals (see Table 21).

Table 21. Mistake #5: Not Preserving an Audit Trail

Problems and Risks Introduced	Prevention Strategies
× **Lack of Accountability & Traceability** In highly regulated environments like nuclear security, the inability to reconstruct the origin, evolution, and human intervention points for critical	✓ **Mandatory Logging** Implement a strict policy to store all significant prompts, LLM outputs, and the LLM model/version used (if feasible) within the organization's document management system.

Problems and Risks Introduced	Prevention Strategies
documents is a significant compliance and accountability failure.	
× **Regulatory & Audit Issues** Failure to satisfy regulatory requirements for demonstrating due diligence and information provenance can severely impact the defensibility of assessments or policies.	✓ **Robust Version Control** Ensure that final documents retain a clear record of any AI-generated components, along with all subsequent human revisions, reviews, and approval chains.
× **Impeded Learning & Process Improvement** Lack of historical records prevents effective review of LLM performance, identification of best prompting practices, or understanding how LLM outputs were integrated into final work products.	✓ **Document Operational Procedures** Maintain logs of who used the LLM tool, for what specific purpose, and during which phase of the assessment, adhering to internal privacy rules and comprehensive operational security protocols.

The instances highlighted above underscore a fundamental truth: the benefits of LLM integration into security assessment workflows are directly proportional to the rigor with which their limitations and potential for misuse are managed.

Avoiding these common mistakes requires not only an understanding of technical controls and policies but also a continuous cultivation of a security-first mindset among all users. Proactive adherence to these prevention strategies is not merely good practice; it is an essential safeguard against compromising the very security posture LLMs are intended to enhance.

10. PRACTICAL CHECKLIST FOR SETTING UP LARGE LANGUAGE MODEL USE IN NUCLEAR SECURITY ASSESSMENTS

Implementing LLMs in nuclear security assessments requires a structured and comprehensive approach to ensure both effectiveness and robust security. The checklist in Table 22 synthesizes the key considerations discussed throughout this guide, providing actionable steps for organizations to establish a secure and compliant LLM integration program. It is designed to assist security practitioners in systematically managing the associated risks and maximizing the benefits.

Table 22. LLM Implementation Checklist for Critical Infrastructure Security Assessments

	Checklist Item (Action Required)	Description/Key Considerations (Reference Section)
	CRITICAL: Establish Overarching Data Sensitivity Protocols	
☐	Establish Data Sensitivity Protocols	**Before working with any LLM, define security protocols to classify, handle, redact, and securely process any data to be handled with LLMs.** This foundational step underpins all subsequent security measures and ensures compliance with Section 3.3, Section 7, and Section 8.

	Checklist Item (Action Required)	Description/Key Considerations (Reference Section)
	Organizational Setup & Policy	
☐	Define Approved Use Cases	Clearly identify and document the specific security assessment tasks authorized for LLM use (e.g., generic framework drafting, non-sensitive document summarization, high-level risk brainstorming). (See Section 4.)
☐	Identify Approved LLM Platforms	Explicitly list all authorized LLM platforms and specify for which types of information they can be used. Prohibit public LLMs for sensitive data. (See Section 3.3.2, 7.1.)
☐	Define Prohibited Information	Clearly communicate all types of information (e.g., classified, site-specific operational details, unredacted personal data) that must *never* be entered into *any* LLM. (See Section 3.3.1, 7.2.)
☐	Update Information Security Policies	Integrate LLM usage policies, acceptable use guidelines, and data handling requirements into the organization's existing information security framework. (See Section 7.1.)
☐	Establish Roles & Responsibilities	Clearly assign roles for LLM governance, including policy enforcement, technical oversight, user training, and incident response.
☐	Obtain Legal & Compliance Approval	Ensure all LLM integration plans are reviewed and approved by legal counsel and compliance officers to meet national regulations, international obligations, and internal policies. (See Section 7.1.)
	Technical Controls & Infrastructure	
☐	Select Appropriate Deployment Model	Choose an LLM environment (on-premise, private cloud) commensurate with the highest sensitivity level of information to be processed. Avoid public cloud for sensitive data. (See Section 3.3.2.)
☐	Implement Strong Access Control	Configure single sign-on and role-based access control to ensure only authorized personnel can access and use approved LLM tools.
☐	Data Usage & Retention Configuration	Confirm input data is *not* used for model training. Configure logging and retention policies to align with organizational and regulatory requirements (e.g., disable logging where required). (See Section 7.1.)
☐	Disable Risky Features	Turn off or restrict LLM features that could automatically send data to third parties (e.g., plugins, external connectors) unless explicitly vetted and approved. (See Section 7.1.)
☐	Secure Data Flow	Ensure secure, encrypted channels for all data transmission to/from the LLM and that data at rest within the LLM environment is encrypted.
☐	Implement Audit & Monitoring	Establish comprehensive audit logging for all LLM interactions (user, time stamp, sanitized prompt, response).

