

Light Water Reactor Sustainability Program

Evaluation Methodology of Trust and Trustworthiness for Large Language Models at Nuclear Power Plants

Casey Kovesdi, Brian Wilcken, Anna Hall, Linyu Lin, John Mapes, and Jeremy Mohon

Idaho National Laboratory

Patrick Murray

EQRP, Inc.

Yusuke Yamani and Austin Jackson

Old Dominion University



August 2025

U.S. Department of Energy

Office of Nuclear Energy

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

Evaluation Methodology of Trust and Trustworthiness for Large Language Models at Nuclear Power Plants

**Casey Kovesdi, Brian Wilcken, Anna Hall, Linyu Lin, John Mapes, and Jeremy
Mohon
Idaho National Laboratory**

**Patrick Murray
EQRPI, Inc.**

**Yusuke Yamani and Austin Jackson
Old Dominion University**

August 2025

**Idaho National Laboratory
Light Water Reactor Sustainability
Idaho Falls, Idaho 83415**

<http://lwrs.inl.gov>

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

Page intentionally left blank

EXECUTIVE SUMMARY

The nuclear industry is under increasing pressure to modernize and reduce costs to remain economically competitive with other energy sectors. Large language models (LLMs), a form of generative AI, offer promising capabilities to automate and enhance tasks; however, broad deployment in nuclear power plants is hindered by critical challenges related to trust and trustworthiness. These include risks such as hallucinations, harmful bias, data privacy breaches, and lack of transparency—issues that can lead to inappropriate user trust, misuse of unreliable LLMs, or disuse of reliable LLMs. Without a structured methodology that evaluates and ensures both the trustworthiness of LLMs, and user trust are adequately addressed, the adoption of these technologies in the nuclear industry remains challenged.

To address this challenge, this research developed a comprehensive, multidisciplinary methodology that improves trustworthiness of LLMs and ensures appropriate user trust when used in the context of nuclear power plant work. The methodology brings forward guidance that addresses both technical and sociotechnical considerations, and is structured across four lifecycle phases:

1. **Work and Data Scoping** – Identifying use cases, system risks, and data integrity.
2. **Model Selection and Data Preparation** – Choosing appropriate infrastructure and models, and cleansing data.
3. **Model Optimization** – Fine-tuning models and implementing Retrieval-Augmented Generation (RAG).
4. **Model Integration** – Engineering prompts and evaluating human-AI interaction.

Each phase includes relevant technical and sociotechnical activities designed to mitigate cross-cutting risks (e.g., data privacy, hallucinations, harmful bias) and ensure calibrated user trust. The methodology is vendor- and model-neutral and includes practical examples to highlight how the methodology can be effectively implemented.

The application of this methodology produces several notable outcomes:

- **Risk Mitigation:** The methodology effectively identifies and addresses risks related to data privacy, hallucinations, harmful bias, information integrity, and information safety and security – thereby improving adoption and use.
- **Improved Trustworthiness:** Through a structured approach offered by the methodology, LLM outputs can become more valid, reliable, safe, secure, usable, useful, and contextually relevant.
- **Improved User Trust:** Through human-centered approaches, including formative and summative assessments, users will be more apt to adopt LLMs and better calibrate their trust when using LLMs for nuclear power plant work.
- **Operational Benefits:** By addressing trustworthiness and trust, the methodology supports accelerated deployment of LLMs, enabling cost

savings and improved decision-making across nuclear power plant work processes.

Collectively, this work provides the nuclear industry with a comprehensive framework for deploying trustworthy LLMs. It bridges the gap between AI capabilities and human-centered design, paving the way for safe, efficient, and confident use of LLMs in the nuclear industry.

The next step in this research is to demonstrate the methodology, presented in this technical report, to develop LLM-enabled solutions that improve nuclear power plant operations. Using the methodology and partnering with industry, future research will benchmark LLM-enabled performance against existing manual performance, comprehensively capturing the operational benefits of deploying LLMs while evaluating and addressing the associated risks with trustworthiness and trust.

ACKNOWLEDGEMENTS

This report was made possible through funding by the U.S. Department of Energy Light Water Reactor Sustainability Program. The authors would like to thank Christopher Bunting of the Department of Energy, as well as Ahmad Al Rashdan and Bruce Hallbert of Idaho National Laboratory for championing this effort.

This research made use of the resources of the High-Performance Computing Center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the U.S. Department of Energy and the Nuclear Science User Facilities under contract no. DE-AC07-05ID14517.

Page intentionally left blank

CONTENTS

EXECUTIVE SUMMARY	iii
ACKNOWLEDGEMENTS.....	v
ACRONYMS.....	xiv
1 INTRODUCTION.....	1
1.1 Scope of Research.....	2
1.2 Report Structure.....	2
2 BACKGROUND: OVERVIEW OF LARGE LANGUAGE MODELS	3
2.1 Defining LLM.....	3
2.2 How LLMs Work	3
2.3 State-of-the-Art Models	5
2.4 Functions of an LLM	5
2.5 LLM Capabilities and Opportunities for Work Efficiencies	5
3 CHARACTERIZATION OF TRUSTWORTHINESS AND USER TRUST	6
3.1 Trustworthiness: A Characteristic of the LLM.....	6
3.1.1 Background: Trustworthiness in AI.....	6
3.1.2 Factors that Influence Trustworthiness.....	9
3.1.3 Cross-Cutting Trustworthiness Issues with LLMs	12
3.2 Trust: A Characteristic of the User.....	14
3.2.1 Background: Human-Automation Trust	14
3.2.2 Factors that Influence Trust.....	16
3.2.3 Cross-Cutting Trust Issues with LLMs.....	19
4 METHODOLOGY	21
4.1 Overview.....	21
4.2 Informative Examples.....	22
5 WORK AND DATA SCOPING.....	25
5.1 Describe the Automated Process and User Tasks.....	25
5.1.1 Relevant Cross-Cutting Trust Issues	26
5.2 Identify System Risks to Automated Process	27
5.2.1 Relevant Cross-Cutting Trustworthiness Issues	27
5.3 Qualify the Integrity of Available Data	30
5.3.1 Relevant Cross-Cutting Trustworthiness Issues	32
6 MODEL SELECTION AND DATA PREPARATION.....	37
6.1 Select Hosting Infrastructure	37
6.1.1 Relevant Cross-Cutting Trustworthiness Issues	39
6.2 Select Pretrained Model.....	40
6.2.1 Relevant Cross-Cutting Trustworthiness Issues	40

6.3	Complete Data Preprocessing and Cleansing	42
6.3.1	Relevant Cross-Cutting Trustworthiness Issues	43
7	MODEL OPTIMIZATION	46
7.1	Fine-Tuning	46
7.1.1	Relevant Cross-Cutting Trustworthiness Issues	47
7.2	Retrieval-Augmented Generation	49
7.2.1	Relevant Cross-Cutting Trustworthiness Issues	50
8	MODEL INTEGRATION.....	53
8.1	Engineer the Prompt	53
8.1.1	Relevant Cross-Cutting Trustworthiness Issues	54
8.2	Evaluate Model and Human-AI Interaction.....	60
8.2.1	Relevant Cross-Cutting Trust Issues	60
8.2.2	Relevant Cross-Cutting Trustworthiness Issues	63
9	SUMMARY OF ISSUES AND REQUIREMENTS ADDRESSED.....	70
10	CONCLUSIONS.....	72
11	REFERENCES.....	73
	Appendix A Benchmark Questions and Results for LLM trustworthiness Evaluations with RAG.....	84
	Appendix B Example Formative Assessment of Trust.....	91
	Appendix C Example Summative Assessment of Trust.....	103

FIGURES

Figure 1. Illustrative conceptualization of LLM (adopted and enhanced from Department of Energy Generative Artificial Intelligence Reference Guide, 2024).	1
Figure 2. The error of several LLMs decreased during training (adapted from Touvron et al., 2023).	4
Figure 3. A basic flowchart of an LLM during the word generation process.....	4
Figure 4. Relationship between trustworthiness and user trust.....	6
Figure 5. Characteristics of trustworthy AI systems (adopted and enhanced from NIST AI 100-1 [2023]).	9
Figure 6. Characteristics of trustworthy AI systems for nuclear power plant applications.....	12
Figure 7. Cross-cutting trustworthiness issues associated with LLMs (from NIST AI 600-1 [2024] and tailored for nuclear power plant work).....	14
Figure 8. Three-layered framework for conceptualizing trust (adapted from Hoff and Bashir, 2014).....	15
Figure 9. Characteristics of trust for nuclear power plant use.	19
Figure 10. Cross-cutting trust issues associated with LLMs.	20
Figure 11. Overview of methodology: An approach to design and evaluate trustworthy AI systems and ensure appropriate user trust.....	21
Figure 12. Cross-cutting trust issues addressed when describing the work process and user tasks.....	26
Figure 13. Histogram of AI Literacy responses from a sample of psychology undergraduate majors.....	27
Figure 14. Cross-cutting trustworthiness issues addressed when identifying system risks.....	28
Figure 15. Net-HARMS applied to process of scoping work based on an incoming equipment failure.	29
Figure 16. Risk matrix as a function of hazard significance to safety or security and technical accuracy of the data.	32
Figure 17. Cross-cutting trustworthiness issues addressed when identifying system risks.....	33
Figure 18. Example condition report.	34
Figure 19. Photograph of the Bitterroot supercomputer located at Idaho National Laboratory (credit https://www.energy.gov/ne/articles/national-lab-brings-new-bitterroot-supercomputer-online).	37
Figure 20. Simplified decision chart to determine use of local hosting versus cloud hosting.....	38
Figure 21. Cross-cutting trustworthiness issues addressed when selecting hosting infrastructure.	39

Figure 22. Cross-cutting trustworthiness issues addressed when selecting a pretrained model.....	40
Figure 23. Detailed benchmarks for common autoregressive LLMs (data adopted from Hugging Face https://huggingface.co/microsoft/phi-4).	41
Figure 24. Process for preprocessing text information for fine-tuning an LLM.	42
Figure 25. Cross-cutting trustworthiness issues addressed when preprocessing the data.	43
Figure 26. High level overview of the LLM fine-tuning process.	46
Figure 27. Cross-cutting trustworthiness issues addressed when fine-tuning the model.	47
Figure 28. Cross-cutting trustworthiness issues addressed with using RAG.....	50
Figure 29. Example application of FActSCORE confidence indication on LLM response.....	51
Figure 30. Cross-cutting trustworthiness issues addressed during prompt engineering.	54
Figure 31. PoS tagging for prompt #1 (a poorly crafted prompt).	56
Figure 32. PoS tagging for prompt #2 (an improved prompt).	57
Figure 33. Cross-cutting trust issues addressed at validation.	60
Figure 34. Cross-cutting trustworthiness issues addressed at validation.	63
Figure 35. Criterion for SME review of explainability and interpretability.	64
Figure 36. Criteria for SME review of relevance.	64
Figure 37. Criterion for SME review of usefulness.....	65
Figure 38. Criterion for SME review of valid and reliable.....	66
Figure 39. Radar chart of inter-rater reliability (Cohen’s Kappa indices) to evaluate level of agreement between ChatGPT 4o to an SME. Green represents agreement for no RAG and blue represents agreement with RAG.	67
Figure 40. Criterion for SME review of fair without bias.	68
Figure 41. Example of multiple choice question, LLM response and explanation.	94
Figure 42. Example of source data the LLM deemed relevant for the Vector RAG condition.	95
Figure 43. Example of Graph RAG rationale whereby each node represents data sources used by the LLM.....	96
Figure 44. Explanation ratings across RAG conditions.....	96
Figure 45. Trust ratings across RAG conditions.....	97
Figure 46. Time in minutes to complete each block.....	98
Figure 47. Explanation ratings across LLM correctness (a) and congruency (b).	99
Figure 48. Example stimuli of color-coding/ highlight (experimental condition).	105
Figure 49. Trust and Trustworthiness Characterization for Evaluating LLMs.....	107

Figure 50. Hypothetical signal and signal + noise distributions using SDT as a framework.	108
Figure 51. Illustration of atomic facts in an LLM response.	110
Figure 52. Evaluation of a single atomic fact.	111
Figure 53. Mean sensitivity in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	112
Figure 54. Mean bias in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	113
Figure 55. Mean trust measured on Jian et al. (2000) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	113
Figure 56. Mean distrust measured on Jian et al. (2000) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	114
Figure 57. Mean performance-based trust measured on Chancey et al. (2017) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	114
Figure 58. Mean process-based trust measured on Chancey et al. (2017) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	115
Figure 59. Mean purpose-based trust measured on Chancey et al. (2017) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	115
Figure 60. Mean workload measured on NASA-TLX in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	116
Figure 61. Mean perceived risk scores in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.	116

TABLES

Table 1. Factors that are likely to influence trust in AI-infused applications at nuclear power plants.	18
Table 2. Benefits and shortcomings of different system characterization techniques.	25
Table 3. Net-HARMS risk mode taxonomy.	29
Table 4. Critical work process considerations of the data that impacts its integrity.	30
Table 5. Data source qualification example.	35
Table 7. Summary of scenarios and corresponding best RAG approaches.	52
Table 8. Common types of prompt engineering techniques.	53
Table 9. General prompt engineering principles for trustworthy LLM results.	54

Table 10. Degree of explainability as a function of different prompt engineering techniques.	55
Table 11. Comparison of readability measures.....	58
Table 12. CoT template example (adapted from Diamant, 2025).	59
Table 13. Inter-rater reliability (Cohen’s Kappa indices) to evaluate level of agreement between ChatGPT 4o to an SME.	68
Table 14. Mapping of cross-cutting issues addressed by the methodology.....	70
Table 15. Mapping of requirements addressed by the methodology.	71
Table 16. LLM question set.....	93
Table 17. Participant demographics information for each condition.....	105

Page intentionally left blank

ACRONYMS

ADAMS	Agencywide Documents Access and Management System
AI	Artificial Intelligence
API	Application Programming Interface
ASME	American Society of Mechanical Engineers
BM25	Best Matching 25
CAP	Corrective Action Program
CFR	Code of Federal Regulations
CI	Notation for ‘Confidence Interval’
CLM	Causal Language Modeling
CoT	Chain-of-Thought
CR	Correct Rejection
CWS	Circulating Water System
DOE	Department of Energy
DPR	Dense Passage Retrieval
EPRI	Electric Power Research Institute
FA	False Alarm
FActSCORE	Factual precision in Atomicity Score
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HELM	Holistic Evaluation of Language Models
HPC	High Performance Computing
Hs	Hits
IBM	International Business Machines
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
INL	Idaho National Laboratory
INPO	Institute of Nuclear Power Operations
IP	Intellectual Property
IRIS	Industry Reporting and Information System
ISO	International Organization for Standardization
LER	Licensee Event Report

LLM	Large Language Model
LoRA	Low-Rank Adaptation
MIRACLE	Machine Intelligence for Review and Analysis of Condition Logs and Entries
ML	Machine Learning
Ms	Misses
NASA-TLX	National Aeronautics and Space Administration-Task Load Index
Net-HARMS	Networked Hazard Analysis and Risk Management System
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NRC	Nuclear Regulatory Commission
O&M	Operating and Maintenance
p	Notation for ‘Probability Value’
PDF	Portable Document Format
PII	Personally Identifiable Information
PoS	Parts-of-Speech
PRA	Probabilistic Risk Assessment
PST	Perceived System Trustworthiness
R&D	Research and Development
RAG	Retrieval-Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SAP	Systems, Applications and Products
SDT	Signal Detection Theory
SME	Subject Matter Expert
SMOG	Simple Measure of Gobbledygook
STPA	Systems Theoretic Process Analysis
T	Trust
t	Notation for ‘t-distribution’
t-SNE	t-Distributed Stochastic Neighbor Embedding
TAM	Technology Acceptance Model
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TPU	Tensor Processing Unit
TRL	Technology Readiness Level
TruncatedSVD	Truncated Singular Value Decomposition
UMAP	Uniform Manifold Approximation and Projection

UTP	User Trust Potential
VIPER	Visualization for Predictive Maintenance Recommendation
XML	Extensible Markup Language

Page intentionally left blank

Evaluation Methodology of Trust and Trustworthiness for Large Language Models at Nuclear Power Plants

1 INTRODUCTION

The nuclear industry has largely operated and maintained existing plants through a labor-centric approach. While this approach has served the industry well for decades, it now carries forward economic challenges, as other electricity generating sources have modernized, leveraging machine learning (ML)/artificial intelligence (AI) and advanced automation, to significantly reduce their cost. The nuclear industry must also adopt these technologies to remain economically competitive. That is, the development, demonstration, and deployment of AI/ ML applications to perform operations, maintenance, and support functions at nuclear power plants will be essential in reducing operating and maintenance (O&M) costs while empowering plant personnel to accomplish their tasks effectively and efficiently. For instance, the use of AI/ML shows significant economic benefit across a range of areas such as equipment health monitoring (Walker et al., 2023), work management (Al Rashdan et al, 2024), and outages to name a few.

There has been extraordinary growth in the development and use of AI. Generative AI, and specifically large language models (LLMs), are one of the most recent advancements. LLMs are a specific type of AI/ML that falls under the umbrella of deep learning techniques (Figure 1).

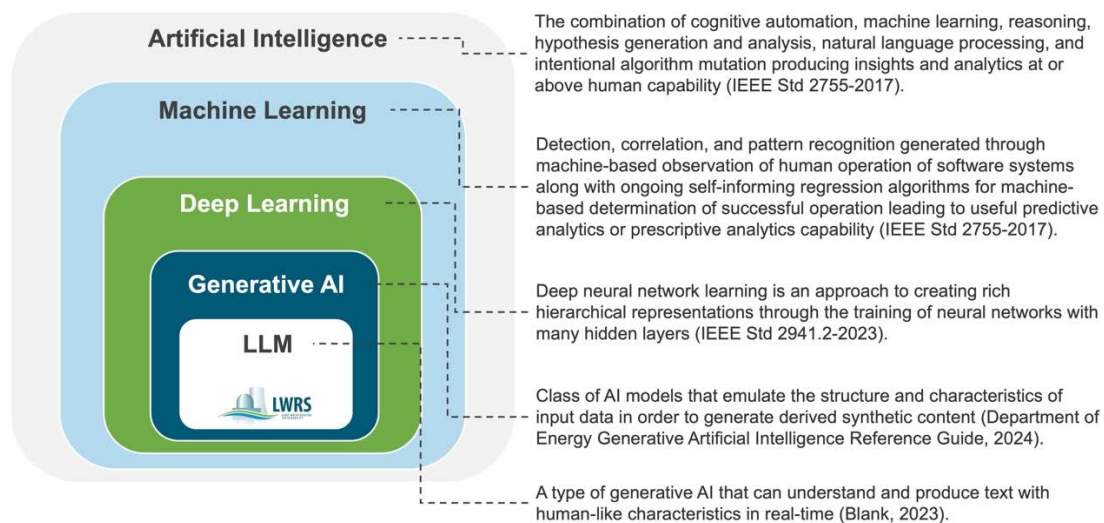


Figure 1. Illustrative conceptualization of LLM (adopted and enhanced from Department of Energy Generative Artificial Intelligence Reference Guide, 2024).

LLMs offer significant potential in improving work efficiencies through automating commonplace tasks that demand text generation, summarization, code generation, or writing. LLMs are being used across many industries, with some related to nuclear electricity generation. Most recently, Kovesdi et al. (2025) had completed an initial business case for the use of LLMs at nuclear power plants. This work identified that the impact of LLMs used at nuclear power plants could see as much as 35% of the workforce benefiting, spanning across a diverse range of functional areas including training, projects, oversight and audits, emergency planning, and maintenance planning. In Kovesdi and colleagues' work, a key finding from the business case analysis was that the use of LLMs ultimately can provide personnel with enhanced access to important plant information to better support decision-making, enabling them to be more proactive when performing work duties.

While LLMs offers significant potential to transform the way that businesses, organizations, and industries perform work, the broad deployment of this technology in the nuclear industry is not without its challenges. For instance, among other issues discussed later in this report, currently available LLMs can hallucinate, where the response (i.e., their output) is neither factually accurate nor logically sound. Hallucinations are one type of major threat to an LLM’s trustworthiness. An untrustworthy response may appear logically sound to the user at face value, and traceability of a response to an original data source may be impossible to obtain. The potential consequences of inappropriately trusting an LLM can have deleterious consequences, such as misusing unreliable LLMs and disusing reliable LLMs (Parasuraman & Riley, 1997), especially when it is being used for critical applications, such as in operating or maintaining a nuclear power plant. Furthermore, the impact of a user’s experience with untrustworthy AI could have negative ramifications on future adoption of any new AI-based technology (Kovesdi, 2021). Indeed, identifying methods that support the design and evaluation of trustworthy AI and resulting user trust is critical to adoption and is also an open area that demands future research to successfully build safe, secure, and trustworthy AI.

1.1 Scope of Research

Following a multidisciplinary approach (comprising data science and human factors scientists), this work developed a methodology that improves trustworthiness of LLMs and ensures appropriate trust when used in the context of nuclear power plant work. The methodology brings forward guidance that addresses both technical and sociotechnical considerations across the development, deployment, and evaluation of LLMs. This research is pertinent for the accelerated adoption of LLMs to enable the U.S. nuclear industry in achieving energy dominance. The methodology is vendor- and model-neutral; its application spans the lifecycle of developing an LLM, beginning with work and data scoping and through application integration and use.

The users of this methodology may include stakeholders in the nuclear industry who are responsible for either the development or integration of LLM technology. For these users, this work ensures that LLM technology is implemented in a way that is trustworthy for the specific work that is being supported at the nuclear power plant. Further, plant personnel who will ultimately use the LLMs may refer to this work in understanding what factors contribute to LLM trustworthiness and user trust, in an effort for making iterative improvements to their current work processes or any future work enhancement opportunities.

1.2 Report Structure

Sections 1 and 2 provide an introduction to this research space including a background of LLMs, followed by a discussion of state-of-the-art models. Sections 3 and 4 provide the backdrop to the current methodology development by defining the important concepts of LLM trustworthiness and user trust. It structures its guidance around technical activities that occur across the LLM application’s lifecycle, represented as four general phases: 1) Work and Data Scoping, 2) Model Selection and Data Preparation, 3) Model Optimization, and 4) Model Integration.

The technical activities within these phases, shared in Sections 5 through 8, are discussed in terms of how the most significant risks to trustworthiness and trust (i.e., referred to as cross-cutting issues) that may be encountered can be effectively mitigated through the guidance provided. Each technical activity provides a recommended approach for ensuring trustworthiness and user trust. Section 9 provides a summary of the cross-cutting issues and requirements addressed by the methodology. Finally, Section 10 concludes this research and offers next steps.

2 BACKGROUND: OVERVIEW OF LARGE LANGUAGE MODELS

This section provides background information related to LLMs. Here, information related to how LLMs work, state-of-the-art models, functions of LLMs, and their capabilities and opportunities for enhancing existing work processes is given. This section serves as foundational information in characterizing LLMs.

2.1 Defining LLM

An LLM is a recent and advanced type of AI that can understand and produce text with human-like characteristics in real-time (Blank, 2023). LLMs are used in natural language processing (NLP). NLP is the field of study concerned with identifying characteristics of written language using computational resources. LLMs can be trained on terabytes of data from libraries, websites, and databases. The goal of most LLMs is to predict the next word in a sentence. The LLM applies this one-word-at-a-time approach to generate paragraphs and even pages of text as instructed by a human. Broadly, LLMs are used in various applications such as translation, summarization, and conversation agents like chatbots. The novelty of LLMs relative to traditional NLP is that LLMs understand context and meaning to produce text that is relevant, coherent, sophisticated, and informative.

2.2 How LLMs Work

This section will explain how LLMs work by exploring the computational concepts underpinning LLMs and will begin with an important component of LLMs, the *neural network*. A neural network is a computational model inspired by the human brain consisting of neurons connected to each other. Each connection between pairs of neurons has a strength, known as a weight, and is stored in a *data structure*. A data structure organizes and stores data on a computer. The data structure that stores neurons and their connections is called a *matrix*. A matrix is typically a table of numbers arranged in rows and columns and is widely used in mathematics, engineering, and science. Specialized computer hardware known as graphics processing units (GPUs) and tensor processing units (TPUs) process these matrices efficiently. The computational power and the amount of memory of these processing units is crucial to have cutting-edge LLMs. Another key organization of the neurons is *layers*, one of the fundamental building blocks of neural networks. Layers are stored in a data structure called a *tensor*, a collection of matrices. When there are more than 3 layers of neurons in a neural network it is called *deep learning*, but deep learning is often dozens to hundreds of layers deep.

Neural network data structures are operated on by a collection of *algorithms*. An algorithm is a step-by-step process that a computer uses to solve a problem. Algorithms enable predictions to be made from neural networks as well as the training of the neural networks to be done. Predictions are made after the neural network has been trained on data such as text from Wikipedia. The input to a neural network could be a sentence like “INL is a [MASK] in nuclear energy.” In this context, a token is a basic unit of text, such as a word, subword (a meaningful word fragment, like “believ” in “unbelievable”), or punctuation mark that the neural network processes during training and prediction. The output of a neural network is the [MASK] token. If the active output neuron is “leader” then the neural network has a low error. If the active output neuron is “banana” then the error is high. The neural network learns by adjusting the *strength of the connection* among the neurons to achieve a smaller error. In the context of neural networks, the strength of the connection refers to the weight or degree of influence that one neuron has on another neuron within the network. Going back to the previous example, the output neuron for the word “banana” would learn to have weaker connections to neurons that are active when the topic is “nuclear energy.” This is done to achieve lower error. Error measurement and neuron connection strength are calculated automatically using algorithms. Neural network algorithms, data structures, and hardware are key areas of research for LLMs.

LLMs use deep learning in a novel neural network architecture termed a *transformer*. Transformers were mainstreamed due to a paper written by Vaswani et al. (2017). Transformers connected neurons in a new way that improved their ability to learn language. Training the neurons of a typical LLM is similar to that described in the previous paragraph. It is accomplished by withholding the next word in a sentence

from the LLM and then asking the LLM to predict that word. The model measures the error in its prediction and gradually the errors decrease as the model learns more about language (see Figure 2).

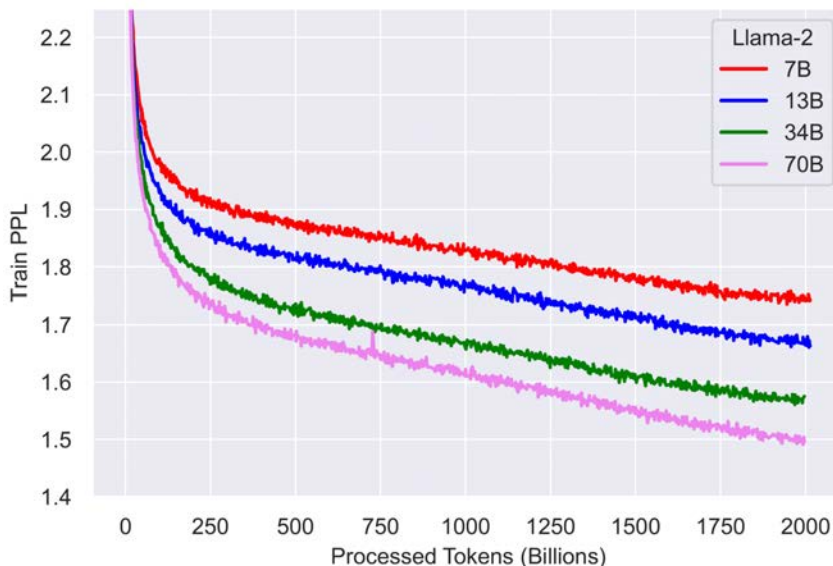


Figure 2. The error of several LLMs decreased during training (adapted from Touvron et al., 2023).

In this figure *perplexity* (PPL) on the y-axis represents error and is how well the model understands the language; lower perplexity means better understanding. The x-axis is how many tokens the LLM has been trained on. This figure shows that the more text seen during training, the smaller the error, and consequently, the quality of the model improves. Notice in this figure the extent of training, 2 trillion tokens. This is the equivalent of approximately 1,000 times the size of the English Wikipedia or approximately 20,000 times the size of the Encyclopedia Britannica. Also note that the lines do not flatten; this indicates that the model is still absorbing information even though it has seen 2 trillion tokens. Datasets this large are necessary to increase the LLMs ability to understand language because the model continues to improve even with these larger datasets.

During inference the LLM chooses the next word to produce text by focusing on the relevant parts of the input text. The concept of focusing on relevant words is based on the attention cognition principle. This occurs in a specialized attention algorithm popularized in an article by Vaswani, et al. (2017) titled “Attention is All You Need.” Attention occurs throughout the transformer by alternating with layers of neurons. Figure 3 presents a very high-level overview of how a transformer works and the role of attention in this process.



Figure 3. A basic flowchart of an LLM during the word generation process.

In step 1 of the transformer, the words are converted to tokens, for example, the word “started” may be tokenized into “start” and “ed”. In step 2, the tokens are converted into a numerical representation consisting of hundreds of numbers, this is called an embedding. In step 3, the embeddings are put into the attention algorithm. In step 4, the attention results are given to a layer of neurons. Steps 3 and 4 repeat alternating with attention and neuron layers. Finally at the end of the alternating attention/neuron layers is the next

word in the output text. The last layer of neurons has one neuron for each word. The output word corresponds to the most strongly activated neuron. This flowchart omits many details but provides the most important aspects of an LLM.

2.3 State-of-the-Art Models

At the time of this report, some of the state-of-the-art LLMs are OpenAI ChatGPT 4o and OpenAI o1, Meta LLaMA 3.2 and 3.3, Microsoft Phi-3 and Phi-3.5. In particular cases it is important to use different size models. Therefore, both Phi-3 and Phi-3.5 are the latest models, the same applies to LLaMA 3.2 and 3.3. The OpenAI ChatGPT 4o is widely available to this effort researchers, but the o1 model is not yet, therefore again both models are the most recent and state-of-the-art depending on their availability. The OpenAI models are cloud-based, and the LLaMA and Phi models can be run locally, this will be an important consideration for this work.

2.4 Functions of an LLM

There are three principal functions in NLP that LLMs support: 1) classification, the categorization and labeling of text data, 2) text comparison which identifies similar texts, and 3) dialogue which facilitates human-AI interactions. LLMs may support additional functions. For instance, for INL-developed software Machine Intelligence for Review and Analysis of Condition Logs and Entries (MIRACLE), a Research and Development (R&D) 100, 2022 award winner, its methods and tools are grouped into 10 LLM functions. The additional functions are estimation for predicting continuous outcomes, generation for creating new content, guidance for text recommendations, insertion for facilitating new information, review for assessing unstructured documents, structuring for data organization and extraction and, summarization for concise overviews of lengthy texts. A more detailed discussion can be found in Al Rashdan et al. (2024).

2.5 LLM Capabilities and Opportunities for Work Efficiencies

LLMs are capable of understanding and generating human language, representing knowledge, and performing reasoning tasks, which offers significant potential in assisting with and automating tasks. These capabilities offer significant opportunity for improving work efficiencies at nuclear power plants, such as:

- 1) autodrafting of various documents (e.g., drafting engineering reports, 50.59 screenings, etc.),
- 2) serving as a conversational interface to support plant personnel query plant-specific information and retrieve contextually relevant responses, as well as
- 3) analyzing and synthesizing large amounts of information from past operational experience, maintenance records, and condition reports to provide insights into equipment failure modes, root causes, and maintenance strategies (Kovesdi et al., 2025).

Significant cost savings are envisioned through these work efficiencies resulting in better prioritization of time spent by personnel in performing high-value tasks. For instance, engineers who are equipped with LLMs will be able to focus on activities like system health monitoring, technical backlog reduction, and strategic equipment upgrades, as opposed to completing labor intensive administrative tasks. This will allow them to plan and act proactively, thereby preventing equipment failures that can lead to forced outages. Though in order to realize these benefits, considerations related to LLM trustworthiness and calibrated user trust must be addressed. The next section discusses these two key concepts, which are central to this research.

3 CHARACTERIZATION OF TRUSTWORTHINESS AND USER TRUST

The purpose of this section is to provide foundational information in characterizing AI trustworthiness (a characteristic of the LLM) and user trust (a characteristic of the user). The conceptual relationship between trustworthiness and user trust is illustrated in Figure 4.

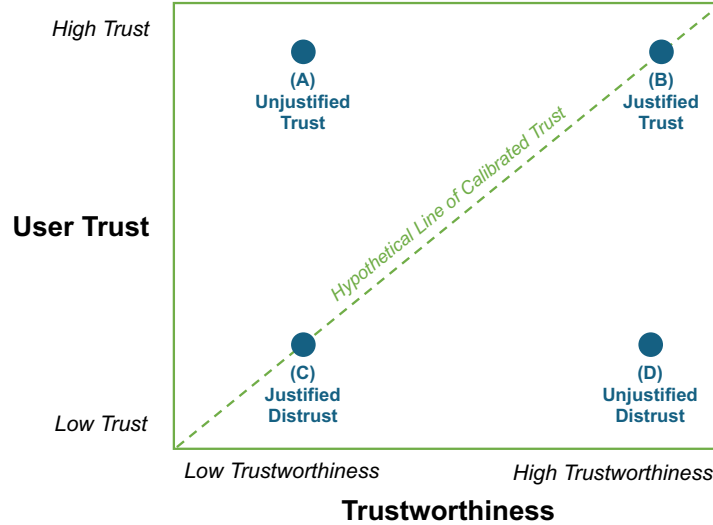


Figure 4. Relationship between trustworthiness and user trust.

A key feature of Figure 4 is the alignment, or calibration, between the level of trustworthiness for the LLM and the level of trust that the user puts in place when interacting with the LLM. Points (C) and (B) portray instances where trust is appropriately calibrated with the trustworthiness of the AI. When the AI is not trustworthy, the user rightly does not trust it (Point C). When the LLM is trustworthy, the user rightly trusts it (Point B). While there are technical considerations required in developing LLM that is trustworthy, factors that contribute to calibrating user trust is largely sociotechnical in nature. Without addressing the sociotechnical factors, even a highly trustworthy LLM may be poorly adopted due to users failing to trust and use the LLM (i.e., known as *distrust* – Point D). Thus, both trustworthiness and trust must be considered to design effective AI systems. Section 3 provides a common reference point that ensures consistency in defining these two constructs, as well as in specifying the considerations that must be addressed. Trustworthiness is first discussed (Section 3.1), followed by user trust (Section 3.2).

3.1 Trustworthiness: A Characteristic of the LLM

In this section, a historical perspective on trustworthiness in AI is given, before characterizing the key attributes (or characteristics) of trustworthiness in LLMs.

3.1.1 Background: Trustworthiness in AI

Trustworthiness of computers was explored in 1950 by Alan Turing using a test popularly called the *Turing Test* (Turing, 1950, reprinted 2009). The goal of the Turing Test is to evaluate if computing machinery can imitate a human being. This is a logical evaluation of trustworthiness because a human being is the gold standard for trustworthiness. In some limited ways the Turing Test has since been successfully completed with modern AI systems. This trustworthiness evaluation laid the groundwork for future trustworthiness evaluations in the field of AI. The term AI was coined five years later in 1955 (McCarthy, 1955 reprinted 2006) and led to the first well-known chatbot developed at MIT in the 1960s called ELIZA (not an acronym) (Weizenbaum, 1976). ELIZA demonstrated trustworthiness because superficially 1) the program had the ability to understand conversations and 2) the program aligned with user expectations. ELIZA was one of the first chatbots capable of attempting the Turing Test (Marino, 2006).

In the 1980's there was a new wave of enthusiasm for AI and in particular *expert systems*. An expert system uses rules to solve problems. Knowledge is transferred from experts to knowledge engineers and then into the expert system. The software MYCIN (not an acronym) was an expert system that helped physicians diagnose and treat bacterial infections. The performance, completeness, and consistency of AI, which are key factors in the trustworthiness of AI are examined in Buchanan and Shortliffe (1984). The performance of MYCIN is discussed here: MYCIN, seven Stanford doctors, one Stanford medical student, and the county hospital selected antimicrobials for 10 patients. Non-Stanford infectious disease specialists then evaluated if the correct antimicrobial was selected. MYCIN recommended the correct antimicrobial 70% of the time. The best human doctor also was correct 70% of the time. But the rest of the human doctors were worse than that. From a trustworthiness perspective MYCIN has been shown to be equal to or better than a human doctor. However, the authors noted that the 70% figure was disappointingly low for the system and 90% was desired for adequate trustworthiness in clinical use. Nevertheless, there was much disagreement even between the experts. This disagreement was apparent because the best human was only 70% in agreement with the infectious disease specialists.

MYCIN uses rules that are easily examined and critiqued by experts, this is different than other methodologies that cannot be easily understood by experts. These rules in MYCIN were checked for completeness and consistency, key aspects of trustworthiness of AI. Incompleteness occurs because the knowledge base on which MYCIN is built is missing rules. As an example, when MYCIN faces a situation that has no rules it cannot produce an appropriate output. To locate these missing rules, it would be necessary to identify every combination of patient symptoms. In practice this is not possible because as an example a male cannot be pregnant. Moving to consistency, there are two basic logical checks performed: 1) identifying conflicting rules and 2) identifying redundant rules. A conflicting rule occurs when the AI decides on two different courses of action for the same situation. A redundant rule occurs when two different rules are applied but the same output is given. These two logical inconsistencies are located by searching for situations that apply multiple rules. Thus, the process of examining MYCIN's rules for completeness and consistency increases the trustworthiness of the AI system despite the challenges mentioned.

In the 1990s to the 2000s, a primary limitation on the success of AI and its trustworthiness was the computing capacity available. Moore's law states that computing capability doubles at regular intervals (Schaller, 1997). This made possible the success of the International Business Machines (IBM) Corporation supercomputer Deep Blue that challenged the world grandmaster in chess (Hsu, 1999). While Deep Blue highlighted a milestone in history as the first computer to defeat a world chess champion, the specific strategy that it used to defeat Garry Kasparov has been long debated. Notably, it was debated whether the defeat was a result of a bug in Deep Blue, that was misinterpreted by Mr. Kasparov as a strategic move. From a trustworthiness perspective, this event invited important questions related to the design of AI regarding ways in which to design such systems to be *explainable* and *interpretable* (Kasparov, 2018).

Moving into the 21st Century, the proliferation of AI technologies has been substantial (Lie et al., 2018). AI can be found in nearly every facet of modern society, including commercial products (e.g., Gkikas & Theodoridis, 2021), medical devices (e.g., Gadde & Kalli, 2020), automobile (Kamran et al., 2022), and industrial processes such as nuclear power plants (Lu et al., 2020) to name a few. One notable study was completed by Khalid and colleagues (2023a), who completed a comprehensive literature review in determining the different ways in which *trustworthiness*, as a construct, has been defined in the nuclear power plant and non-nuclear sectors. Their work characterized each literature source by whether it explicitly defined trustworthiness or not (referred to as implicit), whether the source was nuclear-related or not, and whether the approach used in defining trustworthiness focused on actual observations set by the AI (output deviation-based), or whether trustworthiness focused on different attributes comprising trustworthiness as a construct (attribute based). Challenges in defining trustworthiness were identified. For output deviation-based definitions, actual observed performance of the AI and relevant thresholds for acceptable performance is needed. This can be challenging in defining trustworthiness for AI systems that are yet to be built (e.g., where there is a lack of such data). Likewise, attribute-based approaches have their own

challenges. For instance, the Khalid and colleagues suggest that such definitions often lack objectivity, generally relying on subjective factors that rely on expert opinion. Based on this literature review and identification of challenges, Khalid offer a third approach based on a definition for trustworthiness specifically tailored for the nuclear industry:

- “*The degree of confidence that an automation (AI) system will function as expected.*”

Khalid and colleagues present a risk-informed approach to trustworthiness evaluation; they posit that the “degree of confidence” in this definition is meant to establish a quantifiable measure of uncertainty associated with the AI’s output. They provide an example of how their approach can be applied to an automated firewatch application (Khalid et al., 2023b).

Currently, a diverse set of approaches exist for defining trustworthiness with LLMs (e.g., de Cerqueira et al., 2025; Kowald et al., 2024; Huang et al., 2024; Huang et al., 2025; Wirz et al., 2025). Notably, Huang and his team (2024) proposed probably the most comprehensive benchmark of trustworthiness in LLMs available to date across six attributes: truthfulness, safety, fairness, robustness, privacy, machine ethics, and accountability, in a comprehensive study called *TrustLLM* evaluating 16 mainstream LLMs using 30 datasets. Huang and colleagues define:

- *Truthfulness* as being linked to factuality.
- *Safety* in their definition also incorporates security concerns, defined as the ability to avoid unsafe or illegal outputs; safety in this context highlights how LLMs can be vulnerable to various exploitation attacks, such as jailbreaking (i.e., bypassing safety mechanisms in place for the AI).
- *Fairness* is defined as the ethical principle of ensuring that the LLM is designed, trained, and deployed in ways that not lead to biased or discriminatory outcomes.
- *Robustness* is defined as the ability of the LLM to perform well under varying conditions, handling exceptions, anomalies, and unexpected inputs properly. In this context, ordinary user interactions can be reflective of how robust an LLM is in terms of handling such input to produce an expected outcome.
- *Privacy* refers to the norms and practices that help safeguard human and data autonomy, identity, and dignity.
- *Machine ethics* is loosely defined. The authors refer to it as being broadly characterized by three segments including implicit ethics (judgement of moral decisions), explicit ethics (how the LLM should react in different moral environments), and emotional awareness (capacity to understand its own abilities and recognize human emotion).
- *Transparency* refers to how much information about the LLM and its output is available to users.
- Finally, *accountability* refers to its obligation to inform and justify its conduct to an authority. Such consideration refers to an ability to discern human-produced and LLM-produced data and may relate back to the vetting of data source(s) itself.

These attributes are likely to additively contribute to characterizing overall trustworthiness for LLMs. Thus, acknowledgment is made that there may be many factors that affect the LLM’s trustworthiness, such as those highlighted by Huang and colleagues (2024). In reviewing applicable standards, a complementary definition for *trustworthiness* from the National Institute of Standards and Technology (NIST), the International Organization for Standardization (ISO), and the International Electrotechnical Commission (IEC) in ISO/IEC TS 5723:2022 is given:

- The ability to meet stakeholders (i.e., any individual, group, or organization that can affect, be affected by, or perceive itself to be affected by a decision or activity) expectations in a verifiable (i.e., can be checked for correctness by a person or tool) way.

This definition contains notable attributes worth highlighting in terms of its relevance to this work. First, trustworthiness refers to *meeting stakeholder expectations*. The term *expectations* here suggests that the trustworthiness of AI is largely coupled with its *context for use*. Here, *context of use* is defined as:

- Users, tasks, equipment (hardware, software, and materials), and the physical and social environments in which a product is used (ISO 9241-11:2018).

An important point is that *expectations*, as used in the definition for trustworthiness, is likely influenced by the particulars of the user’s experiences with using AI, the tasks that they are to perform, and their working environment (i.e., accounting for both physical and social factors). An AI system of certain capabilities therefore may have appropriate trustworthiness (i.e., fitting the user’s expectations) in some circumstances, but not in others. This aspect necessitates not only a technical approach to engineering trustworthiness, but to also consider the sociotechnical factors that affect the context for use surrounding the AI and its effect on the user’s propensity in trusting it and therefore using it.

A second attribute worth highlighting from the definition is that trustworthiness must be *verifiable*, meaning that it can be checked for correctness. This attribute necessitates an ability to *measure* and determine what is *acceptable*, or *correct*, in terms of meeting stakeholder expectations. Thus, ensuring trustworthy AI demands testing and evaluation with clear acceptance criteria that directly map to stakeholder expectation. Indeed, these two attributes of the definition for trustworthiness has implications to the methodology discussed in this technical report. Next, the characteristics of trustworthiness as applied to LLMs are brought to attention.

3.1.2 Factors that Influence Trustworthiness

Meeting stakeholder expectations may have many, sometimes conflicting, considerations; thus, the concept of trustworthiness is multifaceted. Similar to the attributes discussed in Huang et al. (2024), trustworthiness is further characterized in the industry standard: NIST AI 100-1, *Artificial Intelligence Risk Management Framework* (2023). The factors that influence trustworthiness, or characteristics, included in NIST AI 100-1 are presented in Figure 5.

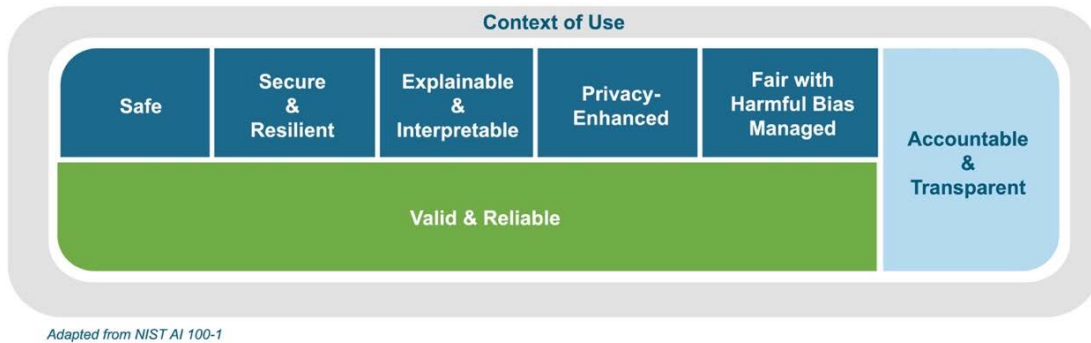


Figure 5. Characteristics of trustworthy AI systems (adopted and enhanced from NIST AI 100-1 [2023]).

These characteristics of trustworthiness inevitably require balancing and managing tradeoffs, which are largely determined by the specific uses of the LLM and the associated risks. The figure thus depicts the characteristics of trustworthiness being subsumed within its context of use. Further, the layout of the characteristics in Figure 5 are deliberately drawn to illustrate their relationship to one another. Namely, *Valid and Reliable* is a necessary characteristic in achieving the five characteristics above; this is reflected by *Valid and Reliable* being horizontally spread across each of the other five characteristics. Also,

Accountable and Transparent is related to all other characteristics and therefore is shown as a vertical slice. These characteristics are described in NIST AI 100-1 as follows.

3.1.2.1 Valid and Reliable

Valid and reliable pertains to ensuring that the AI is as accurate and reliable as possible across a variety of circumstances (i.e., referred to as robustness). The definition of *accuracy* used in this context is described as “closeness of results of observations, computations, or estimates to the true values or the values accepted as being true” (ISO/IEC TS 5723:2022). The measurement of accuracy should consider computational-centric measures, as well as measures of human-AI teaming. *Reliability* used here is defined as “the ability of AI to perform as required, without failure, for a given time interval, under given conditions” (ISO/IEC TS 5723:2022). *Robustness* is defined as “the ability of the AI to maintain its level of performance under a variety of circumstances” (ISO/IEC TS 5723:2022). Robustness thus encompasses ensuring that the AI performs exactly as it does under expected use cases, while also minimizing potential harm (risk) to people if it is operating in an unexpected use setting.

3.1.2.2 Safe and Secure

Safe refers to AI ensuring that human life, health, property, or environment are not endangered (ISO/IEC TS 5723:2022). From this definition, it can be reasoned that safety may be interpreted in a broader sense that includes not only risk of harm, but potentially to aspects like economic loss or reputation in an organization (Kovesdi et al., 2024).

Security refers to the ability of the AI system to maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use. Common security concerns may include data poisoning or exfiltration of models, training data, or intellectual property through endpoints of the AI. While separated in the NIST framework, safe and secure are grouped here as each consideration follows a risk-informed approach in identifying such conditions or circumstances that impact safety or security, and strategizing mitigating controls that reduce risk to an acceptable level.

3.1.2.3 Explainable and Interpretable

Explainable and interpretable assists those using or overseeing the AI system to gain deeper insights into the functionality and other aspects of trustworthiness of the AI system itself. *Explainability* refers to a representation of the mechanisms underlying the AI system’s operation (NIST AI 100-1, 2023). *Interpretability* refers to the meaning of the AI system’s output in the context of their design functional purpose(s) (NIST AI 100-1, 2023). NIST AI 100-1 (2023) notes key distinctions between transparency, explainability, and interpretability. Transparency answers the question “what happened” in the AI system. Explainability answers the questions of “how a decision was made” in the AI system. Finally, interpretability answers the question of “why a decision was made” by the AI system.

3.1.2.4 Privacy-Enhanced

Privacy refers to the norms and practices that help safeguard human autonomy, identity, and dignity (NIST AI 100-1, 2023). Values of privacy include anonymity, confidentiality, and control; threats to privacy can have negative implications on security, model bias, and transparency. Specific design decisions and/or features of the AI may influence privacy.

3.1.2.5 Fair with Harmful Bias Managed

Fairness in AI includes concerns for equality and equity by addressing issues including harmful bias and discrimination. Such bias may present itself in LLMs as concepts that are systemically overrepresented in the output produced.

3.1.2.6 Accountable and Transparent

Trustworthy AI is dependent on accountability, and accountability is a precondition for transparency. *Accountable* refers to being answerable for actions, decisions, and performance (ISO/IEC TS 5723:2022).

Transparency reflects the extent that information about the AI system and its outputs is available to its users (NIST AI 100-1, 2023). Transparency, by this definition, should provide access to appropriate levels of information that is tailored to the role or knowledge of its users interacting with the AI system.

3.1.2.7 Additional Factors: Relevance, Usefulness, and Usability

The characteristics discussed are pertinent to defining trustworthy AI when evaluating it for nuclear power plant applications. Beyond these characteristics of trustworthiness discussed in NIST AI 100-1 (2023), additional considerations that may be relevant to output presented by an LLM pertain to the **relevance, usefulness, and usability** of the model's outputs, for nuclear power plant application¹. These characteristics are defined as:

- **Relevance.** *Relevance* refers to the extent that a response aligns with the expected content (Diamant, 2025).
 - The choice of terminology, *expected content*, is largely driven by the context of use – the user, the task, and the work environment.
 - For the AI to be relevant it must also be valid and reliable. It also must align with the user's needs in terms of the task being performed and the information needed.
- **Usefulness and Usability.** Usefulness of the AI system's output and its associated usability may be important factors to its trustworthiness.
 - *Usefulness* refers to the extent that a user believes the response will enhance their job performance (Davis, 1989).
 - *Usability* refers to the definition given in ISO 9241-11:2018, meaning the extent to which the AI system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.
 - Both usefulness and usability grouped here since they are key components to a related framework, denoted as the Technology Acceptance Model (TAM; Davis, 1989). TAM has been reasoned to be useful for characterizing technology adoption in the nuclear industry (Kovesdi, 2021; Hall & Agarwal, 2024).
 - These two qualities build on the others discussed. That is, the AI must be sufficiently valid and reliable, safe, secure, accountable and transparent, explainable and interpretable, privacy-focused, fair, and relevant before it can be useful and usable. While these considerations are aspects of the AI, they are evaluated within the context of the user's behavior and beliefs (discussed further in Section 3.2).

3.1.2.8 Summary of Factors

Collectively, trustworthiness for AI systems that are used to support personnel perform work at nuclear power plants is characterized by the model presented in Figure 6. The figure illustrates each consideration of AI trustworthiness. The ability of AI to be valid and reliable is emphasized as being centered in the figure, signifying that this characteristic is necessary before other trustworthiness considerations can be met. As such, the methodology discussed later places particular emphasis on ensuring valid and reliable output.

¹ The basis for these characteristics have been identified through repeated discussions with SMEs who had evaluated LLM output by these added considerations.

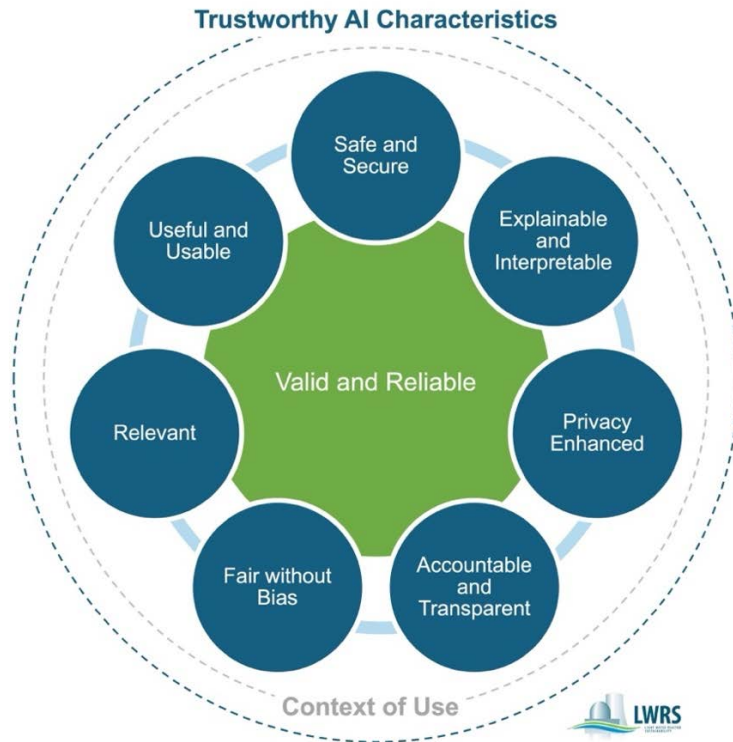


Figure 6. Characteristics of trustworthy AI systems for nuclear power plant applications.

The definitions for each characteristic of trustworthiness previously described apply to the model presented in Figure 6. Like the NIST model presented in Figure 5, *validity and reliability* is a foundational requirement to trustworthiness, that is necessary for the other characteristics to be achieved. As such, validity and reliability has broader applicability across the phases and activities of the methodology, discussed further in Section 4. Further, the *context of use* is all encompassing in defining what levels of each characteristic is needed for the AI to be trustworthy by its end users. The purpose of this model is to provide a consistent way in characterizing trustworthiness of the AI systems that can be verified and validated.

3.1.3 Cross-Cutting Trustworthiness Issues with LLMs

The NIST had published complementary guidance to NIST AI 100-1 (2023), namely *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1, 2024), which is tailored to the application of generative AI systems such as LLMs. NIST AI 600-1 (2024) identifies fundamental trustworthiness issues associated with LLMs and mapped each issue to the corresponding trustworthiness considerations described in NIST AI 100-1 (2023). This work reviewed these issues and identified the ones that would likely apply when supporting critical infrastructure such as nuclear power plant work. These issues include: 1) data privacy, 2) hallucination, 3) harmful bias, 4) human-AI configuration, 5) information integrity, and 6) information safety security.

These issues identified in NIST AI 600-1 (2024) can be summarized as follows:

- **Data privacy** refers to the risk of exposing sensitive data of individuals (e.g., personally identifiable information [PII]), which in some circumstances, can lead to harm to an individual or group of individuals. The risk of privacy here is likely to occur during the ingestion of data that is used for training or fine-tuning of the LLM. Further, intellectual property refers to the risks of using plant sensitive data in a model that is outside of its operational context. For example, models trained on specific plant data and then reintroduced into new use cases, such as for a different nuclear power

plant or new application, could create unwanted risks in intellectual property of sensitive plant data. This work consolidates data privacy issues with intellectual property issues as both are similar in nature within the context of nuclear power plant work. For instance, a major risk concerning data leakage of sensitive plant information has both a data privacy and intellectual property component.

- **Hallucination** is a major threat to LLMs. According to Ziwei et al., (2023), “it is also apparent that deep learning based generation is prone to hallucinate unintended text, which degrades the system performance and fails to meet user expectations in many real-world scenarios.” This is particularly relevant in this report that explores when a failure to meet user expectations degrades trust and trustworthiness. While LLMs are powerful tools, they can occasionally produce incorrect or misleading information (hallucinations), especially when used on text that they have not been trained on.