	Checklist Item (Action Required)	Description/Key Considerations (Reference Section)
		Implement monitoring for anomalous activity. (See Section 9, Mistake 5.)
	Data Management & Sensitivity (Classification)	
☐	Conduct Data Sensitivity Assessment	Formally assess the sensitivity and classification levels of all data types intended for use with LLMs. (Ref. Section 3.3.1)
☐	Develop Redaction & Anonymization Protocols	Create and enforce clear, practical protocols for thoroughly redacting or anonymizing sensitive information before LLM input. (See Section 7.2.)
☐	Implement Sanitization Verification	Establish processes to verify data have been correctly sanitized (e.g., second-party review) before LLM submission.
☐	Define Ingress/Egress Controls	Implement controls for how data enters and leaves the LLM environment, preventing unauthorized transfer of sensitive information.
	Human Factors & Training	
☐	Provide Mandatory AI Awareness Training	Cover LLM capabilities/limitations (see Section 3.2.1), human validation/expert review (see Sections 5.1, 6.1), nuclear-specific concerns (see Section 8), and the “assistant, not authority” principle (see Section 3.3.3).
☐	Provide Prompt Engineering Training	Offer practical training on effective prompt construction using the CREATE mnemonic and core principles (Sec. 5), including hands-on exercises.
☐	Provide Data Handling & Redaction Training	Train staff specifically on the organization’s data classification, handling, and redaction protocols for LLM use. (See Section 7.2.)
☐	Train on Incident Reporting Procedures	Instruct staff on the process for reporting accidental disclosure of sensitive information via LLMs, including immediate containment and reporting steps. (See Section 7.2.)
	LLM Use During Assessments	
☐	Define Planning Phase Protocols	Use LLMs for generic templates, checklists, and training materials, strictly avoiding site-specific sensitive content. (See Section 4.1.)
☐	Establish Document Review Protocol	Redact all sensitive documents, use LLMs for summaries, gap analysis, simplification—never for decisions. (See Section 4.2.)
☐	Guide Risk & Scenario Brainstorming	Define how LLMs can be used for high-level threat scenario brainstorming, ensuring no exploitable details are exposed. (See Section 4.3.)
☐	Outline On-Site Support Procedures	If used on-site, ensure LLMs are in offline/secure mode and process *sanitized* notes into structured findings only. (See Section 4.4.)

	Checklist Item (Action Required)	Description/Key Considerations (Reference Section)
☐	Implement Reporting & Follow-up Protocols	Define LLM usage for drafting phrasing, formatting, and non-technical summaries, always ensuring human verification of claims. (See Section 4.5.)
☐	Maintain Audit Trail	Establish and enforce procedures for preserving an audit trail of all significant LLM interactions, prompts, outputs, and human review/approval. (See Section 9, Mistake 5.)

This comprehensive checklist underscores that the successful and secure integration of LLMs into critical infrastructure security assessments is not a singular event but an ongoing, multi-faceted commitment. By systematically addressing each of these areas—from foundational organizational policies and robust technical controls to meticulous data management, continuous training, and disciplined operational protocols—organizations can harness the efficiency and analytical power of LLMs while rigorously safeguarding against their inherent risks. Ultimately, this structured approach ensures that LLMs serve as powerful force multipliers for human expertise, enhancing the overall security posture without compromising sensitive assets or operational integrity.

11. IMPROVEMENTS & ADDITIONAL RESEARCH

The testing and analysis presented in this guide have demonstrated the significant potential of LLMs to enhance critical infrastructure security assessments. However, they have also highlighted inherent limitations and underscored the paramount importance of robust security protocols and ongoing human oversight. To further advance the utility and safety of LLMs in this sensitive domain, several key areas for improvement and additional research are identified.

11.1. ENHANCING LARGE LANGUAGE MODEL CAPABILITIES FOR NUCLEAR SECURITY

Improved Multimodality and Information Integration: While advanced LLMs show promise, further research is needed to seamlessly integrate and reason across all forms of assessment data: text, code, images, and specialized binary files (e.g., PCAPs, system logs). This includes developing LLMs that can directly parse and interpret such proprietary or binary formats, reducing the reliance on external tools and improving correlation capabilities.