Haung and colleagues (2025) provide a taxonomy for hallucinations in LLMs. They differentiate between two types of hallucinations, including *factuality hallucinations* and *faithfulness hallucinations*. Factuality hallucinations pertain to a discrepancy between the output generated by the LLM to the verifiable truth; this type is essentially a product of inaccurate information provided by the LLM. Faithfulness hallucinations pertain to a discrepancy of output generated by the LLM from the user input (or prompt). Here, factuality hallucinations are tied to validity and reliability considerations while faithfulness hallucinations are tied to relevance of the output. Haung and colleagues discuss the causes of these two types of hallucinations, residing from the data itself, from the training, or from the inference made by the LLM.

- **Harmful bias** refers to the types of biases that can propagate in an AI system. The implications of bias for LLMs used to support work for nuclear power generation may result in incorrect decision-making. Depending on the specific use of the LLM in terms of work function supporting the plant, the results could be significant (e.g., inadvertent plant trip resulting in economic loss).
- **Human-AI configuration** is a major risk and pertains to misconfigurations and poor interactions between the LLM and human interacting with the system. Human-AI configuration ultimately refers to the poor calibration of user trust (this topic is further covered in Section 3.2). A user may distrust the LLM (i.e., not trust the LLM when it is trustworthy) leading to disuse. This can have a negative impact on work performance, such as inefficiencies. Moreover, a user may mistrust the LLM (i.e., trust the LLM when it is not trustworthy) leading to misuse, which can result in error. The issue of human-AI configuration is treated in this work as a property of trust.
- **Information integrity** refers to the pedigree of data quality ingested by the LLM, in which the data is not technically accurate or is incomplete. Data that comes from manually data entry and from self-report can result in providing technically inaccurate information. For nuclear power plant applications that require high integrity information, data that is of high integrity is necessary when training a model to ensure accurate and reliable information produced.
- **Information safety and security** pertains to potential risks associated with safety and cybersecurity that may occur if the LLM is misused and is associated with risk-important decisions and actions. Additionally, any new risks introduced resulting from lowering the barriers for easing integration of the LLM into other work systems (e.g., integration of the model into other software), or from the LLM itself (e.g., malicious prompt injection) are considered.

Figure 7 shows the relationship of trustworthiness considerations previously discussed as they pertain to these cross-cutting LLM issues.

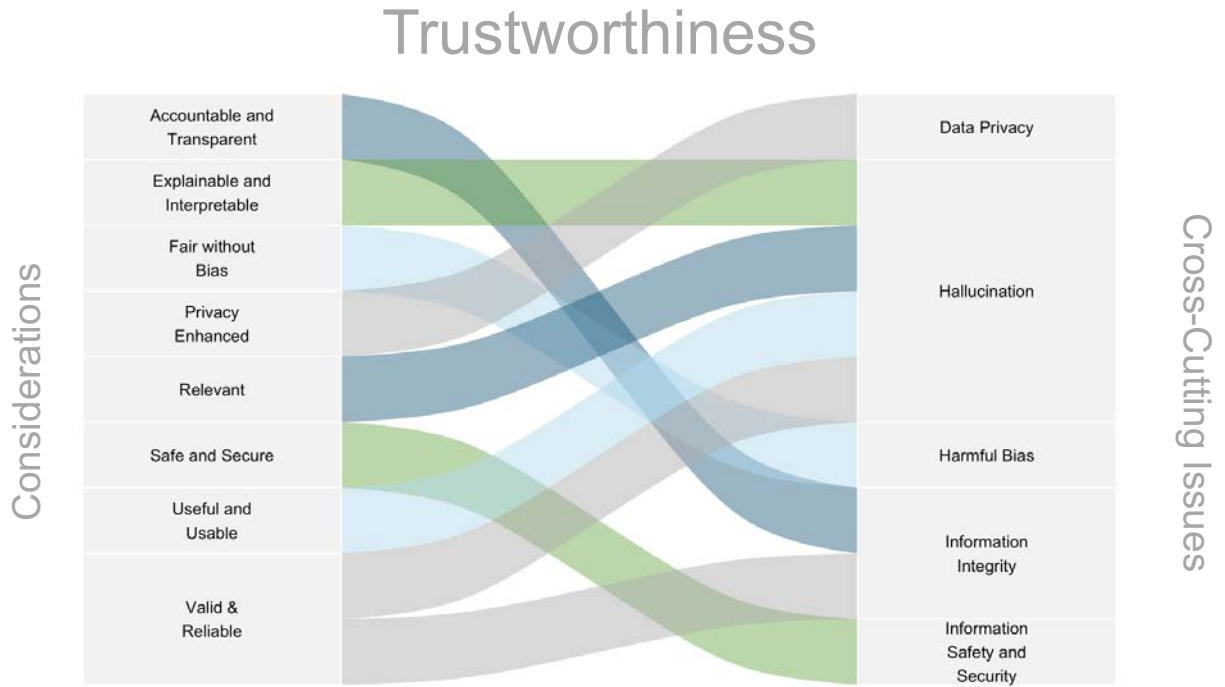


Figure 7. Cross-cutting trustworthiness issues associated with LLMs (from NIST AI 600-1 [2024] and tailored for nuclear power plant work).

Figure 7 presents the relationship between the characteristics of trustworthiness as used in this research to the specific cross-cutting issues they address. This mapping presents a structure to how the methodology discussed in Section 4 addresses relevant cross-cutting issues systematically by each related trustworthiness characteristic (therein referred to as a ‘requirement’ for trustworthiness). Section 3.2 provides a complementary perspective to this section for user trust.

3.2 Trust: A Characteristic of the User

This section discusses the importance for considering user trust when designing and evaluating trustworthy AI. It therefore discusses relevant literature related to human-automation trust and the factors that contribute to user trust, particularly with LLM use. A summary of key factors are then presented as a framework for addressing the relevant human factors considerations in the methodology presented in Section 4.

3.2.1 Background: Human-Automation Trust

The human factors literature indicates that trust is crucial for developing and maintaining successful human-automation interaction (Hoff & Bashir, 2015; Lee & Moray, 1992; Muir & Moray, 1996; Parasuraman & Riley, 1997; Sato et al., 2023; Yamani et al., 2020). Research on human-automation trust was originally initiated as a comparison to interpersonal trust from the 1980s (Rempel et al., 1985). Specifically, Muir (1987) examined whether human-automation trust would follow the same trajectory of the evolution across predictability, dependability, and faith (see also Muir & Moray, 1996). *Predictability* means the degree to which an operator can perceive consistency of observable operation by a machine. *Dependability* means the degree to which the operator is willing to rely on the machine when they are not performing the task. Finally, *faith* means the degree to which the operator can generalize direct experiences they had with the machine beyond the behavioral evidence of the machine’s performance. Muir (1987) predicted that, following the trajectory of interpersonal trust development, operators will develop

predictability, then dependability, and finally faith toward the machine; however, the data showed the opposite pattern indicating that operators first develop faith, then dependability, followed by predictability. Muir's (1987) and Muir and Moray's (1996) seminal work indicate that human-automation trust may evolve following a different trajectory than predicted by interpersonal trust theories (also see Long et al., 2022 for replication studies of Muir & Moray, 1996).

Following Muir and Moray (1996) and related work on automation trust, Lee and See (2004) proposed a theoretical and testable framework across three levels of attributional abstraction – performance, process, and purpose. Briefly, performance means what automation is doing in an observable manner, process means the underlying logic and algorithm that dictate performance of automation, and purpose means the automation designer's intent of why automation is even developed and expected to do. Recent experiments show the three levels of attributional abstraction can be dissociated. For example, operators performing concurrent tasks in a multitasking flight environment showed increased trust on the performance dimension but not on the process nor purpose dimensions (Sato et al., 2023), while operators receiving some context about an automated support system indicated greater levels of process-based trust than those who did not receive the information.

Lee and See (2004) defined trust as:

- “...an attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability.”

A number of previous works indicate that trust influences the performance of human operators interacting with technologies equipped with varying levels of automation (Mayer et al., 1995), visual scanning patterns (Karpinsky et al., 2016), and behavioral intention to accept the technology (Chancey et al., 2017; Yamani et al., 2024). While trust is known to influence multitudes of behaviors of operators interacting with the technology, a psychological construct critical for the safety of high-consequence domains such as nuclear power plant operation, a number of factors can also influence trust in automation.

Building on the previous work on automation trust, one of the most comprehensive model of automation trust currently available in the literature is a three-layered model of trust offered by Hoff and Bashir (2015; see Figure 8).

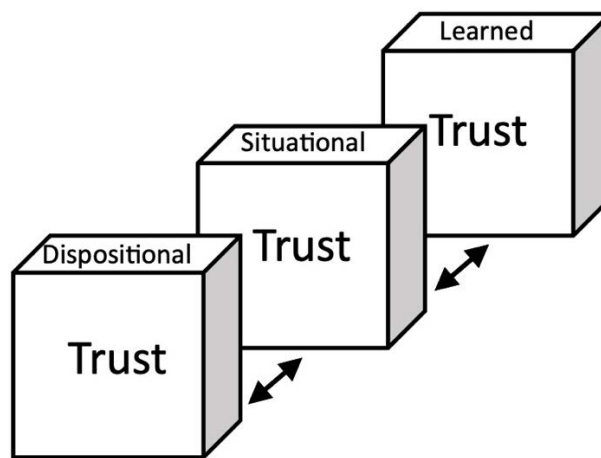


Figure 8. Three-layered framework for conceptualizing trust (adapted from Hoff and Bashir, 2014).

Briefly, the model distinguishes dispositional trust, situational trust, and learned trust, to capture both the static and dynamic nature of trust evolution and maintenance during human-automation interaction. First, *dispositional trust* refers to an individual's general tendency to trust automation, independent of context or specific characteristics of the system. Culture, age, gender, and personality traits of the operator are factors that fall under this category.

Second, *situational trust* captures the development of trust in automation that highly depends on variability in each situation. Factors influencing situational trust are classified into either external or internal variability. External variability refers to performance characteristics of the system such as transparency (Jamieson et al., 2022), complexity (Bailey & Scerbo, 2007), and difficulty of the task (Madhavan et al., 2006) that determine the strengths and weaknesses of the system. Internal variability refers to characteristics of the operator that are more transient than dispositional, such as self-confidence, subject matter expertise, automation complacency, and attentional capacity (Parasuraman & Manzey, 2010; Yamani et al., 2024).

In the realm of nuclear power plant operations fused with LLM, subject matter expertise is likely to play a significant role in accounting for performance variabilities in the success of task operations. Finally, *learned trust*, perhaps the hallmark of Hoff and Bashir's (2014) model, contrasts initial and dynamic learned trust that develops prior to and during human-automation interaction, respectively. Preexisting knowledge such as expected system characteristics, its reputations, and the operator's existing mental model of the system sets their initial learned trust, followed by trust calibration based on their perception and experiences on system performance such as reliability (Dzindolet et al., 2003; Bliss & Acton, 2003), predictability and dependability (Muir, 1994; Muir & Moray, 1996) and type of errors (Chancey et al., 2017).

The development of learned trust is modeled as an evolving psychological state that adapts to dynamic experiences with the automation. Interestingly, Lee and See (2004) already conceptualized this aspect of learned trust in their model using three attributional abstracts of trust – performance, process, and purpose. Namely, performance-based trust represents trust that develops solely based on observed behaviors of automation such as reliability, types of errors, and response speed. Process-based trust denotes trust that develops based on the underlying algorithm of automated processes while purpose-based trust refers to trust defined by the ultimate goal of why the automation was developed. Both Hoff and Bashir's (2015) and Lee and See's (2004) models identify trust as a psychological construct that is at times dynamic, develops during human-automation interaction, can be influenced by several factors, and characterizes human performance in complex task environments.

3.2.2 Factors that Influence Trust

It is important to identify which of the various factors notably influence trust with use of LLMs for nuclear power plant applications. This research conducted a systematic review of the human factors literature surrounding autonomous systems such as AI and LLMs to identify various factors related to trust.

3.2.2.1 Trustworthiness

While trustworthiness and trust are sometimes used interchangeably in the literature, trustworthiness of a system can refer to an intrinsic property of the system (i.e., the trustee) manifested to the user (i.e., the trustor; Sheridan, 2019). This was discussed earlier in Section 3.1.

For instance, plant personnel may develop their performance-based trust in an automated system especially when they have access to perceive system behaviors (Sato et al., 2023). A meta-analysis study that examined factors predicting trust in AI showed that both the performance (accuracy) and reliability of AI significantly predict trust in AI (Kaplan et al., 2023), aligning with the common findings in human-automation interaction literature. Interestingly, other perceived AI attributes such as personality, anthropomorphism, and reputation, in addition to system performance (i.e., being valid and reliable), significantly predict trust in AI. Beyond ensuring that AI performance is optimal, this also suggests that AI performance should be clearly communicated to human users to gain their trust. These findings may also extend to trust in LLMs as responses from a LLM may contain statements with various levels of accuracy and precision along with nuanced delivery depending on the context. Further research should focus on how performance and its presentation to users affect trust in LLM.

Coupling of trustworthiness and trust, Stanton and Jensen (2021) proposed trust (T) as a function of user trust potential (UTP) and perceived system trustworthiness (PST) as the following equation:

$$T(u, s, a) = f(UTP(u), PST(u, s, a)) \quad (1)$$

where a user u interacts with an LLM with characteristics s within a context a . Here, with incorporating the characteristics of trustworthiness (e.g., those discussed in Section 3.1.2), the parameter s may be defined as:

$$s = \sum characteristic_i \quad (2)$$

where $characteristic_i$ denotes i^{th} characteristic comprising trustworthiness (e.g., validity and reliability, explainability and interpretability, etc.). Further research is necessary for investigating the function between UTP and PST that form user trust in LLMs and the weights among the attributes of trustworthiness that drive optimal trust in LLM. Though it is worth noting that, at least conceptually, addressing each aspect of trustworthiness should have positive influence on trust.

3.2.2.2 Adequate Knowledge, Skills, and Abilities

In the same meta-analysis study by Kaplan and colleagues (2023), ability-based factors related to personnel skills such as their competency and understanding of the AI system and their expertise in performing the associated task were significant predictors of trust in AI. This suggests that personnel with more experience with LLMs are more likely to trust an LLM.

3.2.2.3 Oversight

Endsley’s work that extended Bainbridge’s seminal paper, *Ironies of Automation*, to focus on AI systems highlights that system error is still plausible for modern AI technologies, like LLMs (2023). As such, there is need to provide adequate oversight for the user to enable effective decision-making, allowing a user to appropriately trust an LLM. This notion of oversight reflects a continued need for the user to remain ‘in-the-loop’ so that they are positioned to make informed decisions at all times. The concept of oversight is also a key requirement for the European Commission’s Ethic Guidelines for Trustworthy AI (2019).

3.2.2.4 Other Considerations

Beyond these three factors that influence trust, internal factors (i.e., individual differences) and situational factors (i.e., risk perception) may also influence trust and thus should be considered.

3.2.2.4.1 Individual Differences

One consideration in the modern nuclear industry is that personnel with extensive experience in reactor operations are also aging, potentially facing issues of cognitive aging (Hall et al., 2024; Yamani et al., 2024), which calls for designing emerging technologies that consider the characteristics of an aging workforce. However, the results of the meta-analysis show no significant relationship between users’ ages and their trust in the AI system. More research should examine the effect of expertise on trust in AI independent of aging, because the results may reflect the cohort effect. Relatedly, their findings suggest individual differences — culture, gender, and personality traits — significantly predict trust in AI, suggesting that there is considerable variability in how personnel skills interact with individual characteristics to shape their trust in AI. For example, individuals with more innovative mindsets tend to trust in the AI system more, implying that future LLMs may be perceived as more trustworthy to those who hold more positive attitudes toward novel technologies.

3.2.2.4.2 Aligned Risk Perception

Decisions are often characterized by an element of risk, in which risk is defined as “the extent to which there is uncertainty about whether potentially significant and/or disappointing outcomes of decisions will be realized” (Sitkin & Pablo, 1992). Perceived risk is an essential component of trust development, where trust in an agent is influenced by how risky a situation is. Depending on the context, a high-risk situation can cause personnel to rely on automation more than human teammates (Lyons & Stokes, 2012), elevate trust in high-workload environments (Sato et al., 2020), and increase the accuracy of participants at the cost of spending more time verifying the automation responses (Loft et al., 2023). Yet, in other contexts, increased risk can cause a decrease in use behavior and trust in automation (Perkins et al., 2010). For example, if users are concerned that the chatbot may provide incorrect information or are concerned that the technology will threaten their job security (Walker et al., 2024), then they may be less likely to use the technology. System aspects, such as automation transparency, have been shown to mitigate the negative effects of risk (Loft et al., 2023). However, a further investigation of risk within the context of human-automation interaction is necessary. Moreover, the unique risks of nuclear power plant operations have yet to be studied within the context of trust in LLMs.

3.2.2.5 Summary of Factors

In the vast literature of human-automation trust, our literature review has revealed three major factors crucial for the success of using AI applications at nuclear power plants (Table 1), and two additional factors that, while are difficult to address through LLM design, are important for consideration.

Table 1. Factors that are likely to influence trust in AI-infused applications at nuclear power plants.

Factors	Areas of Further Study
<i>LLM Trustworthiness</i>	Validity and reliability, safety, security and resilience, explainability and interpretability, privacy, fairness, relevance, usefulness and usability, accountability and transparency
<i>Personnel Knowledge, Skills, and Abilities</i>	Competency and understanding of the AI system, task expertise,
<i>Oversight</i>	Level of supervisory control and situation awareness over AI
Other Considerations	
<i>Individual Differences</i>	Individual differences (culture, gender, and personality traits)
<i>Perceived Risk</i>	The effects of risk on reliance, trust, and use behavior

First, trustworthiness is an LLM characteristic that influences the growth of trust in the LLM through interactions between the human user and the LLM. The characteristics of AI trustworthiness (Section 3.1.2) are considered pertinent constituents of trustworthiness that shape trust in LLMs. Second, user skills, specifically their competency and understanding of the AI system and expertise in the very tasks that they routinely perform in the nuclear power plant, also predict trust. Third, the degree of oversight that a user has over the AI is a predictor to trust. Further, various individual differences such as culture, gender, and personality traits were also found significant predictors of trust, calling for a large study to model relationships among these variables known to shape human-automation trust. With this, risk in general plays an essential role for trust to emerge, and previous research showed that personnel under higher levels of perceived risk use and trust the system more (Sato et al., 2020).

Figure 9 characterizes the key factors that influence trust in LLM for nuclear power plant use.

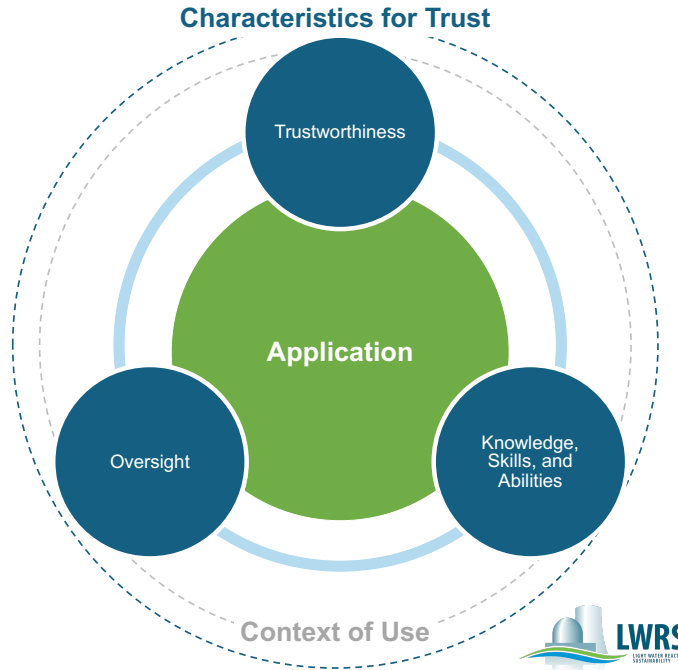


Figure 9. Characteristics of trust for nuclear power plant use.

It should be noted that perceived risk and individual differences discussed earlier are omitted from the figure because they are factors that are external to the design of LLM. For instance, individual differences in users are factors that influence trust, but cannot be changed through design of the LLM. Further perceived risk is a factor influenced by a number of other factors like the individual, the task, or any particulars of the situation itself that grounds interaction with an LLM. While individual differences and perceived risk are considerations that are regarded in this work, the methodology focuses on the factors that can be controlled from an engineering perspective, whether through implementing personnel training and familiarization, establishing guidance for adequate oversight, or through the design of LLM itself (i.e., designing for trustworthiness).

3.2.3 Cross-Cutting Trust Issues with LLMs

Cross-cutting issues related to trust in LLMs for nuclear power plant use can be summarized into two fundamental issues:

- **Adoption** refers to the attitude towards using the LLM. Before users begin interacting with the LLM (and therefore trusting, mistrusting, or distrusting it), they must decide whether to use the system. Their attitude towards using the LLM can be influenced by a number of factors (Davis, 1989; Davis et al., 1989). From the factors characterized in Figure 10, their experiences using similar systems (i.e., their knowledge, skills, and abilities) can influence their initial attitude towards adoption. Further, the level of oversight that the user needs over the LLM based on the user's responsibilities for a given task can also influence adoption.
- **Use** refers to the extent that a user appropriately trusts and uses the LLM. The key issue is the extent that a user's trust is appropriately calibrated so that they trust and use when the system is trustworthy and do not trust and do not use when the system is not trustworthy. The evaluation of trust through interaction with the LLM is central to this issue. Users' knowledge, skills, and abilities, their level of oversight, and the LLM's trustworthiness all contribute as seen in Figure 10.

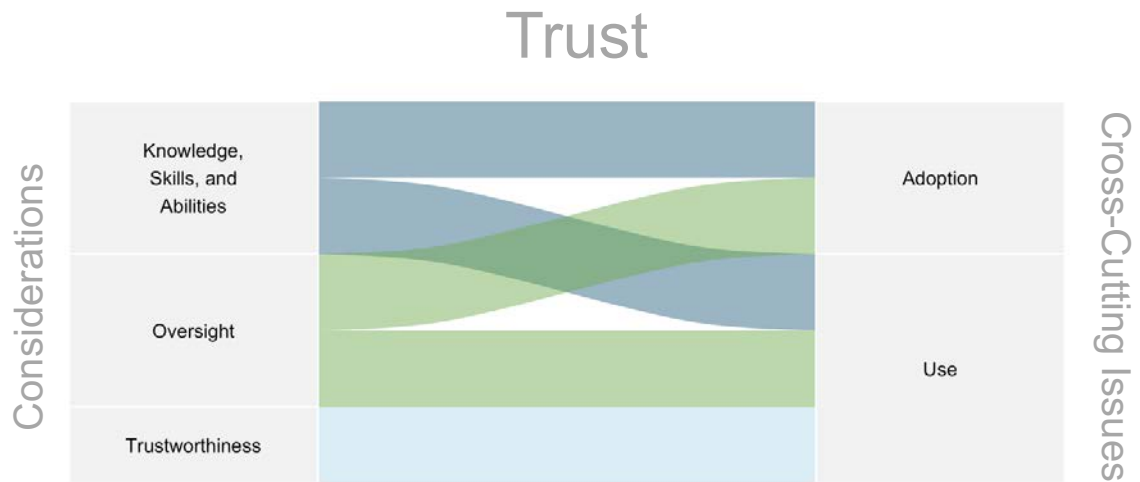


Figure 10. Cross-cutting trust issues associated with LLMs.

Complementary to Figure 7, Figure 10 presents the relationship between the characteristics of trust as used in this research to the specific cross-cutting issues they address. This mapping presents a structure to how the methodology discussed in Section 4 addresses relevant cross-cutting issues systematically by each related trust characteristic (therein referred to as a ‘requirement’ for trust).

4 METHODOLOGY

An integrative approach is needed to address the considerations discussed in Section 3 for designing and evaluating trustworthiness in LLMs while ensuring appropriate levels of user trust. At the most fundamental level, the LLM technology being investigated should embody the technical considerations for designing trustworthiness into the LLM, as well as the sociotechnical considerations for ensuring calibrated user trust, both discussed in Section 3. Indeed, while it is critical to develop an LLM to provide output that is as trustworthy as possible, it is improbable and likely unrealistic to assume that an LLM will consistently provide perfect information in all cases. Importantly, consideration must be made for cases when the LLM provides the user with imperfect information. A well designed LLM should readily communicate to the user of such instances to allow them to appropriately calibrate their trust to the conditions presented to them – i.e., this representing transparency and accountability.

This work therefore presents a methodology that addresses these considerations by applying a multidisciplinary team to address these technical and sociotechnical considerations. An overview of the methodology is shared in Section 4.1, followed by an introduction to informative examples that aim to help demonstrate different aspects of the methodology (Section 4.2).

4.1 Overview

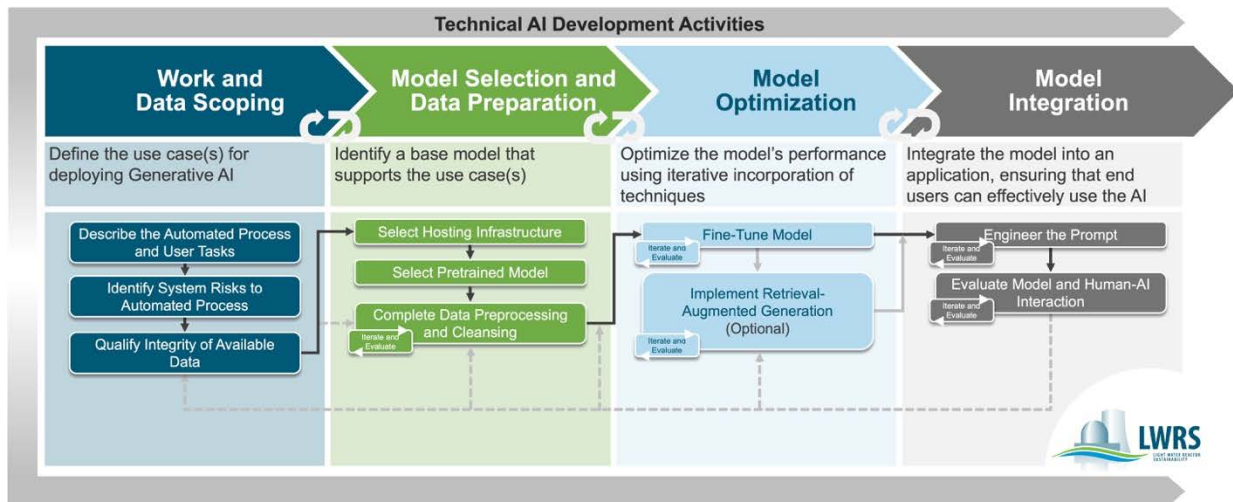


Figure 11. Overview of methodology: An approach to design and evaluate trustworthy AI systems and ensure appropriate user trust.

An overview of the methodology is presented in Figure 11. Attention should first be made to the center material of the figure where there are four phases (in different colors) of LLM development and deployment: 1) Work and Data Scoping, 2) Model Selection and Data Preparation, 3) Model Optimization, and 4) Model Integration.

- **Work and Data Scoping** refers to defining the use case(s) for deploying LLMs to support nuclear power plant work. Here, common activities that are completed include describing the automated process and user tasks, identifying system risks to the automated process, and qualifying the integrity of available data that will be used to prepare the LLM. Activities related to this phase are discussed in Section 5.
- **Model Selection and Data Preparation** refers to the activities associated with establishing hosting infrastructure, identifying a prospective base LLM that supports the use cases under consideration, and completing preprocessing and cleansing of the data. Activities related to this phase are discussed in Section 6.

- **Model Optimization** refers to optimizing the selected model’s performance through fine-tuning and applying retrieval-augmented generation (RAG; optional). These activities are discussed in Section 7.
- Finally, **Model Integration** refers to the activities associated with integrated the prepared model into the work process. Key steps include engineering the prompts and performing iterative evaluation. It should be noted that while evaluation of the model and its use appears linear in Figure 11, the application of evaluation is meant to be applied throughout the process of developing the model. This is emphasized in Figure 11 by the lighter gray dotted arrows looping to relevant technical activities to which evaluation supports. These activities are discussed in Section 8.

These phases are not intended to be prescriptive in terms of specific steps that are required of a utility per se; rather, they are intended to support characterizing major steps in the process of developing a LLM to discuss key activities that address specific cross-cutting issues associated with trustworthiness and calibrated user trust. For instance, an organization may have a specific process to follow when implementing an LLM. The guidance from this methodology, organized by the four general phases, should broadly map to the specific processes that an organization may have in place to prepare a model. Moreover, while the phases are read from left-to-right can be interpreted as being completed sequentially, it is anticipated that these phases (and activities embodied in each phase) are completed iteratively, and thus the circular loops at each phase represent the intent of iteration.

4.2 Informative Examples

To improve the interpretability of the guidance delivered in this technical report, this research uses a set of informative examples to highlight various aspects of the methodology. It should be prefaced that these examples do not, nor should not, reflect the entirety of what work processes LLMs can support at nuclear power plants. Rather, they are intended to highlight how the methodology can be used to address important considerations related to ensuring trustworthiness and user trust.

Example 1: Implementing RAG into an LLM for a Predictive Maintenance Tool

The first example highlights select activities in preparing RAG for an LLM and evaluating user trust for a chatbot that is part of a predictive maintenance tool. This example is based on the Idaho National Laboratory (INL)-developed predictive maintenance tool, Visualization for PrEdictive maintenance Recommendation (VIPER).

Example 1: Implementing RAG into an LLM for a Predictive Maintenance Tool

VIPER was developed to support nuclear power plant personnel understand comprehensive plant health diagnostics by providing diagnosis and prognosis information using various ML approaches (Walker et al., 2023). VIPER recently integrated an LLM chatbot to enhance the explainability and interpretation of the diagnosis and prognosis information provided through its ML capabilities (Walker et al., 2024). While a discussion of VIPER goes beyond the purpose of this report, key technical reports associated with VIPER can be found at INL/RPT-23-74159 and INL/RPT-24-80727, respectively. This example appears at select sections of this report, presented in yellow highlighted text.

Example 2: Demonstrating Data Qualification Process – Reporting and Resolution of an Equipment Deficiency

The second example presents a realistic but fictional scenario pertaining to reporting and resolving an equipment deficiency at a nuclear power plant. This example was developed by an industry subject matter expert (SME) and is used in select sections of this report.

Example 2: Demonstrating Data Qualification Process – Reporting and Resolution of an Equipment Deficiency

Example 2 is presented in green text. The following presents a short narrative that provides context to this example, written by the SME. This narrative first provides background information related to the importance of the Corrective Action Program (CAP) and its relevance to this example. Next, a narrative storyboard is given for this fictional example.

Background into the CAP

Issue reporting is a process at a nuclear plant that is both regulatorily required and is important to safe, reliable, and economical plant operations, and is an element of the CAP. Although issues can be reported within many processes at a nuclear power plant, the main reporting system is the CAP. When reporting issues at a nuclear power plant in the CAP, an event report, commonly known as a *condition report*, is initiated by an employee using a computer-based form (previously a paper template). Nuclear power utilities encourage the reporting of any or all issues that employees encounter so that these issues can be raised up and resolved through the CAP process. The importance of what is reported by the employees ranges from a minor enhancement to some aspect of the plant or its processes, to an issue critical to plant operations. The trustworthiness, such as the integrity and quality (i.e., level of detail) of what is written, of what is reported within a condition report is directly proportional to its importance to the plant.

Throughout the past few decades, the CAP process has been refined such that maximum value is obtained from condition reports, with as little effort as possible to the individuals reporting issues, so that they will report issues freely and frequently. Intelligence has been built into the design of the entry form for condition reports, and the program ties to many databases to maximize the effectiveness of the CAP process. By design, many of the fields are formatted specifically to ensure that the data that is captured can be used with minimal data pre-processing; however, the freeform text fields within the condition report will generally accept anything entered into those fields including special characters, spelling or typographical errors, or even incorrect information, which can create potential issues when using the data to prepare the LLM. By design, most text fields do not filter the information in any way, so that when unusual acronyms or other unusual text is entered, it will be accepted so that the reporting process will not be encumbered in any way thus causing frustration and reducing the issues reported.

Especially for important issues, in order to improve the value and integrity² of the data to the plant, aside from those engineered functions to validate the information within the computerized entry form, the condition report undergoes several levels of “screening” or validation which adds a degree of administrative burden to the CAP process. As the administrative burden added to each condition report varies based on the quality of what is originally entered and the importance of the issue (i.e., when multiplied by all condition reports captured yearly by each nuclear power plant), the overall aggregate administrative burden can be substantial in terms of staffing and data storage and processing.

However, the risks of inaccurate or incomplete information can also be substantial for complex and critical issues, where if left unresolved, the issue can add negative impact on the plant. Therefore, the goal of the screening process is to determine 1) *what*, 2) *when*, 3) *where*, 4) *how* the data will need to be used and 5) *by whom* in a timely manner. Once these are derived, then the significance, regulatory impact, urgency, and level of analysis and correction of the issue can be derived. And with this, based on the plant procedures (and technical specifications), the issue can be resolved.

In this informative example, a condition report is written to document an equipment defect out in the field. From the initial screening of the condition report to the actual scope of the work within the work

² The *integrity* of the data is a fundamental characteristic related to trustworthiness – validity and reliability. It is also a cross-cutting issue identified from NIST AI-600-1 (2024).

management process, various intelligence is added or derived from the condition report and subsequent other plant databases so that the equipment deficiency can be evaluated and eventually corrected commensurate with the potential impact to the station.

Storyboard Narrative

A field operator is performing their rounds, and they notice some water dripping from the outboard seal on the Unit 1 A turbine driven feed pump. The operator documents the deficient condition in the operator rounds program and will need to report the equipment defect in the reporting system. Some plants capture all of their issues in a single system and some capture human related items in one system and equipment related defects in a different system. Based on the plant's digital infrastructure, internal plant intranet access, data quality and accessibility, various information included in the condition report, master engineering equipment list, and inventory database, among real time information, there a significant amount of information that will help to qualify the issue, the trustworthiness of the data, and subsequent automation of this process.

Example 3: Developing a Fine-Tuned LLM Trained on General Nuclear Power Plant Knowledge

The final example highlights different aspects of preparing an LLM on broad nuclear power plant knowledge for general use. This effort is part of a strategic effort in developing tailored LLMs that support specific tasks that will enhance the work management process through further fine-tuning of the nuclear-specific base model that is described in this example. The available LLMs will be part of MIRACLE (Al Rashdan et al., 2024); MIRACLE was developed to support nuclear power plant personnel in performing work management tasks more efficiently and with less error.

Example 3: Developing a Fine-Tuned LLM Trained on General Nuclear Power Plant Knowledge

Example 3 is presented in blue text. This example used INL's available High Performance Computing (HPC) infrastructure to fine-tune the Phi-3.5-mini-instruct model on the publicly available nuclear documents contained in the Nuclear Regulatory Commission (NRC) Agencywide Documents Access and Management System (ADAMS) database, for general use. The ADAMS database contains millions of portable document format (PDF) documents, which were invaluable for training the model on nuclear-specific knowledge.

5 WORK AND DATA SCOPING

The first phase includes three primary activities including:

- Describing the automated process and user tasks
- Identifying system risks to the automated process
- Qualifying the integrity of available data

These activities are discussed next.

5.1 Describe the Automated Process and User Tasks

The first activity entails describing the work process in terms of where automation will be used to improve efficiencies. This can be accomplished through a variety of approaches, such as those in Table 2³.

Table 2. Benefits and shortcomings of different system characterization techniques.

Technique	Benefits	Shortcomings
<i>Process Charting</i>	- Used across disciplines, so likely familiar to many. - Capture actions and decisions of the process. - May be used as a ‘base’ technique to apply other techniques (e.g., cognitive task analysis).	- Can become unwieldy for complex processes. - Limited in scope. Unable to analyze cognitive and trustworthiness requirements at face value.
<i>Hierarchical Task Analysis</i>	- Goal-based, allowing comparison of existing to new ways of working. - May also be used as a ‘base’ technique to apply other techniques.	- Similar to process charting
<i>Cognitive Task Analysis</i>	- Explicitly captures cognitive and trustworthiness requirements.	- Requires subject matter expert (SME) availability and someone versed in human factors.
<i>Cognitive Work Analysis</i>	- Very comprehensive, detailing the work domain, decisions and strategies used, and work competencies needed.	- Requires SME availability and someone versed in human factors. - Can be time consuming to complete.

Regardless of the specific method selected, the nature of information collected from these methods broadly include identifying the goals and functions of the work process, understanding how work is done, identifying the cognitive and trustworthiness requirements of performing tasks within the process, understanding the competencies of personnel, and understanding the communication and coordination requirements for performing tasks with multiple agents (i.e., whether person or AI).

This information is important to understand the level of impact that the LLM will have on existing processes. It also provides requisite information for assessing potential adoption considerations such as the needed knowledge, skills, and abilities, and level of oversight needed over the LLM to ensure personnel maintain adequate situation awareness. For instance, recent work from Dorton et al. (2025) highlight the importance in understanding the broader sociotechnical system, including the people across the organization who interface with users, other interfacing technologies, as well as the broader context of workplace factors including procedures, policies, and organizational culture that may influence trust.

³ The reader may refer to INL/RPT-24-77684 for details of the system characterization techniques shared in the table.

5.1.1 Relevant Cross-Cutting Trust Issues

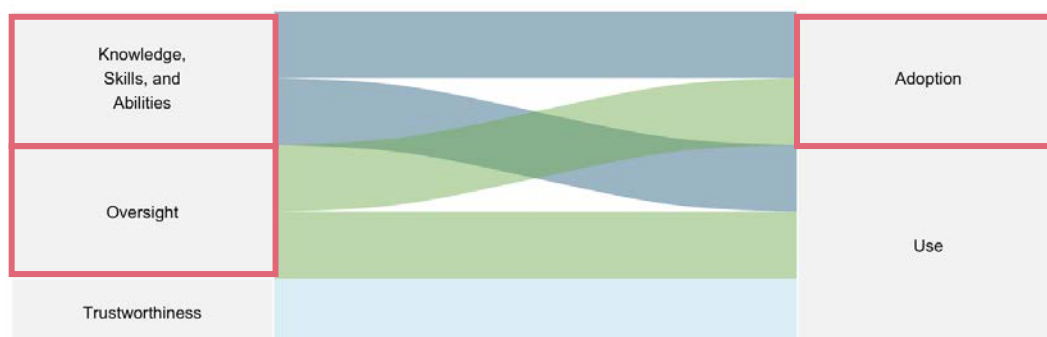


Figure 12. Cross-cutting trust issues addressed when describing the work process and user tasks.

5.1.1.1 Issue 1: Adoption

At this stage, the implementation of LLM is being scoped, so observation of appropriate LLM use is not possible yet; hence, trustworthiness cannot be directly addressed yet. Rather, an important aspect of this activity is a gained understanding the prospective users' attitude towards adoption of LLM, as well as the implications of their knowledge, skills, and abilities and needed level of LLM oversight, as it impacts their attitude (Endsley, 2023; Kaplan et al., 2023).

5.1.1.1.1 Requirement 1.1: Knowledge, Skills, and Abilities

Knowledge, skills, and abilities using LLM are evaluated using the AI Literacy Survey (Tully et al., 2025). The AI Literacy Survey is presented in Appendix C-2 of this report. It comprises a standardized assessment, using a set of multiple choice questions of various AI-focused topical areas (i.e., including programming, algorithms, and terminology of technical concepts), to evaluate aspects of AI competencies. An aggregate score can be captured from the items in the survey. This information can be used to benchmark different groups of users across the organization to inform where training may be particularly appropriate to ensure users have the requisite knowledge, skills, and abilities. The following example highlights how this survey can be used to establish a benchmark score for AI literacy.

Example 1: Implementing RAG into an LLM for a Predictive Maintenance Tool

AI literacy was collected in a study that evaluated aspects of the VIPER LLM (Appendix C provides specific details of the evaluation). The evaluation recruited undergraduate college students from a Psychology Department at a major U.S. university. A benchmark for AI literacy from this population was made (mean AI Literacy score was 13.78). Figure 13 presents a histogram of these scores with the green vertical line representing the average AI literacy score of 13.78. An increase in AI literacy scored by the survey indicates greater competency of using AI technologies like LLMs. Lower values may indicate a greater likelihood of a user poorly calibrating their trust; thus, a significant number of users with lower scores (i.e., when comparing across different organizations) may suggest a need to implement training or familiarization of the respective LLM. From Figure 13, it can be seen that there was a notable range of AI literacy scores, suggesting that there may be different types of users (e.g., novice and power users). The results suggest implementing some training program to ensure all users have a common knowledge base for using AI systems like LLMs.

A similar approach can be used for other use cases in which different groups can be evaluated and compared across an organization. Where literacy is lower, efforts can thus be prioritized to those groups for training to ensure AI adoption is not hinders and later use corresponds to appropriately calibrated trust.

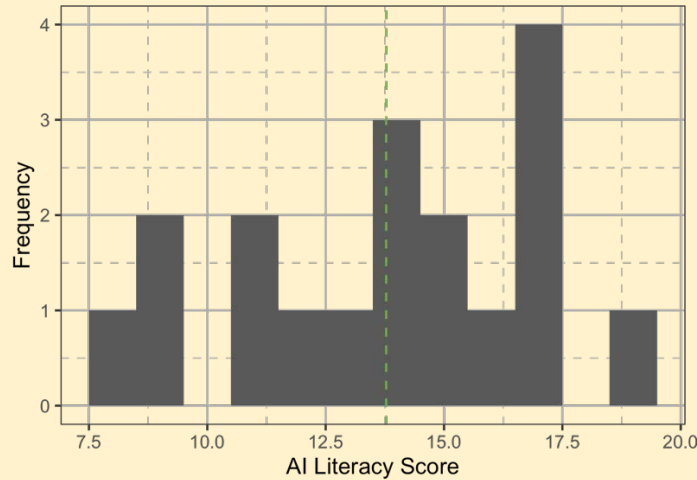


Figure 13. Histogram of AI Literacy responses from a sample of psychology undergraduate majors.

5.1.1.1.2 Requirement 1.2: Oversight

Humans ultimately must be the responsible decision maker for work at a nuclear power plant, and thus incorporating the appropriate level of oversight is critical for adoption. This aspect is addressed here in two ways. First, characterizing the work process and identifying where the LLM will be used provides a direct mapping of who in the work process will be impacted. Second, the specific decisions and actions required of personnel is captured at this step through the techniques shared in Table 2. The assimilation of this information provides added context into identifying what aspects will be important for adequate oversight – through developing an understanding of who in the work process will need oversight of the LLM and for what purpose. The degree of oversight has implications on the extent of tasks completed by the LLM. For instance, greater task oversight may be achieved if the user has a more direct role in decision-making in the automated process. Thus, a user would achieve greater oversight in a redesigned work process that integrates the LLM to support human decision-making rather than with replacing human decision-making.

5.2 Identify System Risks to Automated Process

Risk analysis should be integral to informing the preparation of the model leading up through its deployment (Salmon et al., 2024). Thus, this activity entails identifying both LLM safety and security risks associated with the automated process. Because hazards can result from deficient interactions between the LLM and end users, the use of systems thinking methods are discussed here (Salmon et al., 2022). Methods like System Theoretic Process Analysis (STPA) and the Networked Hazard Analysis and Risk Management System (Net-HARMS) offer a proactive approach to identifying potential ‘risk modes’ that are both technical and sociotechnical in nature (Dallat et al., 2018; Kovesdi et al., 2024; Leveson & Thomas, 2018; Mylius, 2025;). Though, this methodology does not mandate a specific type of risk analysis method; other risk analyses can certainly be applied.

Regardless of the methodology selected, there are two key benefits to risk analysis in the context of addressing trustworthiness. First, it provides a direct way of identifying potential risk modes associated to safety and security. A second benefit is that it provides a way to prioritize tasks with greatest risk when developing a trustworthy LLM and ensuring calibrated trust. Thus, efforts to address trustworthiness and trust can be risk-informed and allow for grading of subsequent activities in the methodology.

5.2.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

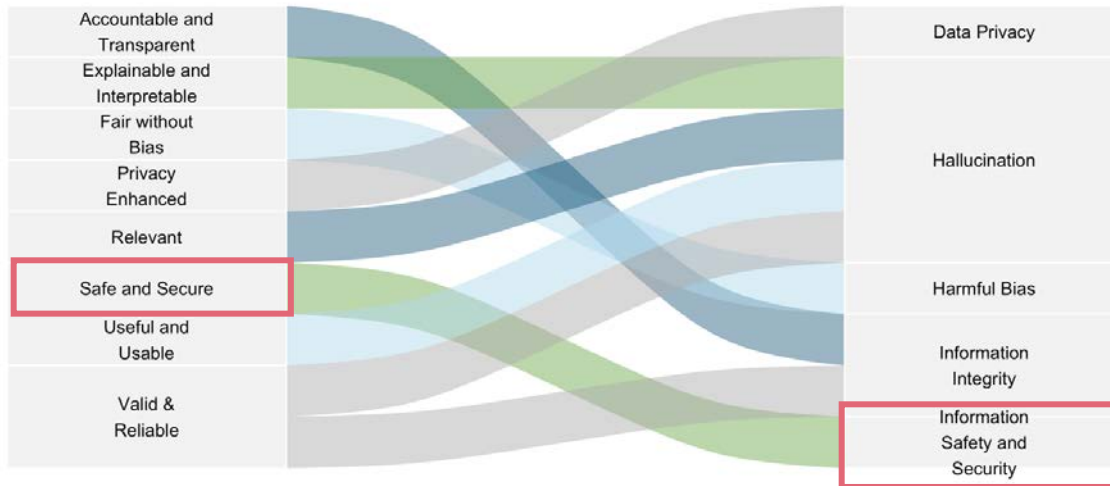


Figure 14. Cross-cutting trustworthiness issues addressed when identifying system risks.

5.2.1.1 Issue 1: Information Safety and Security

Risks associated with safety and security can have implications to downstream activities and design decisions such as infrastructure hosting, model selection, optimization, and integration. Thus, potential for hazards associated with LLM use that have consequences to safety (e.g., potential hallucinated information used in troubleshooting critical equipment resulting in plant trip) and security (e.g., potential vulnerabilities in the access control of the LLM leading to hazard of prompt injection thereby allowing access to sensitive plant data) must be identified and addressed to ensure trustworthiness. The scope of risk identification should expand beyond the LLM itself, encompassing where it interacts with data, other systems, and prospective users.

5.2.1.1.1 Requirement 1.1: Safe and Secure

Risk analysis methods, like STPA or Net-HARMS, identify potential hazards in the work process that have important implications to safety and security. Once hazards are identified, controls (i.e., guardrails) can be identified and mitigated; the identification of risks can also inform subsequent model development activities following a risk-informed graded approach. For instance, the guidance developed by the Electric Power Research Institute (EPRI) in technical report 3002004310 (2015) provides criteria for grading, based on a three-tier categorization scheme:

- *High* – Plant trip or damage to major equipment causing prolonged shutdown, death or serious injury to personnel, or significant radiation exposure to the public.
- *Medium* – Impact to plant performance, cause major equipment damage, or injury to personnel.
- *Low* – Neither impact to plant health, safety, or personnel safety (other).

The following example illustrates a simple application of Net-HARMS to evaluate the process of identifying and resolving an equipment failure at a nuclear power plant (Example 2).

Example 2: Demonstrating Data Qualification Process – Reporting and Resolution of an Equipment Deficiency

A Net-HARMS analysis was applied to a process chart developed when identifying the user tasks for Example 2. Net-HARMS identified the following potential risks (e.g., T3) and emergent risks (i.e., risks that compound from the presentation of a given risk – eT3s), shown in Figure 15. This analysis used a risk taxonomy from Dallat and colleagues (2018), shown in Table 3.

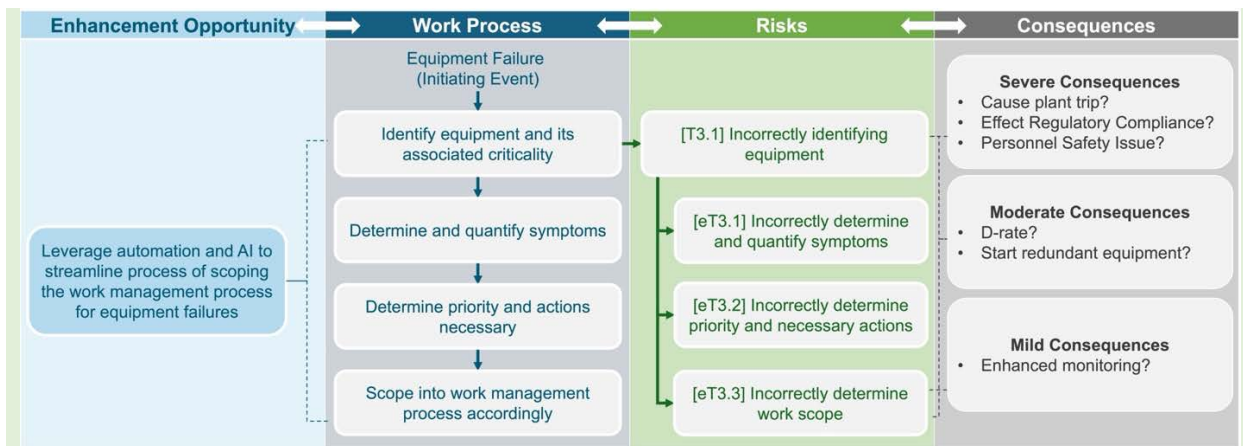


Figure 15. Net-HARMS applied to process of scoping work based on an incoming equipment failure.

The risks and emergent risks identified (the Risks column), using the Net-HARMS taxonomy, were mapped to each task from the process chart (the Work Process column). The immediate risk of incorrectly identifying equipment (T3.1) may have compounding effects that surface as emergent risks, including:

- eT3.1: Incorrectly determining and quantifying symptoms.
- eT3.2: Incorrectly determining the priority and necessary actions to remedy the equipment.
- eT3.3: Incorrectly determining the needed work scope.

The consequences of this cascade of risks can have a variety of impact depending on the specific plant equipment (the Consequences column). Severe consequences, such as those that have significant impact to the safety or economics of the plant, may include plant trips, impacts to regulatory compliance, or personnel safety issues. In this example, a conservative approach was taken where the hazard of the greatest significance was used to frame risk (potential consequence of plant trip – thus adding risk as *high*). The implications from Net-HARMS to ensure safety is not negatively impacted with the inclusion of the LLM included the following guardrails – consider robust fine-tuning, application of RAG, and robust prompt engineering techniques – to minimize the occurrence of biased, irrelevant, or invalid information related to plant equipment. Further, administrative controls including training and policy to perform an independent verification of the information provided is put in place for this specific task where LLM will be implemented.

Table 3. Net-HARMS risk mode taxonomy.

Aspect	Risk Mode
Task	T1—Task mistimed
	T2—Task omitted
	T3—Task completed inadequately
	T4—Inappropriate task performed
Communication	C1—Information not communicated
	C2—Wrong information communicated
	C3—Inadequate information communicated
	C4—Communication mistimed
Environment	E1—Adverse environmental conditions

5.3 Qualify the Integrity of Available Data

The third step is to perform an analysis that qualifies the level of technical accuracy needed from the available data to support the tasks identified and analyzed from the previous steps (Sections 5.1 and 5.2). The process of qualifying the data entails assembling a multidisciplinary team that includes SMEs (i.e., domain experts), data scientists, human factors experts, as well as other relevant stakeholders. This team should then leverage the outputs from previous steps. Where priority should be given based on risk, the relevant data sources that are used by personnel (i.e., and to be ingested by the LLM) are identified and then reviewed in terms of:

- *WHAT* the data is (including its format).
- *WHERE* the data comes from (how it is collected).

These questions can be answered by systematically reviewing the outputs that come out of the previous steps and expanding these results by identifying the relevant data sources and analyzing these sources in terms of 1) how the data was collected and 2) how it will be used. Adding this perspective to the reviewed data sources ensures that all data is not treated equally and helps to provide the context by which the data was created and captured. This allows for a systematic qualification of the data sources. An example rubric intended to be used for data qualification is presented in Table 4.

Table 4. Critical work process considerations of the data that impacts its integrity.

Internal or External Source [PICK ONE] – Internal - Information came directly from a plant source internal to the plant or utility including Condition Reports, CAP Evaluations, Corrective Actions, Work Management Schedule, Procedures, Surveillance, Technical Specifications, Modifications, Projects, Work Package, Training Records, Inventory, Technical Evaluations, Regulatory Reports, Audits, Assessments, Surveys, Instruments and Sensors. (+15) – Information came from sources external to the plant including: Regulatory documents (Licensee Event Report (LERs), Inspections, Info Notices), Institute of Nuclear Power Operations (INPO) originated (Industry Reporting and Information System [IRIS], Evaluations), EPRI, Manufacturer, Supplier, Services, Vendor, Bulletins, Research Papers, Other Utilities/Plants (+5) How It was generated [PICK ONE] – Instrument or Sensor (+15) – Human Interaction (+5) – Machine Generated – Synthesis of sources of information (+5) Technical accuracy (Unreliable Narrator) [PICK ONE] – Low – conjecture, guess (+1) – Moderate – Inference, circumstantial, support refute analysis (+5) – High – Objective Technical analysis performed (+15) Source (Originator's) perspective (Point of View) (First Person, Second Person, Third Person) [PICK ONE] – Operations (+10) – Engineering (technical) (+10) – Corporate or Generic Communication (+1) – Newsletter (+1) – Management Review Meeting (+1) – Vendor/Manufacturer (+7)

Review Level [PICK ONE] – Preparer (+5) – Preparer/Reviewers (+8) – Internal Collegial Review (+10) – External Review (+15)		
Document type (examples) [PICK ONE] – Procedure (+5) – Work Package (+5) – Condition Report (+5) – CAP Evaluation (+10) – Engineering Evaluation (+10) – Modification, Condition Report (+10) – Logbook (+7) – Computer Point or Data Recorder (+10) – Training Lesson Plan (+7)		
Potential Info Distortion at the time of creation, storage or reporting [PICK ANY] – Training and Qualifications of writer (-5) – Organizational Culture (-5) – Reporting or data capture distortion (-5) – Potential Improper synthesis of data sources (-5) – Generative output (-5)		
Origination Date (Creation Date) – At what point in the organization's history was the data or information created vs when it was digitally captured. (-1 per every > 5 years)		
TOTAL POINTS _____		
<25	26 - 60	>60
Low	Medium	High

Using the criteria in Table 4, a composite score can be created to determine the degree of technical accuracy needed from the data⁴. The score can then be used to support grading subsequent model development efforts that address trustworthiness considerations (discussed later in this report). Figure 16 presents a risk framework based on the information collected at risk analysis (y-axis) in conjunction with qualifying the data discussed here (x-axis). Where the work process contains hazards of greatest significance and where the technical accuracy of data is lowest, more robust methods are needed to ensure trustworthiness. For instance, if the use of LLM is to support personnel in decisions that have impact to the safety or economics of the plant (severe) and the technical accuracy of the data is low, the following actions

⁴ The scoring criteria was developed by a SME.

would appear in the ‘red zone.’ Thus, the following should be taken as a robust approach of ensuring trustworthiness:

- Robust data preprocessing
- Robust fine-tuning of the model
- Implementation of RAG
- Robust prompting techniques
- Iterative validation embedded in the process at each activity discussed above

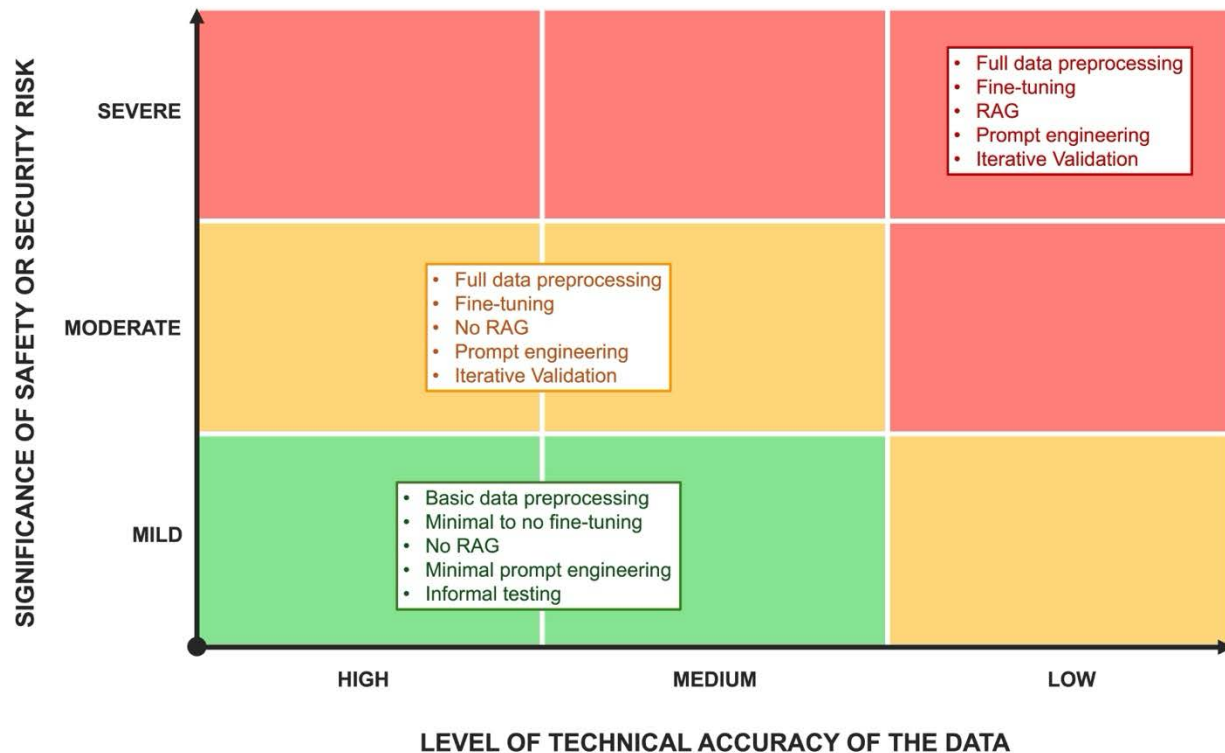


Figure 16. Risk matrix as a function of hazard significance to safety or security and technical accuracy of the data.