Contextual Understanding and Domain Specialization: Research should focus on developing LLMs with deeper contextual understanding of nuclear security, rather than just pattern recognition. This includes

- **"Nuclear Security-Aware" LLMs**: Training or fine-tuning models on curated, non-sensitive nuclear security frameworks, regulations, and best practices to improve domain-specific accuracy and relevance.
- **Reasoning and Causality:** Advancing LLM capabilities to better understand causal relationships and interdependencies within complex systems, enabling more reliable risk analysis and scenario prediction.

Reduced Hallucination and Increased Verifiability: Continued research into techniques that reduce LLM hallucination rates and increase the verifiability of their outputs is critical for nuclear security assessments. This includes developing models that can:

- **Cite Sources More Reliably:** Provide more precise and verifiable citations for generated information.
- **Express Uncertainty:** Explicitly articulate confidence levels or areas of uncertainty in their responses.

11.2. ADVANCING SECURE AND AUTONOMOUS LARGE LANGUAGE MODEL APPLICATIONS

Agentic AI for Assessment Workflows: Investing in agentic AI development holds significant potential. This involves prototyping and testing autonomous LLM agents that can plan and execute assessment workflows (e.g., "collect → interpret → hypothesize → validate → report"). Such agents could utilize various tools (e.g., search engines, parsers, vulnerability databases, diagram extraction) to increase the periodicity and depth of assessments, potentially enabling near real-time monitoring and alerting for changes in the threat landscape. This would move beyond sporadic human-led assessments.

Standalone/Self-Hosted LLMs and Air-Gapped Environments: The current evaluation primarily involved cloud-hosted LLMs configured for controlled unclassified information. Extensive follow-on research is essential to rigorously test the effectiveness, security, and performance of various standalone and self-hostable LLM architectures in truly air-gapped or highly isolated environments. This includes the following:

- **Performance Optimization:** Optimizing smaller, open-source, or commercial LLMs for performance on restricted hardware typical of secure environments.
- **Security Hardening:** Developing specialized security hardening techniques and configurations for on-premise LLM deployments to meet the most stringent classified and sensitive data handling requirements.
- **Maintaining Relevance:** Research into efficient and secure methods for updating and maintaining the knowledge base of air-gapped LLMs without compromising their isolation.

11.3. DEVELOPING ROBUST EVALUATION AND ASSURANCE METHODOLOGIES

Standardized Benchmarking for Security Assessments: Developing standardized benchmarks and metrics specifically tailored to evaluate LLM performance in security assessment tasks (e.g., vulnerability identification accuracy, mitigation recommendation quality, adherence to security constraints, robustness against adversarial prompting).

Ethical AI and Misuse Prevention: Ongoing research into techniques for preventing LLMs from generating harmful "dual-use" content, especially in critical infrastructure contexts, is paramount (refer to Section 8). This includes exploring the following:

- **Proactive Guardrails:** Developing more sophisticated, inherent guardrails within LLMs themselves.
- **Adversarial Training:** Training LLMs to better resist prompts designed to elicit dangerous or exploitable information.

Human-AI Teaming Optimization: Further research into optimizing the human-AI teaming paradigm, focusing on how LLMs can best augment human expertise, facilitate decision-making, and improve assessors' cognitive load and efficiency without fostering overreliance or complacency.

These research directions aim to mature LLM technology and its application and ultimately strengthen the resilience and security of critical and nuclear infrastructure.

12. CONCLUSIONS

The advent and rapid evolution of LLMs present a transformative opportunity to enhance the efficiency, clarity, and consistency of security assessments in the nuclear and critical infrastructure domains. As this guide has demonstrated, LLMs, when deployed and utilized correctly, can serve as powerful force multipliers, streamlining tasks from document review and risk brainstorming to report generation.

However, organizations must proceed carefully and fully understand what LLMs can and cannot do. The central tenet of this guide remains unequivocal: LLMs are assistants, not authorities. Critical security judgments, validation of information, and ultimate decision-making responsibility must always reside with qualified human experts. The unique and catastrophic consequences of a security lapse in nuclear and critical infrastructure environments necessitate an absolute commitment to data protection, rigorous verification of all AI-generated outputs, and stringent ethical considerations.

The path forward involves a continuous commitment to developing and implementing robust technical controls, fostering a culture of vigilant oversight, and providing comprehensive training to all personnel. By adhering to meticulous data sensitivity protocols, carefully selecting deployment models, mastering prompt engineering, and diligently avoiding common pitfalls, organizations can strategically leverage LLMs to bolster their security posture. This forward-looking approach supports the integration of AI to strengthen the resilience and safety of our most vital assets, keeping human expertise firmly at the core of every critical assessment.