5.3.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

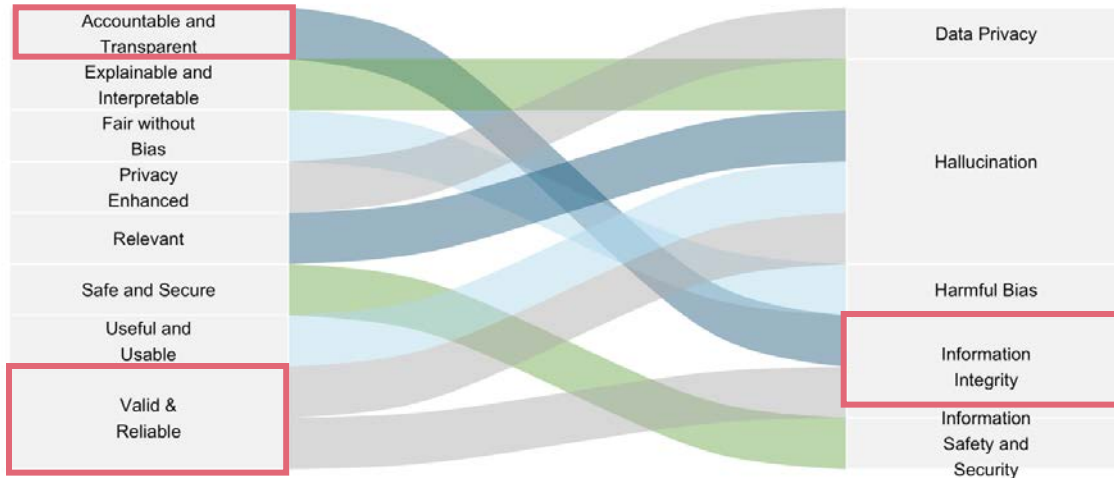


Figure 17. Cross-cutting trustworthiness issues addressed when identifying system risks.

5.3.1.1 Issue 1: Information Integrity

Information integrity issues stem from invalid and unreliable data, as well as from a lack of accountability and transparency of the data used to train the model. Without having a clear mapping of the specific data sources that will be used to train the model and establishing a clear understanding of the technical accuracy of these sources (i.e., concerning validity and reliability), key requirements pertaining to the model’s accountability and transparency, as well as validity and reliability is threatened. Specific factors that need to be considered include:

- *Data Consistency*. Are there inconsistencies in the formatting of data?
- *Availability*. Is the data readily available?
- *Completeness*. Does the data include all possible records for the expected timeframe?
- *Accuracy*. Is the information factually accurate?
- *Representativeness*. Is the data representative of the work process in the manner in which it was captured?

5.3.1.1.1 Requirement 1.1: Accountable and Transparency

This step directly addresses accountability and transparency of the model by providing an explicit mapping of the data sources used in training and the technical accuracy of this data. LLMs are ‘black box’ models by their inherent design, and thus transparency is achieved through thoughtful mapping of data sources and clear documentation of such results (Kalodanis et al., 2024). The data qualification process provides a means of systematically identifying all data sources that will be used for training and providing information of the factual correctness of this data.

5.3.1.1.2 Requirement 1.2: Valid and Reliable

Qualifying the data at this step ensures that any limitations (e.g., data inconsistencies, lack of availability, incompleteness, inaccuracies, or lack of representativeness) with the data itself is clearly understood, so proper mitigation strategies can be implemented throughout model development process (i.e., applying robust fine-tuning, RAG, and prompt engineering techniques). Example 2 is shared to highlight how the data qualification process may be undergone.

Example 2: Demonstrating Data Qualification Process – Reporting and Resolution of an Equipment Deficiency

Building on the results from the risk analysis, this step entails qualifying the data for use. The data qualifying process entails two steps discussed around the use case:

- 1) data source identification and
- 2) data source qualification.

These steps are walked through using the use case as an example.

Step 1 - Data Source Identification

Data source identification entails locating primary sources for data that will be ingested by the LLM for fine-tuning or potentially used for RAG. This information is analyzed in terms of how it will be used, as well as how the data was collected, which is critical in informing what aspects of the data are important for trustworthiness of the model. For example, the fields and information from the condition report (Figure 18) were reviewed in terms of their information integrity.

Condition Report Information

Condition Report							Unit 1
CR#	CR-U1-2025-04321	Event Date:	05/23/2025	Discovery Date:	05/27/2025	Originated Date:	5/27/2025
Title							
CR Title:	Outboard Seal Cooler Leak on 1A Turbine Driven Feedpump						
Originator Information							
Originated by:	P. Welson	Originating Dept:	US-OPS-FLD-U1				
Issue Description:	During operator rounds, it was identified that there was a water leak on the outboard bearing seal cooler of the 1A Turbine Driven Feedpump. The leak appears to be approximately 12 drops per minute. A review of the sump pump integrator revealed that there had been a notable increase in sump pump run time for the past four days. This implies that the seal cooler has been leaking for four days.						
Immediate Actions							
Immediate Actions Taken:	Quantified leakage per minute, notified Shift Manager, initiated a Work Request Notification						
Shift Manager Review							
SM Review:	1A TDFWPP is Operable. Increase field monitoring to 4x per shift. Schedule repair at earliest available opportunity. If leakage level increases, notify shift immediately and prepare for unit derate and taking FWPP off-line.						
Screening							
Classification	CAQ	Significance Level	3	Evaluation	B		
Screening Comments	FWPP is Operable, CR is closed to WR 250012543						
WR Information							
Def Tag#	A-25067	WR#	250012543	WR Title:	1A TD Feedpump Outboard Bearing Seal Cooler Leak		
Equipment Information							
EPN#	1A-TDFWPP-SC-OTBD-1	Equip Name	1A Turbine Driven Feedpump Outboard Bearing Seal Cooler				
System	FW - Feedwater	AP-913 Criticality	CC2				
Tech Spec	4.6.2 Section A1	Tech Spec Related	Yes				
FLOC#	U1-A-TDFWPP	Functional Location	U1-A-Turbine-Feedpump-Outboard-Bearing-Seal-Cooler				
Building	Turbine	Physical Location	Unit 1 Turbine Driven Feedpump Rm				
Corrective Actions							
Recommended Actions:	Perform a Basic Cause Evaluation to determine the cause of the seal 1A-TDFWPP seal cooler leak						
Owner	Plant Engineering						
Due Date	7/31/2025						

Figure 18. Example condition report.

The technical accuracy of this specific condition report was reviewed using the criteria in Table 4. This source was identified as an internal source (+15), developed through synthesis of source information (reviewing sump pump integrator; +5), consisted of objective technical data (+15), reported by operations (+10), contained multiple reviews (+15), was documented as a condition report (+5), and had no detractors. A final score of 55 was given. Further, each field was reviewed in terms of how the data was inputted into the report. Of note, several fields were comprised freeform fields (e.g., Issue Description),

which has potential implications on the necessary level of preprocessing and cleansing needed when preparing the data for an LLM.

Step 2 – Data Source Qualification

The review of each data source and individual data field provided important contextual information to determine the level of effort needed to ensure the trustworthiness of the LLM. Table 5 is an example of this review for a given condition report.

Table 5. Data source qualification example.

Source of Information	Condition Report Information
Typical Software of Source	Typically, a custom made front end usually tied to 1) EAM database (e.g., Asset Suite), 2) Systems, Applications and Products (SAP), or 3) Maximo.
Key Decisions Made	Helps to determine the scope of the work in conjunction with many other fields.
Technical Accuracy Needed of Information	Score 55 – Medium Using criteria: [5 + 5 + 15 + 10 + 15 + 5 – 0 - 0]
Typical Formats of Information	The Condition Report consists of a variety of fields, including dates, structured text, and unstructured text.
Known Trustworthiness Challenges with Source	Many fields from the Condition Report used for determining the scope of work (e.g., problem description field) consist of unstructured text. There are no controls on these fields; however, a user may rely on other fields to help clarify the issue.
Potential Issue(s) Related to Data Integrity	Freeform text fields create potential artifacts that can affect the output of an LLM, challenging the validity and reliability of the output.
Resolution to Issue(s)	– Verify text preprocessing techniques remove potential artifacts that impact technical accuracy of output. – Verify accuracy of output during data preprocessing and model fine-tuning (i.e., that the model accurately interprets the description of the problem and accurately identifies related events). – Perform human-centered evaluation to ensure that users appropriately trust the output from the model.
Candidate Automated Function to Support Use Case from Information	AI could be used to mine the information from the Condition Report to support the user in determining the scope of work needed to resolve an equipment failure.
Description of Automated Function	Use of LLM in interpreting the description and comparing related events of potential significance and priority. It will also help in identifying failure modes.

The review as seen in Table 5 contains four general types of information used to qualify the data. First, data source information is recorded in gray. This information includes general information of the source (e.g., Condition Report) and associated software that is used to manage the data source. Next, specific information regarding the decision that is made from this sample use case is inputted (in green). This decision uses the condition report, so this source was reviewed using the criteria in Table 4. A score of 55 was given, placing this source with a technical accuracy of 'Medium.' Also, in green, a summary of the nature of the content is given per review. These findings thus informs the potential concerns related to ensuring trustworthiness of the LLM, the potential issues, and suggested remedies to mitigate these issues. Finally, in yellow, there are two summary fields that share what specific AI can be incorporated to support this decisions and what functions may be needed of it.

6 MODEL SELECTION AND DATA PREPARATION

The second phase includes three primary activities:

- Selecting hosting infrastructure
- Selecting a pretrained model
- Completing data preprocessing and cleansing

These activities are discussed next.

6.1 Select Hosting Infrastructure

Before a pretrained base (i.e., foundational) model is selected, the infrastructure must be considered. The base model can either be run locally or accessed through an application programming interface (API; i.e., the cloud). Local models do not necessarily require being run on laptops; they can also exist on a server or a supercomputer environment (Figure 19).



Figure 19. Photograph of the Bitterroot supercomputer located at Idaho National Laboratory (credit <https://www.energy.gov/ne/articles/national-lab-brings-new-bitterroot-supercomputer-online>).

Local LLMs are typically pre-trained and provided with a license from a vendor (i.e., Microsoft, Meta, Mistral, etc.). On the other hand, LLMs accessed through an API are typically hosted by a vendor (i.e., OpenAI, Microsoft, Google, Anthropic, etc.). Generally, LLMs are resource intensive, requiring large amounts of energy, costly computing infrastructure, and numerous technical experts to develop and maintain. This makes using LLMs a considerable investment limiting the number of organizations participating in the evolution of LLM applications. This investment is justified because, prior to LLMs, there were very limited possibilities to understand language as a human does. The high resource demands show the need for continued innovation to increase the efficiency of LLMs. The latest OpenAI frontier model (i.e., o5) is only available in a miniature version indicating progress toward heightened efficiency. Also there has been much discussion of potentially resource efficient models. With increased efficiency comes the reduction of resources needed. The resulting reduction of resources required will remove the barriers, thereby increasing the adoption of LLMs across new applications. Cloud-based models increase the accessibility of the LLMs to organizations that could not run an LLM on their resources. Though, cloud-based models have their limitations as well, particularly in terms of data privacy considerations.

There are definite tradeoffs associated with the decision to locally host an LLM versus using a cloud-based service. The implications to this decisions can have a direct effect on data privacy considerations. A succinct decision framework is offered here to capture the high-level decisions that must be made when choosing hosting infrastructure (Figure 20).

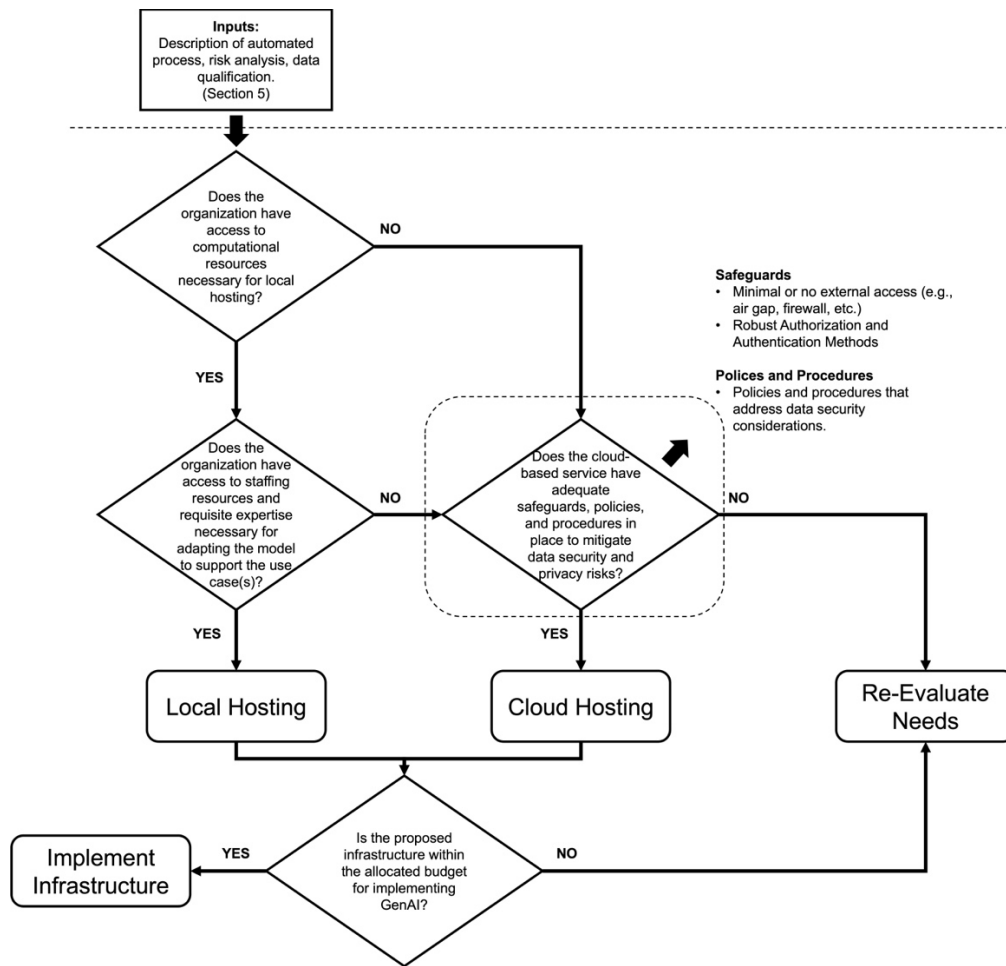


Figure 20. Simplified decision chart to determine use of local hosting versus cloud hosting.

Key inputs into determining whether to host locally or use cloud-based service include the outputs from Section 5 of this report. Cases where sensitive data (e.g., plant procedures, engineering documentation, CAP, etc.) are to be included in model optimization activities (e.g., use for fine-tuning or RAG) could thus trigger need for added protections of the data, and hence invoke the high-level decisions shown in Figure 20.

It should be noted that an organization will have more control over data governance considerations when the model is hosted locally, as opposed to relying on a third-party to host over the cloud. Thus, the decision chart starts with deciding whether or not to host locally. Thus, the first key question to address is whether an organization has access to adequate computational resources that are necessary for performing the requisite model optimization activities, discussed in Section 7 of this report. Without access to such resources, model preparation becomes impractical in terms of the time required to prepare and use the model; hence, answering ‘no’ directs questioning in determining whether a cloud-based service is possible.

Moving down the left side of Figure 20, the next question related to local hosting pertains to having adequate staff and expertise available to support the requisite model optimization activities. This element is equally important for locally hosting and managing the model. If the decision is to use a cloud-based approach, a key question is whether the third-party service has adequate safeguards, policies, and procedures in place to mitigate potential risks to data privacy; the results from risk analysis may inform this decision. The figure provides candidate considerations that should be accounted for, such as engineering controls in place for accessing data (e.g., will the data be secured through some barrier such as air gap, firewall, etc.), as well as robust authorization and authentication methods. The level of control needed will

depend on the nature of content being ingested by the model and its application. Ultimately, if the cloud-based service does not provide adequate protections, then the needs for hosting an LLM should be re-evaluated, or alternative solutions should be considered. Finally, the decision to use a particular hosting infrastructure needs to be economically aligned with the organization. Thus, the final decision point is to verify whether the proposed solution meets the requisite budget. If not, then reconsideration must be made.

6.1.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

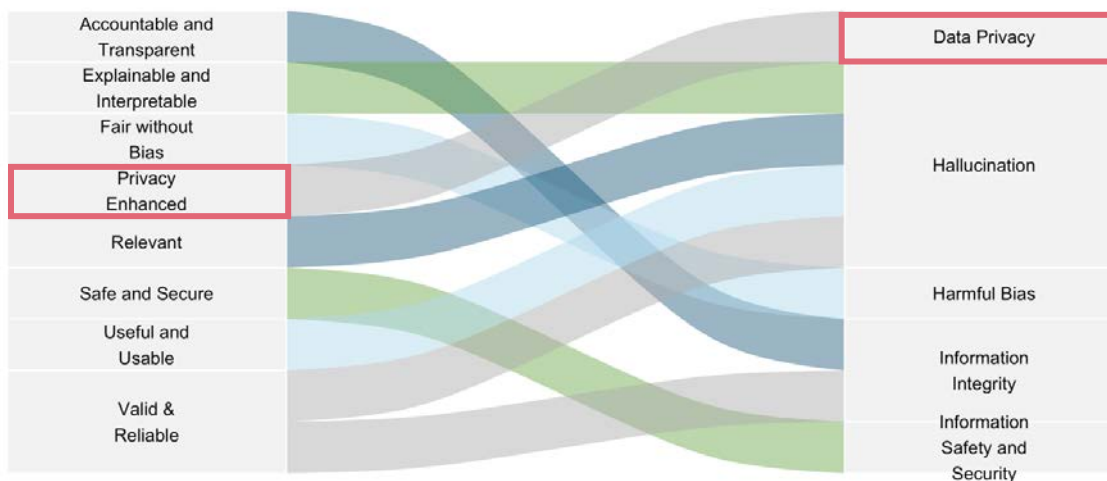


Figure 21. Cross-cutting trustworthiness issues addressed when selecting hosting infrastructure.

6.1.1.1 Issue 1: Data Privacy

Sensitive information such as commercial data and export-controlled data is common in nuclear power plants. Choosing between a cloud-based model and a local model has important implications for data privacy issues. While cloud-based models are generally superior to local models in terms of performance and reduce the large amount of resource demands posed on an organization, they are often hosted on an external server and cannot always guarantee that the data sent will not be used to train an LLM. If an LLM is trained using sensitive information, users beyond the organization who use the trained model could potentially access the data. The access occurs because LLMs memorize the data they are trained on (Fatemehsadat et al., 2022; Zhou et al., 2023). Thus, potential data privacy issues are addressed when identifying the underlying infrastructure to which the model will be hosted.

6.1.1.1.1 Requirement 1.1: Privacy Enhanced

Figure 20 maps key decisions that an organization must consider when selecting hosting infrastructure that has data privacy implications. For cases where the intended use of the LLM requires access to sensitive information, local hosting offers the most robust approach for achieving enhanced privacy. If an organization chooses a cloud-based model service, then key considerations to ensure enhanced privacy is to apply adequate safeguards including external access restrictions and robust authentication. Additional efforts may be required to anonymize data such as by omitting PII or plant-sensitive information that is not appropriate for public view (e.g., final safety analysis reports, CAP, etc.).

6.2 Select Pretrained Model

Selecting a pretrained model is a foundational step in developing and deploying LLMs. Each base model has different capabilities, so thus selecting a model based on the intended use case is important for ensuring valid and reliable output.

6.2.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

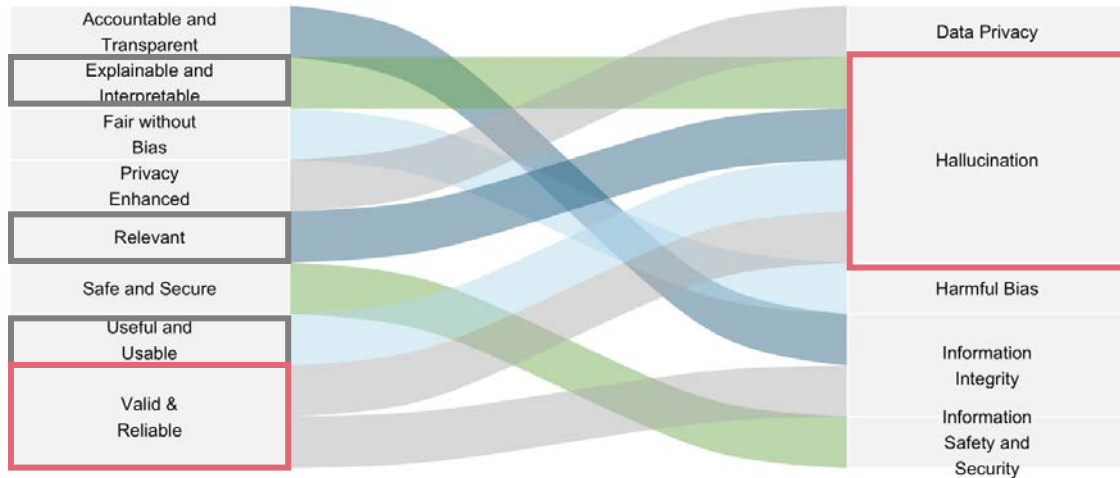


Figure 22. Cross-cutting trustworthiness issues addressed when selecting a pretrained model.

To note, explainable and interpretable, relevant, as well as useful and usable requirements cannot be addressed directly from selecting a foundational model. This is because these requirements are highly dependent on the context of use, which is not addressed until there is iterative refinement and evaluation of the model during model optimization.

6.2.1.1 Issue 1: Hallucination

One of the key threats to trustworthiness is the propensity of particular models to hallucinate. Thus, assessing a model's validity and reliability in performing tasks that are aligned with the respective use case is necessary. Also as discussed in Section 5.3, the determination of infrastructure used may bound the specific base model that can be reasonably selected.

6.2.1.1.1 Requirement 1.1: Valid and Reliable

One way to evaluate the validity and reliability of an LLM is through evaluation of specific models against established benchmarks and truth-seeking questions. Key leaderboards such as Holistic Evaluation of Language Models (HELM) provide benchmark data that can be used to compare models when completing specific truth-seeking questions relevant to general topics like reasoning, multilingual processing, or math tasks (Fourrier et al., 2024). These benchmarks provide a way to discern the capability of a given foundational model when performing tasks related to these topics. This may offer a way to evaluate a model's propensity to provide valid and reliable output based on these general task types. Example 3 is shared next, to highlight the use of various benchmark data to inform model selection.

Example 3: Developing a Fine-Tuned LLM Trained on General Nuclear Power Plant Knowledge

For instance, as illustrated in Figure 23, certain models may excel in specific types of tasks (e.g., reasoning, math, etc.) over other models. This benchmark data was used when selecting the Phi-3.5-mini-instruct model, as it excelled in reasoning tasks when compared to other models of similar size.

It should be noted that higher values across these benchmarks represent better performance, as measured by each common benchmark test. The benchmarks shown in this figure are common standardized benchmarks that can inform the selection of specific models based on general task types: reasoning, multilingual, and math. While this information is not specific to nuclear-related information, these general benchmarks can be used to inform early selection of a specific model based on their performance across general topical areas.

	Phi-4 mini-Ins	Phi-3.5-mini-Ins	Llama-3.2-3B-Ins	Mistral-3B	Qwen2.5-3B-Ins	Qwen2.5-7B-Ins	Mistral-8B-2410	Llama-3.1-8B-Ins	Llama-3.1-Tulu-3-8B	Gemma2-9B-Ins	GPT-4o-mini-2024-07-18
Popular aggregated benchmark											
Arena Hard	32.8	34.4	17	26.9	32	55.5	37.3	25.7	42.7	43.7	53.7
BigBench Hard (0-shot, CoT)	70.4	63.1	55.4	51.2	56.2	72.4	53.3	63.4	55.5	65.7	80.4
MMLU (5-shot)	67.3	65.5	61.8	60.8	65	72.6	63	68.1	65	71.3	77.2
MMLU-Pro (0-shot, CoT)	52.8	47.4	39.2	35.3	44.7	56.2	36.6	44	40.9	50.1	62.8
Reasoning											
ARC Challenge (10-shot)	83.7	84.6	76.1	80.3	82.6	90.1	82.7	83.1	79.4	89.8	93.5
BoolQ (2-shot)	81.2	77.7	71.4	79.4	65.4	80	80.5	82.8	79.3	85.7	88.7
GPQA (0-shot, CoT)	25.2	26.6	24.3	24.4	23.4	30.6	26.3	26.3	29.9	39.1	41.1
HellaSwag (5-shot)	69.1	72.2	77.2	74.6	74.6	80	73.5	72.8	80.9	87.1	88.7
OpenBookQA (10-shot)	79.2	81.2	72.6	79.8	79.3	82.6	80.2	84.8	79.8	90	90
PIQA (5-shot)	77.6	78.2	68.2	73.2	72.6	76.2	81.2	83.2	78.3	83.7	88.7
Social IQA (5-shot)	72.5	75.1	68.3	73.9	75.3	75.3	77.6	71.8	73.4	74.7	82.9
TruthfulQA (MC2) (10-shot)	66.4	65.2	59.2	62.9	64.3	69.4	63	69.2	64.1	76.6	78.2
Winogrande (5-shot)	67	72.2	53.2	59.8	63.3	71.1	63.1	64.7	65.4	74	76.9
Multilingual											
Multilingual MMLU (5-shot)	49.3	51.8	48.1	46.4	55.9	64.4	53.7	56.2	54.5	63.8	72.9
MGSM (0-shot, CoT)	63.9	49.6	44.6	44.6	53.5	64.5	56.7	56.7	58.6	75.1	81.7
Math											
GSM8K (8-shot, CoT)	88.6	76.9	75.6	80.1	80.6	88.7	81.9	82.4	84.3	84.9	91.3
MATH (0-shot, CoT)	64	49.8	46.7	41.8	61.7	60.4	41.6	47.6	46.1	51.3	70.2
Overall											
	63.5	60.5	56.2	56.9	60.1	67.9	60.2	62.3	60.9	65	75.5

Figure 23. Detailed benchmarks for common autoregressive LLMs (data adopted from Hugging Face <https://huggingface.co/microsoft/phi-4>).

6.3 Complete Data Preprocessing and Cleansing

Preprocessing refers to cleaning up data prior to training an LLM on the data. The data preprocessing phase requires multiple steps, such as acquiring the raw data, preparing the data for usage, and running multiple iterations of the model training process to validate that model output was satisfactory (Zhang et al., 2022; Raffel et al., 2020). A high-level overview of the process is summarized in the following figure.

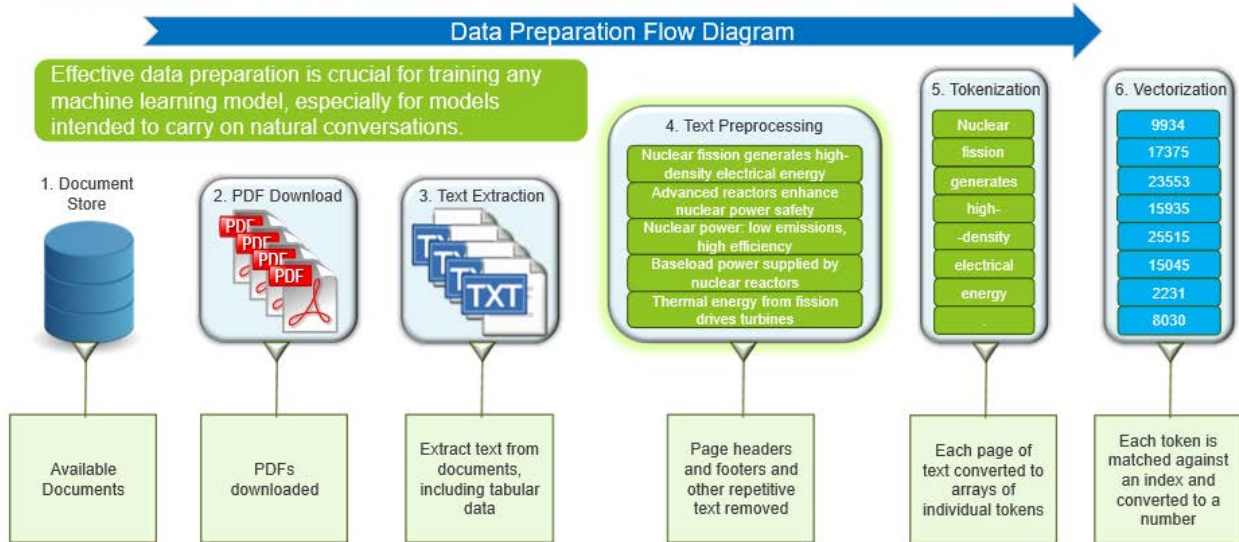


Figure 24. Process for preprocessing text information for fine-tuning an LLM.

Proceeding from left to right, the first step entails identifying the data source(s); this is achieved earlier in the methodology (i.e., Section 5.3). Next, the data from the database must be accessed such as by downloading individual data files. Because natural language data typically is stored in human readable files like PDF documents, an extraction process must then be completed (step 3). Key factors that must be considered during this step include accounting for inconsistent line breaks, malformed text, figures and tabular data, as such inconsistencies can have a direct impact on validity and reliability of the model's output. Step 4, text preprocessing, further removes artifacts from the text that can impact model validity and reliability; text preprocessing is completed through an established pipeline that systematically reviews the text in terms of semantics and grammar. Following text preprocessing, steps 5 and 6 convert every text document into a tokenized and vectorized form, amenable for usage in the model training process (Gururangan et al., 2020).

6.3.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

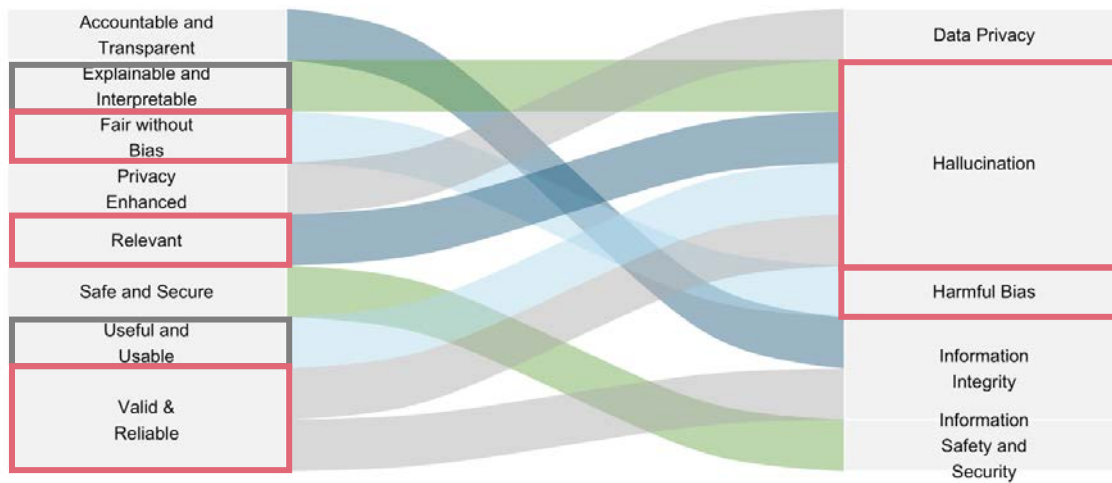


Figure 25. Cross-cutting trustworthiness issues addressed when preprocessing the data.

To note, explainability and interpretability are features of the LLM that are addressed within the context of the users' interactions with the LLM (see Section 8.2). Usefulness and usability share similar qualities.

6.3.1.1 Issue 1: Hallucination

In the context of data preprocessing, early model outputs trained on minimally processed data frequently exhibit hallucination because of unprocessed table references, fictitious section headings, and superfluous signatory blocks, particularly in response to prompts involving those data sources. Contributing factors included structural noise in the training corpus, such as duplicated email signatures, malformed references, and boilerplate formatting patterns that the model incorrectly learns as meaningful completions.

6.3.1.1.1 Requirement 1.1: Relevant

Ensuring relevant output is addressed at step 4 of data preprocessing (i.e., text preprocessing). There are two strategies that can be completed to reduce the likelihood of producing irrelevant output. These strategies include 1) filtering out repeated and low-value content and 2) incorporating deduplication and cluster-based redaction.

Text Preprocessing – Filtering of Repeated and Low-Value Content

Databases that contain a large amount of unstructured natural language text may contain content that is not germane to the intended use case that the model will support. Text such as repeated email heads, signatures, and greetings, for instance, are text that are typical seen in various documents, but do not have a central use when training the model for a specific use case. Thus, to filter this 'low-value' content, a fast noise classifier using term frequency-inverse document frequency (TF-IDF) features can be applied (Salton & Buckley, 1988; Rajaraman & Ullman, 2011), following the use of a ridge regression model (Hoerl & Kennard, 1970) that is trained on a large set of labeled examples (e.g., ~30,000) of noise versus valid text. Once trained, this classifier can achieve a high degree of accuracy in identifying irrelevant content (i.e., over 99% accuracy) to remove repetitive or irrelevant content at scale. This not only reduces irrelevant completions but also helps the model focus its capacity on domain-relevant text.

Text Preprocessing – Deduplication and Cluster-Based Redaction

Corrupted or low-value content can often appear in clusters of documents with highly similar language. For example, groups of documents containing garbled extraction artifacts or very short, non-informative/irrelevant messages can be identified and removed using an embedding-based clustering approach (Vaswani et al., 2017; Johnson et al., 2019) that allows similar documents to be grouped together, reviewed as a set, and appropriately removed from further processing.

6.3.1.1.2 Requirement 1.2: Valid and Reliable

The strategies described below focus on 1) identifying and removing sources of confusion in the training corpus, 2) improving the structural clarity of the data, and 3) providing the model with consistent patterns that encourage valid and reliable output (Weidinger et al., 2022; Bommasani et al., 2021).

Text Extraction – PDF Text Extraction and Structural Repair

Problems such as broken paragraphs, misplaced line breaks, and disrupted table formatting can teach an LLM false relationships between words and phrases. This is addressed during text extraction through the following techniques:

- Apply paragraph reconstruction heuristics (e.g., measuring the average vertical spacing between lines and using thresholds) to detect actual paragraph breaks.
- Remove breaks between lines using lowercase-join rules to address mid-sentence fragmentation.
- Exclude severely malformed or highly complex documents where the above cannot adequately address; for instance, optical character recognition can be very difficult to extract text in a way that ensures valid and reliable output.

These techniques ensure that the text fed to the model preserves logical structure and minimizes opportunities for it to learn spurious sentence or paragraph boundaries that could later lead to hallucinations.

Text Preprocessing – Removal of Problematic References

In-text references to specific figures and tables from the data can cause the LLM to incorrectly refer to tables and figures in the output. To mitigate this risk to validity and reliability, regular expressions can be used to find all explicit table and figure references and then replace them with standardized placeholders (e.g., [TABLE] and [FIGURE]). This standardized treatment prevents the model from overfitting on specific numbering schemes or inventing plausible-sounding but incorrect reference identifiers. It also keeps the representation of these elements consistent across the entire corpus (Friedl, 2006; Kukich, 1992).

Text Preprocessing – Consistent Tabular Data Handling

Inconsistent table formatting can lead the model to produce incorrect or malformed tables. To improve consistency, all tables are converted into a standardized tagged format (e.g., using <TABLE_START>, <TABLE_END>, column separators (|), and explicit <BLANK> tags for empty cells). This helps the model learn a consistent internal representation of tables, preventing it from “inventing” structural variations during generation.

6.3.1.2 Issue 2: Harmful Bias

Bias in LLMs built on specialized datasets can arise when some types of documents or ‘voices’ are more common than others. When working with nuclear power datasets, the plausibility for harmful bias may be heightened if certain licensees, reactor types, vendors, or regions appear much more often, leading the model to lean toward their particular regulatory viewpoints, vocabulary, or ways of doing things. This imbalance can be caused by repeated copies of similar documents, the use of standard form letters, or differences in how many documents each organization contributes.

6.3.1.2.1 Requirement 2.1: Fair without Bias

Fair without bias is addressed by removing content (i.e., documents, perspectives, or terminology) that is disproportionately represented in the training corpus, allowing for heterogeneity of content. The following strategies ensure fair without bias output at the data preprocessing stage by detecting and removing such imbalances before training begins.

Text Preprocessing – Corpus-Wide Deduplication

Near-duplicate detection using a combination of text edit distance (Levenshtein, 1966; used for within document comparison) and embedding-based similarity measures (Vaswani et al., 2017; Johnson et al., 2019; used for corpus-wide comparison) can be performed to remove nearly identical documents. In removing duplications, canonical retention rules should be applied to retain one representative version of each document. Collectively, this prevents the model from being “over-trained” on dominant templates or repetitive language.

Text-Preprocessing – Context-Sensitive Clustering and Outlier Removal

Some forms of bias stem from overrepresented subtopics or document categories that do not meaningfully contribute to balanced model behavior. To mitigate this, context-sensitive clustering can be applied. This entails reducing dimensionality of sparse embeddings using Truncated Singular Value Decomposition (TruncatedSVD; Halko et al., 2011). Semantically similar clusters and outliers can then be identified using Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) and the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithms (McInnes et al., 2017) for removal. By removing or down-weighting such clusters, bias in the dataset is mitigated by maintaining a more heterogeneous distribution of content.

Text-Preprocessing – Balancing Representational Diversity

Even when documents are unique, some categories can still appear more frequent than others. The preprocessing pipeline should incorporate sampling adjustments to ensure that no single category dominates the training set. Key steps to removing potential sources of bias resulting from homogeneous categories (i.e., uneven distribution) of data entail identifying overrepresented topics by reviewing cluster label distributions derived from TF-IDF keyword analysis (Salton & Buckley, 1988) and filtering documents in categories that noticeably exceed the mean representation across clusters. These adjustments help align the dataset’s topic distribution more closely with the intended balance of perspectives, ultimately leading to a fair representation of the data without harmful bias.

Text-Preprocessing – Removal of Structural Bias Sources

Structural artifacts—such as repeated headers, footers, and routing metadata—can bias the model toward memorizing irrelevant elements instead of focusing on substantive content. The noise classifier described in the hallucination-focused section is also effective here (in removing bias) – i.e., by removing artifacts consistently, the classifier reduces structural patterns that could otherwise skew the model’s outputs toward certain document formats or institutional voices.

7 MODEL OPTIMIZATION

The third phase includes two primary activities:

- Fine-tuning the model
- Implementing RAG

These activities are discussed next.

7.1 Fine-Tuning

Fine-tuning an LLM refers to the process of continuing the training of a pretrained model on a domain-specific corpus so that it internalizes patterns, terminology, and structural nuances unique to that domain (Gururangan et al., 2020; Raffel et al., 2020). Whereas foundational models like Phi-3.5 are trained on vast and heterogeneous datasets intended to maximize generalization, fine-tuning enables specialization—allowing the model to adapt its knowledge representation to the subtleties of a new, targeted data distribution. This process is typically supervised and consists of feeding the model input sequences and requiring it to predict the next token in each sequence. Over time, the model’s internal weights are adjusted to minimize prediction error, effectively anchoring its outputs to the style and semantics of the fine-tuning corpus. A high level overview of this process is illustrated in Figure 26, which is meant as a continuation in context from data preprocessing and cleansing (from Figure 24). That is, whereas Figure 24 addressed preprocessing of the data, Figure 26 focuses on how the preprocessed data is used to fine-tune the model.

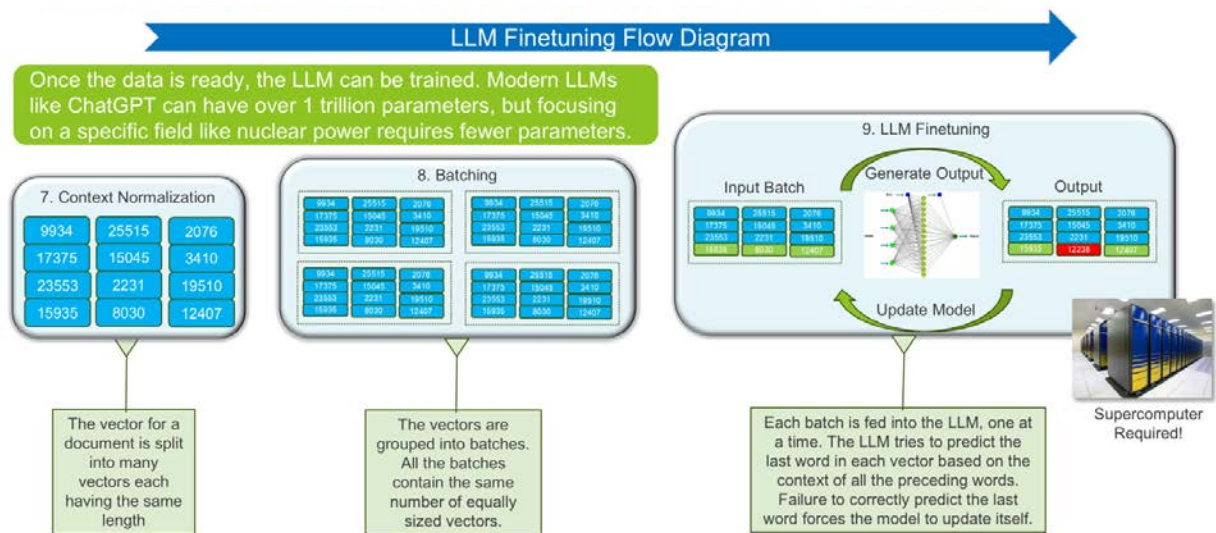


Figure 26. High level overview of the LLM fine-tuning process.

Fine-tuning in the context of LLMs encompasses a range of different techniques (Chung et al., 2022; Hu et al., 2021). Table 6 summarizes three common fine-tuning techniques: full fine-tuning, Low-Rank Adaptation (LoRA), and instruction tuning. Full fine-tuning entails updating all trainable parameters of the base LLM. This is generally done using a large, domain-specific corpus; full fine-tuning is a computationally intensive process compared to lighter-weight approaches that update only a subset of model weights. Full fine-tuning may be done when adapting a general model to a specific domain, such as nuclear power. LoRA is significantly more efficient and is often used for domain-specific adaptation or instruction tuning without the need to retrain the entire model (Hu et al., 2021). LoRA may be used when adapting the model to specific concepts used at an organization (e.g., adapting a fully fine-tuned nuclear model to a specific utility’s conventions for specific tasks). Finally, instruction tuning is a type of fine-tuning that focuses on training the model to follow specific instructions, as the name implies. Instruction

tuning is used to adapt the model for following instructions to complete specific types of tasks (e.g., help populate 50.59 forms or condition reports).

Table 6. A short list of LLM fine-tuning techniques with relevance to this research.

Approach	Description	Parameters Updated	Common Uses
<i>Full Fine-Tuning</i>	All model weights are updated to adapt to new data	All	Full-domain adaptation, high-resource settings
<i>LoRA (Low-Rank Adaptation)</i>	Adds low-rank matrices to attention weights	Low-rank matrices	Efficient fine-tuning, scalable setups
<i>Instruction Tuning</i>	Trains model on broad, task-diverse instructional data	Usually all	General-purpose instruction following

7.1.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

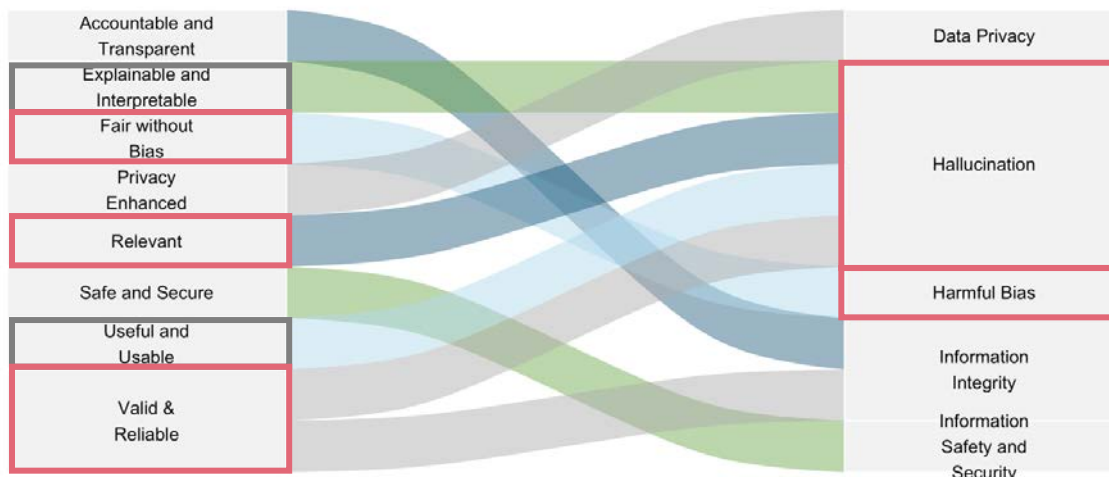


Figure 27. Cross-cutting trustworthiness issues addressed when fine-tuning the model.

To note, explainability and interpretability are features of the LLM that are addressed within the context of the users' interactions with the LLM (see Section 8.2). Usefulness and usability share similar qualities.

7.1.1.1 Issue 1: Hallucination

Fine-tuning ultimately reduces hallucinations by aligning model behavior with language patterns grounded in verified, domain-relevant sources. Without fine-tuning, base models often generate fluent but inaccurate responses when prompted with specialized or technical questions outside their original training domain (Ji et al., 2023). Though, it should be noted that there are still notable risk for hallucinations with fine-tuning. A major threat pertains to reduced generalization whereby the model becomes 'brittle' in terms of how users can access information from it. This can result from the model learning from spurious patterns in the data. Thus, key considerations here pertain to applying appropriate strategies that omit such data patterns to ensure that the model is 'adequately adaptable,' when providing the user with information from its acquired knowledge base.

7.1.1.1.1 Requirement 1.1: Relevant

Key strategies that can be used at fine-tuning to ensure relevant output include 1) deterministic orchestration and 2) chunking with loss masking. These strategies are discussed next in terms of how they address relevance.

Deterministic Orchestration

Training large models across multiple GPUs and nodes needs to be organized with reproducibility in mind. Deterministic orchestration sets forth a predefined sequences of reproducible steps for fine-tuning. By leaving random seeds fixed, using a deterministic job manager, and regular checkpointing, mid-training performance audits can be performed, and training jobs can be restarted from an existing checkpoint when necessary. These steps are essential for effectively identifying and managing hallucinations, by 1) providing a structured approach that ensures the output stays relevant and contextually appropriate based on the prompt and 2) allowing for allow for rapid pivots after minor data adjustments with minimal disruption.

Chunking with Loss Masking

To prepare training inputs, documents can be divided into fixed-length segments with careful padding strategies; this is considered chunking with loss masking. By dividing large documents into smaller, semantically meaningful chunks, LLMs can focus on the information most relevant to the prompt and ignore less relevant information. Thus, chunking with loss masking improves the contextual relevance of the response produced by the LLM and reduces the likelihood of including irrelevant information.

7.1.1.1.2 Requirement 1.2: Valid and Reliable

Key strategies that can be used at fine-tuning to ensure valid and reliable output include 1) semantic spot-checking, 2) causal language modeling (CLM) objective, and 3) loss monitoring and early anomaly detection. These strategies are discussed next in terms of how they address validity and reliability.

Semantic Spot-Checking

Documents can be selectively included through the use of a semantic spot-checking filter. Semantic spot-checking filters work by comparing the meaning of text output produced by the LLM to a reliable curated dictionary. A primary goal of the semantic spot-checking filter is to ensure that the produced output is factually accurate and consistent with the provided context by excluding texts with little coherent language or insufficient vocabulary coverage, which might otherwise introduce noise into the token stream.

CLM Objective

The CLM objective trains the model to predict the next token based on previous context (i.e., preceding tokens). The CLM objective thus reduces the chance of fabricated references and ensures that factual and stylistic patterns from the training data are reproduced consistently.

Loss Monitoring and Early Anomaly Detection

Loss monitoring and early anomaly detection identifies unusual patterns in the data that might result from corrupt, erroneous, or incomplete data. This enables early removal of such context that threaten validity and reliability before the model learns from such peculiarities.

7.1.1.2 Issue 2: Harmful Bias

Bias can arise in two distinct ways: 1) it can be transferred from the base model or 2) it can be introduced during fine-tuning. Transferred bias originates from the pretraining phase, during which the base model is built on large datasets that may contain imbalances or overrepresent certain perspectives, styles, or contexts. If left unaddressed, such inherited biases may persist in the fine-tuned model. Bias introduced during fine-tuning can occur if the adaptation dataset overrepresents certain viewpoints, processes, or conventions, causing these to be reflected disproportionately in outputs.

7.1.1.2.1 Requirement 2.1: Fair without Bias

A key approach for ensuring the output at fine-tuning is fair without bias entails stable optimization and architectural safeguards.

Stable Optimization and Architectural Safeguards

Stable optimization and architectural safeguards prevent the model from overfitting on biased patterns. This is achieved by reducing instability in gradient updates that can skew the output of the model. Through this intervention, the risk of producing outputs that reflect only dominant precedents is reduced, and the model is better positioned to represent the full diversity of perspectives within the target domain (Weidinger et al., 2022; Bender & Friedman, 2018).

7.2 Retrieval-Augmented Generation

RAG is a technique that enhances LLMs by integrating external knowledge into the response generation process. Instead of relying solely on pretrained knowledge, RAG enables the LLM to access and incorporate relevant information from external sources, which allows for grounding the outputs in explicit factual content. RAG can be implemented through several architectures and retrieval strategies. The two primary types include:

- 1) *in-context RAG* – which appends retrieved documents to the prompt without retraining the model, and
- 2) *integrated RAG* – which jointly trains the retriever and generator for better alignment and accuracy.

There are a variety of different retrieval strategies used in RAG. Sparse methods, like Best Matching 25 (BM25), rely on keyword matching. Dense methods, like Dense Passage Retrieval (DPR) and sentence transformers, use semantic embeddings. Hybrid methods combine both sparse and dense approaches. Finally, recursive methods iteratively gather context through document structure or metadata. There are also emerging techniques like GraphRAG (Han et al., 2025) and agentic RAG, which further expands the capabilities of RAG. For instance, GraphRAG incorporates knowledge graphs to structure RAG data, allowing for richer associations of concepts from the data (i.e., using graphs to represent relations in the data). Agentic RAG also improves relevance, as well as validity and reliability by enabling multi-hop reasoning, dynamic retrieval, and tool-based interaction.

It should be noted that not all LLMs may need to use RAG. The decision to implement RAG depends on the tasks that the LLM is supporting, as well as the nature of the data itself. RAG may be used for cases where the data is dynamic (updated frequently); this offers a practical approach to data management rather than continuously fine-tuning a model. A final point is that one notable limitation to RAG is that the retrieval of information is more computationally intensive when compared to LLMs that do not include RAG. This degree of resources needed for using RAG significantly increase with larger RAG datasets, thus lending RAG to be practically infeasible in some cases.

7.2.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

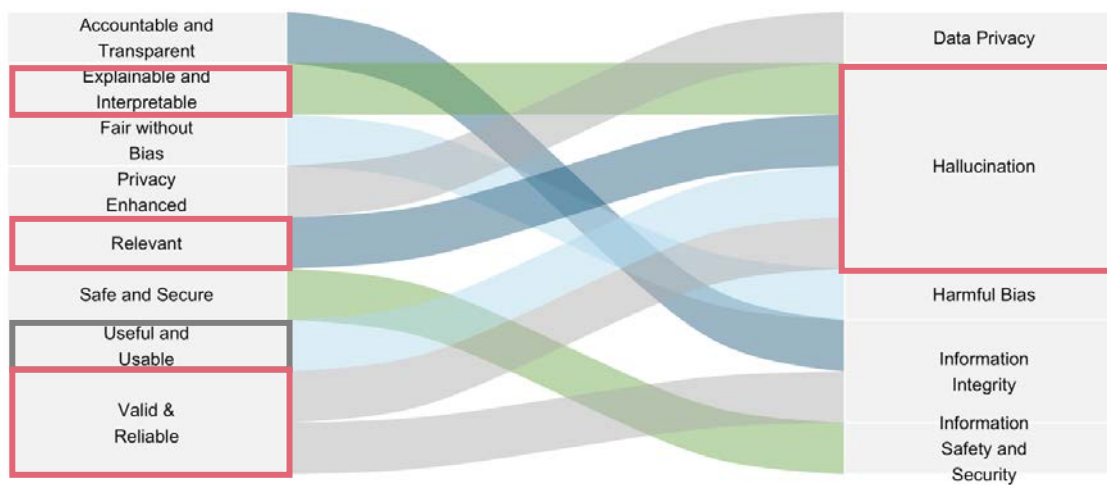


Figure 28. Cross-cutting trustworthiness issues addressed with using RAG.

To note, usefulness and usability are features of the LLM that are addressed within the context of the users' interactions with the LLM (see Section 8.2). Usefulness and usability of RAG can therefore be addressed by applying evaluation methods discussed in Section 8.2.

7.2.1.1 Issue 1: Hallucination

The direct benefit of RAG is to significantly improve the model's validity and reliability, relevance, as well as explainability and interpretability, thus reducing the risk of hallucination. However, while RAG was designed to reduce the propensity of hallucinations, it is not immune. Faithfulness hallucinations are the greatest threat for RAG. These hallucinations persist when the retriever omits the exact fact that needs retrieved. The result can be a response that appears convincing but is either incorrect or irrelevant.

7.2.1.1.1 Requirement 1.1: Explainable and Interpretable

The inherent capabilities of RAG in using explicit knowledge allows it to have greater explainability and interpretability when compared to models that do not use RAG. One promising approach is the application of GraphRAG, which uses knowledge graphs to represent information. A knowledge graph structures the data as a graph database to which data entities (graph nodes) are interconnected as representing relations (graph edges) between concepts. This explicit structuring of the data provides a direct means to the user in identifying the patterns of relationships that the model uses, as indicative to the knowledge graph structure, providing greater explainability and interpretability.

7.2.1.1.2 Requirement 1.2: Relevant

To address faithfulness hallucinations, Factual precision in Atomicity Score (FactSCORE; Min et al., 2023; Walker et al., 2024) evaluations can be added to evaluate the confidence in each atomic fact based on the prompt and relevant context from the database. The FactSCORE provides a means to present to the user the level of confidence the LLM has for a given response by evaluating its atomic facts in relation to the prompt and the atomic facts in the RAG database.

Example 1: Implementing RAG into an LLM for a Predictive Maintenance Tool

Figure 29 presents an example of how the FActSCOREs may be used (i.e., using the predictive maintenance chatbot). The response produced by the LLM is color coded as a function of confidence that the LLM has for each atomic fact. The colors used in this example show that if the atomic fact is greater than 80% agreement to the reference content, it will appear green. Agreement between 40% to 80% appear yellow, and anything below 40% appears red. This approach provides a means of communicating to the user its confidence in the response the LLM produces, potentially highlighting what content may most relevant based on the prompt and external dataset.

User: How does the primary loop in a Pressurized Water Reactor (PWR) work?

Bot: The primary loop in a Pressurized Water Reactor (PWR) is a closed circuit that transports the heat generated by the nuclear fuel to the steam generators. Here's a simplified explanation of how it works:

1. **Core**: The nuclear fuel is located in the core, where the nuclear fission process produces heat.
2. **Primary Coolant**: Pressurized water, called the primary coolant, is continuously pumped through the core by the reactor coolant pumps (RCPs), also known as circulators. In the Salem NPP, there are six RCPs per unit.
3. **Heat Exchange**: As the water passes through the core, it absorbs heat and becomes hot, high-pressure water.
4. **Steam Generators**: The hot water then flows into the steam generators. Here, the heat is transferred through a heat exchanger to the secondary loop, which turns the secondary coolant (feedwater) into steam.
5. **Cool Down**: The cooled water from the steam generators then flows back to the reactor core, where the process repeats. This cycle continues as long as the reactor is operating, maintaining a consistent temperature and pressure in the primary loop. The primary loop is isolated from the outside environment and the secondary loop to ensure safety and efficiency.

Figure 29. Example application of FActSCORE confidence indication on LLM response.

7.2.1.1.3 Requirement 1.3: Valid and Reliable

The following approaches ensure validity and reliability with RAG.

Apply a Multi-Method RAG Framework

A multi-method RAG framework that leverages an agentic and hybrid retrieval engine to dynamically orchestrate between different RAG methods (e.g., sparse, dense, or graph-based), based on the user's prompt, may ensure that valid and reliable content is retrieved. Based on benchmarking the performance of various RAG methods when competing various task types, Table 7 summarizes top performing RAG methods based on specific types of task. Such multi-method framework leverages the capabilities of each RAG method and applies the appropriate approach based on the type of task being performed.

Table 7. Summary of scenarios and corresponding best RAG approaches.

RAG Use	Best Approach	Rationale
<i>Exact numeric/facts</i>	BM25 or Hybrid	Sparse term matching reliably captures precise values and exact facts.
<i>Semantic paraphrase retrieval</i>	Vector or Hybrid	Embedding surface related text even with different wording.
<i>Mixed keyword and semantic queries</i>	Hybrid	Weighted combination balances term precision and semantic similarities.
<i>Multi-hop/entity-chain questions</i>	GraphRAG	Graph-based traversal links entity nodes across documents, enabling reliable multi-step reasoning.
<i>Document-level Question and Answer</i>	MetadataRAG	Metadata indexing guides question and answer to the right document, improving context selections for focused questions.
<i>Safe abstention on unsupported queries</i>	GraphRAG or metadata-RAG	Strictly abstains if no relations/nodes exist.

Iteratively Benchmark RAG Performance

Finally, to verify that RAG-based models produce valid and reliable results, benchmarking performance should be completed as part of iterative refinement. The use of an evaluation dataset, such as a multiple-choice question and answer set, should be maintained for testing the validity and reliability of different RAG approaches. Appendix A provides a list of 55 nuclear-related items that were used to evaluate the validity and reliability of several RAG models. These questions and answers provide a direct means to evaluate accuracy of the different models, allowing for an effective way to benchmark validity and reliability of RAG.

8 MODEL INTEGRATION

The fourth phase includes two primary activities including:

- Engineering the prompt
- Evaluating the model and human-AI interaction with the model

These activities are discussed next.

8.1 Engineer the Prompt

Prompt engineering is an emerging area in AI that entails the construction of prompts (i.e., instructions) that guide an LLM to a desired output (Geroimenko, 2025). In essence, prompt engineering allows intersection between the human and the LLM via communication in the form of natural language. At the time of this report, prompt engineering is very much an art as it is a science. That is, while prompt engineering is systematic in nature, there still remains an element of ‘craft’ skill in developing well engineered prompts, which requires experimentation and iteration. Nonetheless, there are many types of prompt engineering techniques (Geroimenko, 2025; Diamant, 2025). For example, Geroimenko (2025) cites 44 different techniques. Some common approaches, as described in Geroimenko (2025) and Diamant (2025), are summarized in Table 8.

Table 8. Common types of prompt engineering techniques.

Technique	Summary
<i>Zero-shot prompting</i>	The LLM is provided a prompt without prior examples or demonstrations. This technique can handle a broad range of tasks but relies on the clarity and specificity of the user’s prompt.
<i>Few-shot prompting</i>	The LLM is provided a few examples of the desired behavior in the prompt itself with the objective to guide the LLM’s responses by demonstrating the task or its format through examples.
<i>Chain-of-thought (CoT) prompting</i>	The LLM is encouraged to break down a task into smaller steps, like how humans problem solve. This approach provides a means of adding greater explainability to the model’s reasoning process.
<i>Constraint-based prompting</i>	The LLM is provided a set of constraints within the prompt to ensure that its response is focused, relevant, and contextually aligned with the intent of the prompt.
<i>Role-based prompting</i>	The LLM is assigned a specific role to guide its responses (e.g., “You are a reactor designer. Please provide ...”). This is distinct from persona-based prompting where personality traits are given.
<i>Persona-based prompting</i>	The LLM is given a specific persona to embody when interacting with a user (e.g., “You are a seasoned reactor designer who has a firm understanding in thermodynamics and has a strong questioning attitude. Please provide ...”).

The selection of a specific technique, such as those listed in Table 8, may be based on the specific purpose of the interaction with the LLM. For example, if a user is exploring the capabilities of the LLM, zero-shot prompting may be most practical. For more complex interactions, such as using an LLM for purposes of troubleshooting and diagnosis, an approach like CoT prompting may be better suited, as this technique provides greater granularity into the reasoning used by the LLM in formulating its response. It is also worth pointing out that selecting a specific technique is likely not “one-size-fits-all.”

Interestingly, Wang and colleagues (2024) examined the performance of a combination of nine different LLMs and four different prompt engineering techniques, applied to the healthcare domain. The authors found that the consistency in performance when using each of the different prompt engineering techniques varied significantly as applied to each of the nine LLMs. Their findings may suggest that the right prompt engineering technique likely requires some trial-and-error to achieve optimal performance for a given problem space. Thus, rather than prescribing a specific set of techniques, the following set of general principles (from Geroimenko, 2025) in Table 9 may be considered to ensure a trustworthy outcome.

Table 9. General prompt engineering principles for trustworthy LLM results.

<i>Principle 1</i>	The prompt should be clear and precise.
<i>Principle 2</i>	The prompt should be concise and use straightforward language.
<i>Principle 3</i>	The prompt should be specific but avoid over-specifying.
<i>Principle 4</i>	The prompt should contain contextual information, including examples where possible.
<i>Principle 5</i>	The prompt should employ action verbs.
<i>Principle 6</i>	The prompt should include domain knowledge.
<i>Principle 7</i>	The prompt should ask direct questions without using leading questions.
<i>Principle 8</i>	Prompt templates should be used for routine tasks.
<i>Principle 9</i>	Prompt components should be separated so that the LLM can better understand and execute its request.
<i>Principle 10</i>	Prompts should use an iterative approach, starting with more simple prompts and progressively adding layers of detail.

8.1.1 Relevant Cross-Cutting Trustworthiness Issues

The following trustworthiness considerations are relevant to this activity.

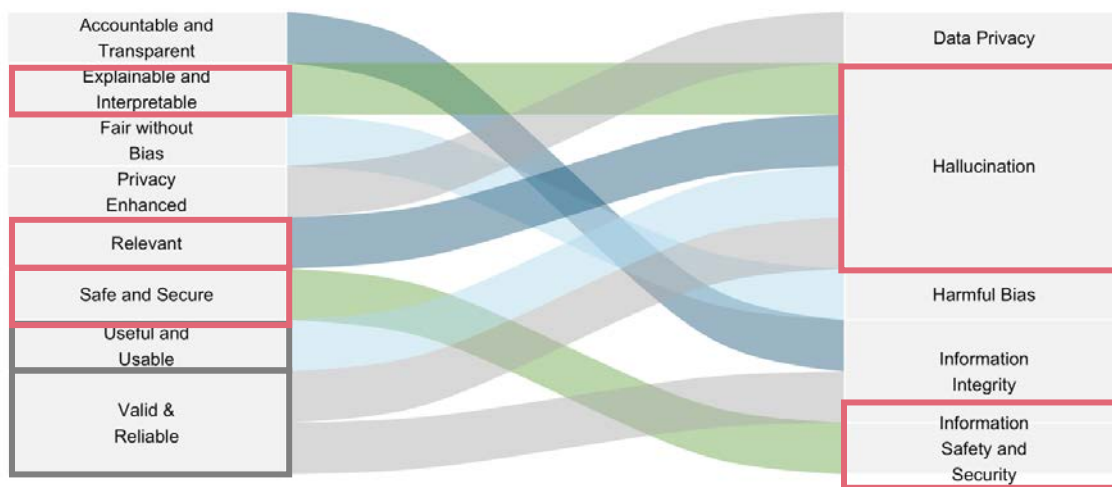


Figure 30. Cross-cutting trustworthiness issues addressed during prompt engineering.

To note, valid and reliable is not directly addressed during prompt engineering because at this stage, the model’s internal weights are not changed. Therefore, prompt engineering does not have an effect on the knowledge base of the model itself. Further, usefulness and usability are features of the LLM that are addressed within the context of evaluation (see Section 8.2).

8.1.1.1 Issue 1: Hallucination

Prompt engineering allows the human user to communicate to the LLM via natural language. Like human-to-human communication, human-to-LLM communication via prompt engineering can be negatively impacted by ambiguity in a natural language request from the user. The potential consequences of ambiguous prompts can result in hallucinations, such as those that relate to the relevance of an LLM’s response respective to the user’s original prompt (Haung et al., 2025).

8.1.1.1.1 Requirement 1.1: Explainable and Interpretable

The ability for an LLM to produce explainable and interpretable output is a key requirements for addressing negative implications of hallucinations and is relevant to prompt engineering. Depending on the complexity of the task that LLM must perform, a greater degree of explainability and interpretability may be necessary to allow users to readily identify incorrect logic used by an LLM, resulting in hallucination. Table 10 provides a summary of the different prompt engineering techniques first shared in Table 8 in terms of their degree of explainability and interpretability. Example uses are provided.

Table 10. Degree of explainability as a function of different prompt engineering techniques.

Technique	Degree of Explainability and Interpretability		Potential Uses
<i>Zero-Shot</i>	Low–Medium	There are no reasoning steps provided.	Simple question & answer tasks.
<i>Few-Shot</i>	Medium	Examples in the prompt show how the model is expected to respond, making its behavior somewhat predictable.	General content summarization tasks.
<i>CoT</i>	High	The model produces explicit reasoning steps, making its decision-making process more transparent.	In-depth reasoning tasks – troubleshooting, diagnosis, or root cause analysis.
<i>Constraint-Based</i>	Medium–High	Clear rules (formatting, logic constraints) make outputs more interpretable, but internal reasoning is not explicitly provided (like CoT).	Content summarization tasks that entail following specific formats.
<i>Role-Based</i>	Medium	Specifying the role gives context for the type of reasoning expected, but not the reasoning itself.	Inspection and summarization tasks following specific plant personnel roles.
<i>Persona-Based</i>	Low–Medium	Adds style and tone context, which may obscure the rationale rather than clarify it.	Inspection and summarization tasks following a specific persona.

Determining whether a specific prompting technique provides sufficient explainability and interpretability ultimately requires feedback from actual users. Thus, human-centered evaluation discussed

in Section 8.2 should be applied to validate whether the prompt provided an appropriate level of explainability and interpretability.

8.1.1.1.2 Requirement 1.2: Relevant

Prompt engineering also has an important role in generating relevant output from an LLM. The guidelines for prompt engineering presented in Table 9 has a direct application to ensuring that the output produced by the LLM is relevant to the original intent. The following content discusses how these guidelines can be applied to ensure relevant output. Each guideline from Table 9 is discussed in terms of how it can be addressed with prompt engineering.

Application of General Prompt Engineering Principles

Each principle is discussed in terms of how it addresses relevance.

Principle 1 – Is the Prompt Clear and Precise?

Determining whether a prompt is clear and precise can significantly benefit having a domain expert review the prompts and resulting output. While such review can be qualitative in nature, it will likely capture nuances in the prompt as related to the specific context it applies to, that automated measures simply cannot easily achieve. With that said, there are potential supplementary approaches that may support in determining whether the prompt is clear and precise, particularly if compared to alternative prompts through experimentation and iteration. For instance, Viveros-Muñoz and colleagues (2025) evaluated how linguistic complexity, grammatical variations, and informal language impacts the output quality produced by LLMs. Their findings showed that verbal moods and complex sentence structures can significantly affect response quality.

Assessment of complex sentence structures may account for several factors including lexical diversity (Engber, 1995), lexical sophistication (Juhasz, 2008), and lexical density (Lu, 2011). Automated measures that address these aspects may be used to support evaluation of prompts against these considerations and then be used to inform an assessment of a given prompt to the criteria described in this work. Thus, precision of a prompt can be assessed by lexical diversity, sophistication, and density. Diversity may be assessed by identifying the total number of unique words and normalizing by the length of text (Engbar, 1995).

Sophistication of language used in a prompt may be evaluated by various metrics that calculate the number of syllables in a word for a given prompt (Juhasz, 2008); this is discussed further for measures of readability later in this section. Finally, diversity can be measured by accounting for the diversity (i.e., variation) of lexical items such as nouns, verbs, adjectives, and adverbs. In the latter, we propose using parts-of-speech (PoS) tagging, an area in NLP, as one potential approach that may be used in evaluating whether a prompt is clear and precise, as a proxy measure of lexical diversity. The following figures illustrate the use of PoS tagging for prompt assessment.

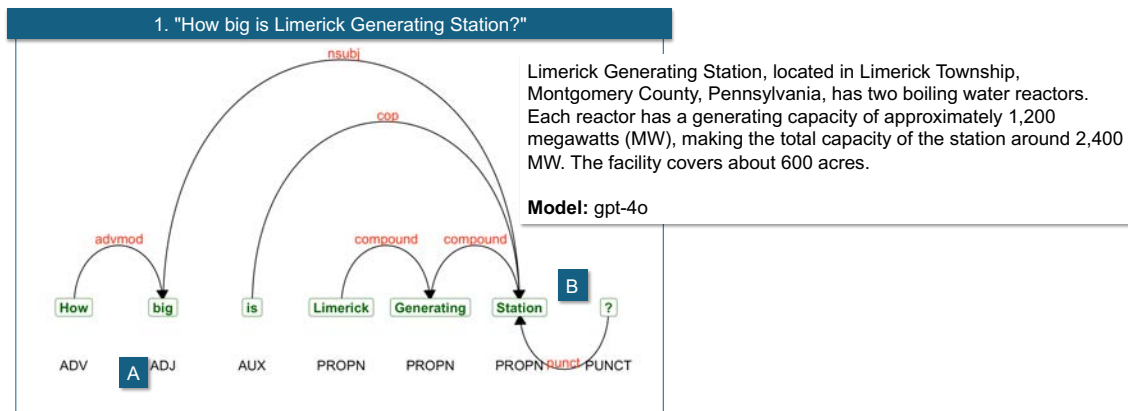


Figure 31. PoS tagging for prompt #1 (a poorly crafted prompt).

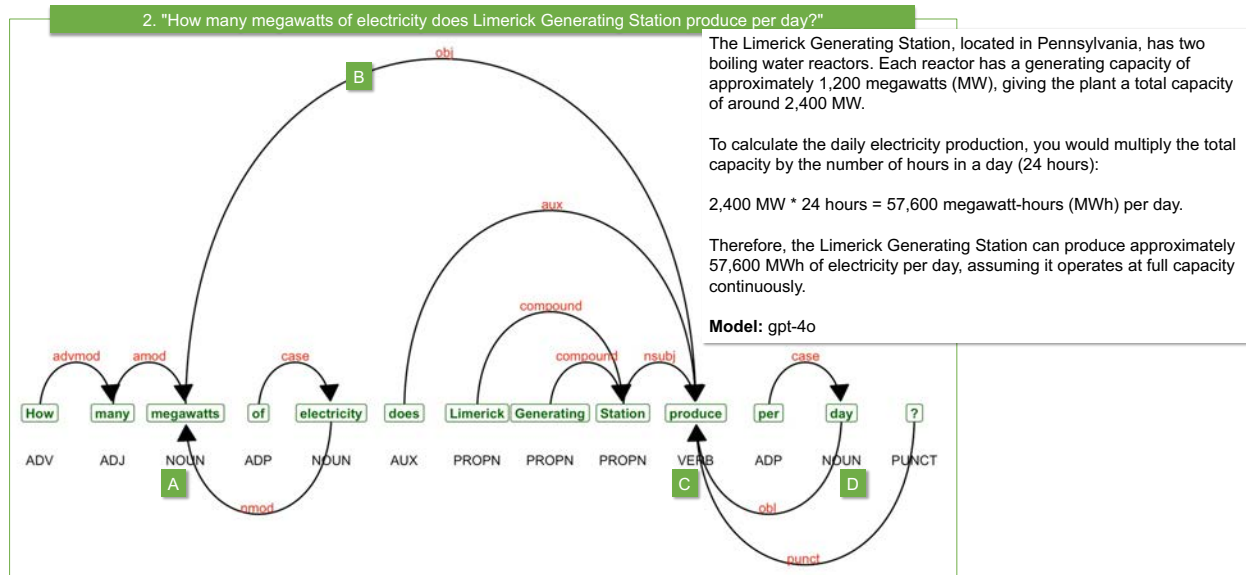


Figure 32. PoS tagging for prompt #2 (an improved prompt).

At face value, prompt #1 in appears to be more precise than prompt #2 in, based on word count. However, upon closer inspection, Figure 31 actually illustrates a poorly constructed prompt (prompt #1) while Figure 32 illustrates an improved prompt (prompt #2). From the perspective of clarity and precision of the language used, prompt #1 uses ambiguous word choices. For instance, the prompt uses an auxiliary verb without a main verb as shown in Figure 31(A). Moreover, the word choice of ‘big’ can take many different meanings without any added context. Here, ‘big’ may refer to the physical size of the nuclear power plant facility, the size of the organization running the facility, its electricity that is produced, or other meanings. Hence, such prompt may create very different results than what was intended by the user.

For prompt #2, the question is presented with added context, which improves the clarity and precision of the prompt. For instance, the object of the question is explicitly defined as ‘megawatts’ (Figure 32 – [A] & [B]). There is a main verb (‘produce’) as seen in (C). Further, the main verb (‘produce’) is further clarified by the unit of time (‘day’) as seen in (D). These enhancements provide added clarity and precision that avoids potential ambiguity.

Principles 2 & 3 – Does the Prompt Use Concise/ Straightforward Language, Avoiding Over-Specifying?

Similar to determining whether a prompt is clear and precise, determining whether a prompt uses concise and straightforward language can benefit from having a domain expert manually review prompts. A potential approach to evaluate the degree of straightforwardness of a word may be to review word used in a prompt against a word frequency list generated from a corpus of relevant documents (e.g., collection of nuclear power plant documents, procedures, etc.). In such case, more common words (i.e., words more frequently used) may reflect use of more straightforward terminology. Though, it must be emphasized, that having a human domain expert review the prompt is really the ‘gold standard’ in terms of assuring that this criterion is met.

Finally, it is worth noting that measures of readability may also be considered as a proxy for evaluating the level of straightforwardness of a given prompt. Such measures are typically based on some tracking of syllable frequencies for the words used in the text string. This can have limitations depending on the domain (i.e., in some case, words of greater syllables may be unavoidable), so the application of readability measures for the purpose of evaluating a prompt’s straightforwardness should be used with caution. Referring to the two example prompts (Figure 31 and Figure 32), Table 11 presents calculated readability measures from a variety of indices including Flesch, Flesch Kincaid, Linsear Write, and the simple measure of gobbledygook (SMOG).

Table 11. Comparison of readability measures.

Prompt	Flesch	Flesch Kincaid	Linsear Write	SMOG
Prompt #1	31.55	<u>10.35*</u>	<u>8.00*</u>	<u>11.21*</u>
Prompt #2	<u>11.36*</u>	14.66	16.00	14.55

Each of these measures were developed for the purposes of measuring the readability of text under different contexts of use. For instance, Flesch Kincaid estimates the grade level of text, intended to assess appropriate reading materials for students (Kincaid et al., 1975) whereas SMOG developed by McLaughlin (1969) has been used extensively in the healthcare community (e.g., Shieh & Hosei, 2008). Interestingly, as seen in Table 11, the directionality of readability varies depending on which measure is used (i.e., illustrated by the green values). Collectively, these results suggest the implementation of readability metrics to assess relevance of a prompt should be done in combination with a multi-measure approach (i.e., along with human-centered evaluation).

Principle 4 – Does the Prompt Contain Enough Contextual Information?

Addressing whether enough contextual information is provided in a prompt refers to determining whether a prompt contains sufficient information for the LLM to readily interpret the request without ambiguity. Beyond expert review, this question may also be addressed through PoS tagging, which decomposes the prompt into its respective constituents. The analytical breakdown and comparison of the two prompts previously discussed thus equally apply to this question.

Principle 5 – Does the Prompt Employ Action Verbs?

Ensuring that the prompt uses action verbs reduces ambiguity in interpretation for the LLM (Geroimenko, 2025). Verifying that the prompt uses clear verbs (e.g., prompt #2) provides added clarity for the LLM to more easily interpret the prompt without misinterpretation.

Principle 6 – Is Domain Knowledge Used?

In complex domains like nuclear power plant work, the vocabulary that is used can be highly specific and unique to the domain. For example, terms like “SCRAM” or “trip” have distinct meanings in the nuclear industry that are not relevant to other domains. Just as preparing the LLM should focus on data relevant to the industry (Sections 5 – 7), prompting the LLM should also include such vocabulary.

Principle 7 – Does the Prompt Avoid Leading Questions?

Leading questions should be avoided in prompting an LLM to avoid bias. Rather, prompts should take the form of being neutral. One potential tool to aid in evaluating whether a prompt contains leading questions is sentiment analysis. Take for example the following questions that may be presented as prompts to an LLM:

- 1) "Why is Vendor A better than Vendor B in terms of cost and schedule?"
- 2) "How does Vendor A compare to Vendor B in terms of cost and schedule?"
- 3) "Explain the factors that indicate why Vendor A is either better or worse than Vendor B in terms of cost and schedule?"

Question #1 contains a leading question, as it is suggesting that Vendor A is ‘better’ than Vendor B in terms of cost and schedule. This is not a neutral question if the user is trying to understand the capabilities of each vendor in terms of cost and schedule. When using sentiment analysis, the term ‘better’ is in fact detected as being of *positive* sentiment. Question #2 contains no positive or negative sentiment words, so it is therefore neutral. Question #3 takes a neutral stance because it uses an equal number of positive (‘better’) and negative (‘worse’) sentiment words.

Principle 8 – If the Prompt is to be Reused, is a Standard Template Used?

In some cases, prompts may be reused to inquire specific plant information at different times. In such cases, having a standard prompt template is suggested. For example, a use case may involve using an LLM to support a routine maintenance task that involves systematically decomposing the problem and calculating some value. In such case, a template may be used to provide consistent responses that allow the user to understand the decision-making undergone by the LLM. An example of this can be found in Table 12.

Table 12. CoT template example (adapted from Diamant, 2025).

```
advanced_cot_prompt = PromptTemplate(  
    input_variables=["question"],  
    template="""  
        Solve the following problem step by step. For each step:  
        1. State what you're going to calculate  
        2. Write the formula you'll use (if applicable)  
        3. Perform the calculation  
        4. Explain the result  
  
        Question: {question}  
  
        Solution:  
        """  
)
```

Principle 9 – Are Prompt Components Clearly Separated?

When developing prompt templates, it is also important to delineate components of the prompt (Geroimenko, 2025). As seen in Table 12, each component (e.g., `input_variables` or `template`) are separated by a newline character. By doing so, the potential for confusing the LLM is reduced to ensure more relevant output.

Principle 10 – Was the Prompt Constructed through Iterative Refinement, Starting with a Simpler Prompt?

Finally, it should be re-emphasized that steps in ensuring trustworthy results during prompt engineering requires experimentation and iteration. Generally, prompts should begin with more general, or simple, information, allowing to gain a better understanding of the LLM's capabilities and understanding before delving more deeper into a topic.

8.1.1.2 Issue 2: Information Safety and Security

Another key issue pertains to prompt injection. Prompt injections entails malicious input presented in the form of prompts that influence the LLM to bypass any built-in safeguards for sharing sensitive information. Prompt injections work by coercing the LLM to provide responses that are beyond its intended scope through inputs (prompts) that mimic system level prompts.

8.1.1.2.1 Requirement 1.1: Safe and Secure

Primary safeguards to mitigate prompt injections include limiting access to the LLM through a combination of engineered controls, policy, and procedures; these safeguards are generally identified through risk analysis, as discussed. However, with prompt engineering, designing prompts in ways that clearly delineate system-level prompts from user prompts provide an added layer of protection. Liu and colleagues (2023) offer several approaches that may be considered, including Random Sequence Enclosure and Extensible Markup Language (XML) tagging of user prompts. Both approaches offer unique ways in which user input is uniquely delineated from system level prompts.

8.2 Evaluate Model and Human-AI Interaction

Evaluation is organized by two primary activities. The first aspect that must be evaluated is the LLM itself, which entails ensuring that model is performing at the level of accuracy required of it, based on the use case(s) that it is intended to support. The second aspect that should be evaluated is the LLM being used within its context of use, as emphasized throughout this report. In other words, a model may be well-quipped to perform at a level of accuracy needed but be perceived as untrustworthy by users for reasons beyond its validity and reliability, such as LLM’s usability. The evaluation of the LLM within its context of use is referred to here as *human-AI interaction evaluation*. Both types of evaluations are undergone iteratively throughout the technical activities previously discussed, rather than in one test. This was indicated by the dotted arrows from Figure 11.

Model Evaluation

Model evaluation is a critical step in preparing an LLM for real-world use. The primary goal of model evaluation is to ensure that the model produces valid and reliable output to support the use case(s) that the model is intended to support. From a trustworthiness perspective, model evaluation is meant to provide reasonable confidence that the LLM will produce trustworthy results. Interestingly, evaluating LLMs pose unique challenges. For instance, unlike more ‘quantitative’ AI algorithms (e.g., regression or support vector machines), well-structured standardized evaluation criteria can be difficult, due to the complex nature of LLMs (de Hond et al., 2024). Similarly, LLMs can provide diverse output, which can make identifying a single measure insufficient. As a result, evaluating the LLM typically rely on a combination of both automated and human-based evaluations (van der Lee et al., 2021).

Human-AI Interaction Evaluation

Human-AI interaction evaluation is the key method for addressing trust; it comprehensively confirms, through a combination of qualitative and performance-based tests, that the LLM can be used by plant personnel (i.e., the users of the system) successfully, and is used (i.e., appropriately trusting the LLM), to perform their tasks safely, reliably, and efficiently. Human-AI interaction evaluation is performed through a variety of methods. When the model is being developed, qualitative methods provide feedback on what aspects of the model are most important for user trust; these types of evaluations are regarded as formative assessments. An example of a formative assessment is provided in Appendix B. As the model progresses through each development activity, performance-based tests are used to evaluate the calibration of trust; these types of evaluations are regarded as summative assessments. An example of a summative assessment is provided in Appendix C.

8.2.1 Relevant Cross-Cutting Trust Issues

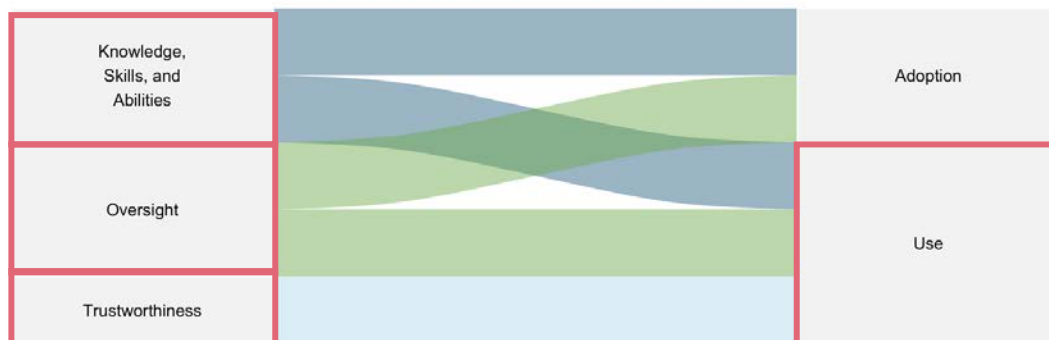


Figure 33. Cross-cutting trust issues addressed at validation.

8.2.1.1 Issue 1: Use

Trust is a key consideration that is addressed during evaluation. Misalignments with competencies, appropriate oversight, or deficiencies in one or more considerations of trustworthiness may have negative implications on trust, leading users to either mistrust (trust the LLM when it is not trustworthy) or distrust (not trusting the LLM when it is trustworthy). Such miscalibrations can have deleterious effects on use, resulting in inefficiencies. The following requirements are therefore addressed through evaluation.

8.2.1.1.1 Requirement 1.1: Knowledge, Skills, and Abilities

Competencies (i.e., the knowledge, skills, and abilities of users using LLM) is assessed through focused elicitation of users' experiences and understanding of LLMs. As discussed in Section 5.1.1.1.1, the results from the AI Literacy Survey (Appendix C-2) can provide an understanding of the competencies that users have with using AI. This information can inform the appropriate level of training needed to support their use at this step. Simulated training and familiarization is applied during evaluation of the LLM as part of a diverse approach to ensuring calibrated trust. The adequacy of training and familiarization is evaluated holistically in conjunction with the LLM under study.

8.2.1.1.2 Requirement 1.2: Oversight

Appropriate levels of oversight is assessed through inspection of whether end users have formed appropriate situation awareness of the work process and have an adequate mental model of how the LLM supports the process. There are several ways that situation awareness can be assessed. Two of the most common ways are through freeze-probe techniques and self-report (Stanton et al., 2017). An in-depth discussion of these methods is beyond the scope of this work. Though, it is worth highlighting here that freeze-probe techniques are done through simulated tasks in which the dynamics of the task are 'frozen' and key indications that are important to the task are captured; the users are then provided a series of targeted questions about the state of certain system states, which offers an objective way of assessing one's understanding of the dynamics of the task and environment. To connect situation awareness with oversight, it can be reasoned that poor awareness as scored by freeze-probe techniques provide a means of addressing whether an appropriate level of oversight was used.

Related to situation awareness, mental models are also an important aspect of appropriate oversight. If oversight is deficient, it can be reasoned that users' mental models of the work process and application of LLM will suffer. While situation awareness captures a user's ability to perceive and understand the dynamics of the environment and task, mental models capture a user's internal representation of how the process and systems work. Both considerations are important for adequate oversight. The assessment of mental models is an open research area. However, Hoffman and colleagues (2023) offer a practical approach to evaluating mental models – termed ShadowBox Lite. ShadowBox Lite simply provides the user a target statement (i.e., this could be the output from an LLM) and asks the user to evaluate what is correct and what is incorrect from the statement. The output of this exercise is transcribed into propositions (i.e., typically presented in the form of a concept map), and the propositions are analyzed against an SME's mental model (i.e., in the form of a concept map). A user's mental model is thus evaluated in terms of completeness and accuracy. Poorly formulated mental models may be indicative of inadequate oversight. Such approach can be used to assess the output from LLM and serve as a proxy for assessing mental models of the domain, as part of assessing whether the appropriate levels of oversight are engaged.

8.2.1.1.3 Requirement 1.3: Trustworthiness

Trustworthiness is a clear driver for user trust. While all cross-cutting issues have some level of impact on trust, a salient issue that impacts trust during use is the LLM's capability of being factually accurate or not (i.e., the extent of hallucinating). Thus, hallucinations are a primary consideration discussed in Section 8.2.2. For the assessment of user trust, a number of approaches can be utilized, depending on when in the process of model development that evaluation is being undergone. Two classes of assessment techniques are offered below, as formative and summative approaches, respectively.

Formative Trust-Focused Usability Assessments

Formative assessments help identify design issues that may contribute to uncalibrated trust. Formative assessments are therefore used earlier in the process when the model is being developed to help with refinement (e.g., done to assess the benefits of fine-tuning or determining the best RAG approach to use). There is a degree of flexibility regarding the ways in which a formative assessment can be applied. The design and implementation of formative assessment is based on the maturity of the system under study and specific objectives. Though, with its primary focus being on issue identification rather than validation, the scope of formative assessment are generally less onerous when compared to summative tests. The degree of maturity needed of the LLM can be a prototype and the number of users sampled can be far less.

Appendix B provides an example of a formative assessment. This assessment evaluated the effectiveness of two different RAG approaches compared to a baseline (no RAG). The approach used a mixture of qualitative (verbal feedback) and quantitative measures (time-on-task and ratings of explanation suitability for instances where the LLM was accurate and inaccurate). Measures of trust were collected in terms of verbal feedback provided by the user in combination with qualitative feedback that provided a basis for their perceived trust. This approach offers rich information regarding the elements of the LLM that are most important to the user in trusting it, providing a way of collecting design feedback that can be used for iterative refinement.

Summative Trust-Focused Usability Assessments

Summative assessments evaluate the overall effectiveness of the designed system, including any training and procedures that are used in part with the system. Summative assessments are completed with a completely designed system and generally applied towards deploying the system into use. Summative assessments apply some form of pass/fail criteria determined by the task or by comparing against a baseline system. Quantitative performance-based measures are generally used. For trust, the selection of such measures may include behavioral observations, such as measures of use, that therefore require a user to trust the system when appropriate.

Appendix C provides a detailed study that captures the essence of summative assessment. The evaluation compared the use of color to present confidence of an LLM's response using its FActSCOREs. This study used a novel approach to evaluating trust, by incorporating signal detection theory as the evaluation framework; the approach collects behavioral data of whether the user believed the atomic fact or not, by responding to a binary true/ false statement that reflects the atomic fact. This approach was used in combination with self-report trust from a battery of standardized trust surveys (see Appendix C-3 and Appendix C-4). A summative assessment such as seen in Appendix C provides a multi-method/ multi-measure approach for evaluating trust, allowing for triangulation of findings to improve overall validity of the evaluation.

8.2.2 Relevant Cross-Cutting Trustworthiness Issues

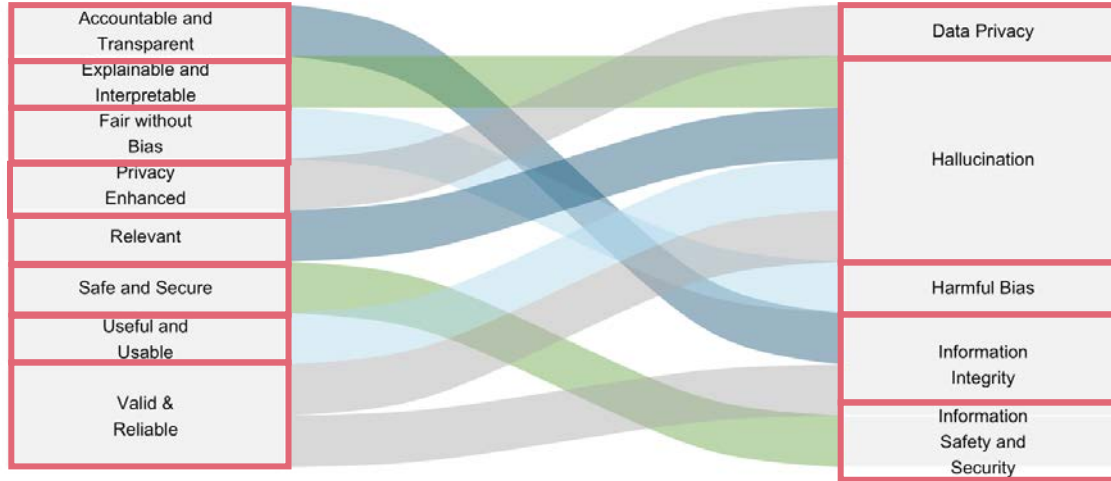


Figure 34. Cross-cutting trustworthiness issues addressed at validation.

8.2.2.1 Issue 1: Data Privacy

This issue was discussed in Section 6.1. It is discussed here because evaluation provides a means to identify potential latent issues (those that were not previously identified) related to privacy.

8.2.2.1.1 Requirement 1.1: Privacy Enhanced

Privacy issues can be identified indirectly from other evaluations of the model (e.g., when assessing hallucinations), or through formal approaches (e.g., AI red teaming; Bullwinkel et al., 2025). When an issue is identified, a different set of decisions from what was previously shared in Section 6.1 may be needed; because it is likely infeasible to change infrastructure once implemented, alternative safeguards may be necessary. For instance, Asthana et al. (2025) discuss a PII mitigation framework for LLMs that provides dynamic contextual PII detection and masking from the output of an LLM before reaching a user. This approach provides a post hoc method for remedying privacy issues for as-built models.

8.2.2.2 Issue 2: Hallucination

Hallucinations are a significant concern with LLMs and thus is a central focus for evaluation. This issue is met by addressing the following requirements at each iterative cycle of evaluation, using a combination of human-based and automated approaches.

8.2.2.2.1 Requirement 2.1: Explainable and Interpretable

Explainability and interpretability are each addressed through a human-based approach.

Human-Based Approaches

SME review is the primary human-based approach to address explainability and interpretability.

Explainable and Interpretable-Focused SME Reviews

The extent that the model is explainable and interpretable is ultimately a perceived quality of the system by the end user. As emphasized in Hoffman and colleagues (2023), the property of being an explanation is not a property of the statement itself, but the interactions among the system and user. Thus, the degree of being explainable is a function of what the user needs to know (i.e., their objectives) and what they already know. In the context of LLM evaluation, this interaction between user and system is done through conversational interface (i.e., a dialog produced by the model). The degree of explainability and

interpretability of this output depends on what specifically the user needs to do with that information. To address explainability and interpretability, human evaluation is thus emphasized. Moreover, the human evaluations should be grounded by the use case to which the LLM is intended to be used for. Application of rating the response produced by the model is thus done using a format such as seen in Figure 35.

The reasoning and logic in formulating the response is clear and interpretable.

1	2	3	4	5
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure 35. Criterion for SME review of explainability and interpretability.

8.2.2.2.2 Requirement 2.2: Relevant

Relevance is addressed through both human-based approaches and automated approaches.

Human-Based Approaches

SME review is the primary human-based approach to address relevance.

Relevance-Focused SME Reviews

Like explainability and interpretability, relevance requires grounding evaluation in the context in how output from the model will be used by the user. The model may produce two correct responses, but with one being of greater detail than the other. Depending on the context of use, one response may be more relevant than another. Human evaluation using standardized criteria such as seen in Figure 36 provides a gold standard.

The level of detail is satisfying.

1	2	3	4	5
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

The response seems complete.

1	2	3	4	5
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

The explanation is relevant to the question asked.

1	2	3	4	5
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure 36. Criteria for SME review of relevance.

Automated Approaches

Automated approaches include applying relevance-focused LLM reviews, answer relevance, and faithfulness (for RAG).

Relevance-Focused LLM Reviews

Relevance can be evaluated by having an independent LLM review output from the LLM under evaluation, using similar criteria as in the SME review (e.g., Figure 36).

Answer Relevance

One automated measure for relevance is known as *answer relevance*. This measure characterized as the mean cosine similarity of the original prompt to a number of artificial questions that are generated by an LLM based on the LLM's response under review. If the generated questions are similar to the original prompt based on the measure, than relevancy is considered high. An example of its use can be found at:

- https://docs.ragas.io/en/v0.1.21/concepts/metrics/answer_relevance.html

Faithfulness

For RAG specifically, the measure of *faithfulness* can be applied to address relevance. Faithfulness refers to the idea that an LLM's answer should be faithful to source documents as expected and not extrapolated or hallucinated (Saad-Falcon et al., 2023; Yu et al., 2024). Any answer or claim the LLM makes that is not derived from a source document decreases faithfulness. The metric faithfulness is defined by two terms, (1) how many claims the LLM makes, and (2) how many of those claims are supported by source documents (Ru, et al., 2024; Es et al., 2023). These two terms are then converted to a faithfulness score by dividing (2) into (1).

8.2.2.2.3 Requirement 2.3: Useful and Usable

Usefulness and usability are each addressed through human-based approaches. Usefulness can also be evaluated using automated approaches.

Human-Based Approaches

SME reviews and usability tests are primary human-based approaches to address usefulness and usability.

Usefulness-Focused SME Reviews

The usefulness of a model's response is evaluated using SME reviews with established criteria (Figure 37). While simple, such approach provides an effective way for a SME or user to review samples of responses generated by LLM.

The explanation is useful for someone in the nuclear industry.

1	2	3	4	5
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure 37. Criterion for SME review of usefulness.

Usability Tests

The usability of the model's produced output can be assessed through formal assessments discussed in Section 8.2.1.1.3. Upon completing the assessment, usability can be assessed through measures that are characteristic of usability (i.e., the extent that LLM is effective, efficient, and satisfying when used to complete the intended task in the operational environment). An in-depth discussion of usability methods is beyond the scope here, but measures like task completion, time-on-task, and self-report are typical to assess the usability of a system (ISO 9241-11, 2018).

Automated Approaches

Automated approaches include applying relevance-focused LLM reviews, answer relevance, and faithfulness (for RAG).

Usefulness-Focused LLM Reviews

Usefulness can be evaluated by having an independent LLM review output from the LLM under evaluation, using similar criteria as in the SME review (e.g., Figure 37).

8.2.2.2.4 Requirement 2.4: Valid and Reliable

Valid and reliable can be assessed using a combination of human-based and automated approaches.

Human-Based Approaches

Hallucination-focused prompt-based behavioral audits and SME reviews are two types of human-based evaluation approaches that address validity and reliability.

Hallucination-Focused Prompt-Based Behavioral Audits

Hallucination-focused prompt-based behavioral audits entail crafting diagnostic prompts that trigger hallucinations (e.g., fabricated citations). These audits provide a qualitative check that complements automated metrics, showing whether safeguards are translating into more valid and reliable output.

Validity and Reliability-Focused SME Reviews

The reviewers judge the model in terms of set criteria – i.e., validity and reliability (e.g., see Figure 38). Human evaluation is critical for detecting subtle forms of hallucination that automated tools could miss.

The explanation shows me how accurate the LLM is.

1	2	3	4	5
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure 38. Criterion for SME review of valid and reliable.

Automated Approaches

The primary automated approach for addressing validity and reliability entails model comparisons.

Model Comparisons (i.e., A/B Evaluation)

Direct comparisons of models can be carried out through evaluation of outputs from a set of common prompts. The specific strategies in prompting can vary depending on the evaluation's goals. Prompts may be engineered to produce content-rich responses, or they may be engineered to produce focused responses. When used for the format (content-rich responses), metrics focused on text summarization are generally a preferred approach. Measures like Recall-Oriented Understudy for Gisting Evaluation (ROUGE; Lin, 2004) or semantic similarity (Reimers & Gurevych, 2019) can be applied for this reason. For hallucination evaluation, higher scores from these metrics tend to reproduce verified facts without inventing details.

For focused responses, prompts may be engineered to structure multiple choice question and answering. This approach provides a straightforward way to assess accuracy between multiple models. Example questions and answers used for this purpose of presented in Appendix A.

Validity and Reliability-Focused LLM Reviews

Validity and reliability can be evaluated by having an independent LLM review output from the LLM under evaluation, using similar criteria as in the SME review (e.g., Figure 38). Worth noting, Example 3 highlights a comparison of the LLM review to SME review in terms of inter-rater reliability. These results emphasize the importance of maintaining a role for SME reviews.

Example 3: Developing a Fine-Tuned LLM Trained on General Nuclear Power Plant Knowledge

It is emphasized that automated approaches to using ratings like in Figure 37 have been explored, such as providing a set of rating criteria to an independent model with assessing aspects of trustworthiness. For instance, this research explored the inter-rater reliability of ratings from a SME to ChatGPT 4o in rating a set of 25 responses in terms of the response being detailed (aspect of relevance), complete (aspect of relevance), relevant, useful, and accurate (aspect of valid and reliable). The results are illustrated in Figure 39 and in Table 13, and they may suggest a few notable points:

- Not all criteria are equally replicable using automated analyses. Accuracy for instance had very little agreement in either condition. Relevance only showed fair agreement with RAG but not with no RAG.
- The greater inter-rater reliability observed for RAG responses suggests that there was a fundamental difference in how the raters evaluated responses between the two response sets even when using the same criteria for both. It is unclear why there are discrepancies in agreement between GPT 4o and the SME.
- It is possible that both types of reviews provide unique insights when evaluating responses using standardized criteria. While these observations are preliminary, future research is needed to consider the specific reasons that SMEs and GPT 4o agree or disagree when rating a response
- Collectively, these findings suggest that careful consideration should be made if using automated approaches for evaluating. Human evaluations appear to remain the gold standard for assessment.

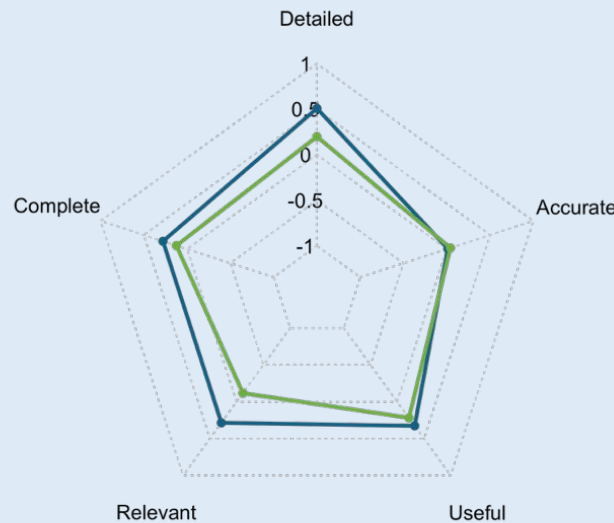


Figure 39. Radar chart of inter-rater reliability (Cohen's Kappa indices) to evaluate level of agreement between ChatGPT 4o to an SME. Green represents agreement for no RAG and blue represents agreement with RAG.

Table 13. Inter-rater reliability (Cohen’s Kappa indices) to evaluate level of agreement between ChatGPT 4o to an SME.

Scale	Default	RAG	Interpreting Kappa Indices
Detailed	0.199	0.509	≤ 0 = Poor Agreement 0.1-0.20 = Slight Agreement 0.21-0.40 = Fair Agreement 0.41-0.60 = Moderate Agreement 0.61-0.80 = Substantial Agreement 0.81-0.99 = Near-Perfect Agreement 1 = Perfect Agreement
Complete	0.123	0.278	
Relevant	-0.117	0.286	
Useful	0.219	0.3271	
Accurate	0.042	0.011	

8.2.2.3 Issue 3: Harmful Bias

Harmful bias is an issue that is addressed through evaluation. Specifically, bias is addressed here through providing users opportunity to review and provide feedback on the fairness of a given response produced by the model.

8.2.2.3.1 Requirement 3.1: Fair without Bias

Fair without bias is addressed through both human-based and automated approaches.

Human-Based Approaches

SME reviews are a key approach for evaluating output for bias.

Bias-Focused SME Reviews

Bias is evaluated using SME reviews with established criteria (Figure 40). While simple, such approach provides an effective way for a SME or user to review samples of responses generated by LLM.

The response is fair without bias.

1	2	3	4	5
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure 40. Criterion for SME review of fair without bias.

Automated Approaches

Automated approaches include bias-focused LLM reviews and prompt-based behavioral audits.

Bias-Focused LLM Reviews

Bias can also be evaluated by having an independent LLM review output from the LLM under evaluation, using similar criteria as in the SME review (e.g., Figure 40).

Bias-Focused Prompt-Based Behavioral Audits

Prompt-based behavioral audits can be extended to also identified potential bias in output produced by the LLM. Specifically, these audits can identify preferences for certain procedural defaults.

8.2.2.4 Issue 4: Information Integrity

Hallucinations identified through evaluation may be due to information integrity issues. As a result, this cross-cutting issue should be continuously monitoring at points in the process where evaluation is done.

8.2.2.4.1 Requirement 4.1: Accountable and Transparent

Accountability and transparency can be assessed in conjunction with evaluating the level of oversight needed (Section 8.2.1.1.2). Specifically, focused assessments of the user's mental model regarding their understanding of the model's underlying architecture and representative data sources used in preparing the model can be assessed to address aspects of accountability and transparency (Liao & Wortman Vaughan, 2024).

8.2.2.4.2 Requirement 4.2: Valid and Reliable

Validity and reliability in the context of evaluation is discussed in Section 8.2.2.2.4. Through the approaches discussed in that section, a key question should be whether any hallucinations identified are a result of inaccurate, misleading, or conflicting data.

8.2.2.5 Issue 5: Information Safety and Security

Finally, the effectiveness of design controls put in place (e.g., fine-tuning, use of RAG, etc.) to mitigate risks associated with safety and security identified with risk analysis should be verified at evaluation.

8.2.2.5.1 Requirement 5.1: Safe and Secure

For tasks that have significant implications to safety and security, usability testing that simulates these tasks should be applied to verify that users can effectively use the LLM without introducing any potential risk modes that have implications to safety or security.

9 SUMMARY OF ISSUES AND REQUIREMENTS ADDRESSED

Collectively, the following tables summarize the cross-cutting issues and associated trustworthiness and trust requirements addressed by the methodology. These tables map each issue/ requirement to the specific activities discussed in Sections 5 through 8.

Table 14. Mapping of cross-cutting issues addressed by the methodology.

		Trustworthiness Issues					Trust Issues	
		<i>Data Privacy</i>	<i>Hallucination</i>	<i>Harmful Bias</i>	<i>Information Integrity</i>	<i>Information Safety and Security</i>	<i>Adoption</i>	<i>Use</i>
Scoping and Data Strategy	<i>Describe the Automated Process and User Tasks</i>						✓	
	<i>Identify System Risks</i>					✓		
	<i>Qualify the Integrity of Available Data</i>				✓			
Model Selection and Data Preparation	<i>Select Hosting Infrastructure</i>	✓						
	<i>Select Pretrained Model</i>		✓					
	<i>Complete Process Processing and Cleansing</i>		✓	✓				
Model Optimization	<i>Fine-Tuning</i>		✓	✓				
	<i>RAG</i>		✓					
Model Integration	<i>Engineer the Prompt</i>		✓			✓		
	<i>Evaluate Model and Human-AI Interaction</i>	✓	✓	✓	✓	✓		✓

Table 15. Mapping of requirements addressed by the methodology.

		Trustworthiness Requirements								Trust Requirements		
		Accountable and Transparent	Explainable and Interpretable	Fair without Bias	Privacy Enhanced	Relevant	Safe and Secure	Useful and Usable	Valid and Reliable	Knowledge, Skills, and Abilities	Oversight	Trustworthiness
Scoping and Data Strategy	Describe the Automated Process and User Tasks									✓	✓	
	Identify System Risks	✓										
	Qualify the Integrity of Available Data	✓									✓	
Model Selection and Data Preparation	Select Hosting Infrastructure	✓										
	Select Pretrained Model									✓		
	Complete Process Processing and Cleansing	✓								✓		
Model Optimization	Fine-Tuning	✓								✓		
	RAG	✓								✓		
Model Integration	Engineer the Prompt	✓										
	Evaluate Model and Human-AI Interaction	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	

10 CONCLUSIONS

LLMs offer promising capabilities to automate and enhance tasks such as document drafting, information retrieval, and decision support. However, their deployment in nuclear power plants is challenged by critical issues related to trust and trustworthiness. This technical report presents a comprehensive methodology that addresses this challenge by evaluating and enhancing the trustworthiness of LLMs in nuclear power plant applications, while ensuring appropriate user trust. By integrating technical and sociotechnical considerations across four lifecycle phases—Work and Data Scoping, Model Selection and Data Preparation, Model Optimization, and Model Integration—the methodology provides a structured, systematic approach to mitigate key technical challenges and human factors challenges. The methodology ensures that LLMs being developed and deployed at nuclear power plants produce outputs that are more valid, reliable, safe, secure, privacy enhanced, usable, useful, and contextually relevant, thereby improving user trust and operational efficiency.

The next step in this research is to demonstrate the methodology, presented in this technical report, to enable trustworthy LLM solutions to be deployed that improve nuclear power plant operations. Using the methodology and partnering with an industry collaborator, future research will select a use case and benchmark LLM-enabled performance that follows this methodology against existing work process performance. By enhancing trustworthiness and user trust, this work collectively supports the adoption and deployment of LLMs across industry to significantly reduce O&M cost.

11 REFERENCES

- ACL 2014 Ninth Workshop on Statistical Machine Translation. 2014. “Statistical and Neural Machine Translation.” Retrieved December 24, 2024, from <https://www.statmt.org/wmt14/index.html>.
- Al Rashdan, A., Giraud, K., Mapes, J., Wilcken, B., Mohon, J., and Murray, P. 2024. “Scalable Methods to Automate Manual Work Management Activities Using Artificial Intelligence.” Idaho National Laboratory, INL/RPT-24-80159. <https://lhrs.inl.gov/content/uploads/11/2024/09/ScalableMethodsAutomateManualWork.pdf>.
- Asthana, S., Mahindru, R., Zhang, B., and Sanz, J. 2025. “Adaptive PII Mitigation Framework for Large Language Models.” *arXiv preprint arXiv:2501.12465*. <https://doi.org/10.48550/arXiv.2501.12465>.
- Bailey, N., and Scerbo, M. 2007. “Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust.” *Theoretical Issues in Ergonomics Science*, 8(4), 321–348. <https://doi.org/10.1080/14639220500535301>.
- Bender, E., and Friedman, B. 2018. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” *Transactions of the Association for Computational Linguistics*, 6, 587–604. https://doi.org/10.1162/tac1_a_00041.
- Blank, I. A. (2023). What are large language models supposed to model?. *Trends in Cognitive Sciences*, 27(11), 987–989. <https://doi.org/10.1016/j.tics.2023.08.006>.
- Bliss, J., and Acton, S. 2003. “Alarm mistrust in automobiles: how collision alarm reliability affects driving.” *Applied Ergonomics*, 34(6), 499–509. <https://doi.org/10.1016/j.apergo.2003.07.003>.
- Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S., et al. 2021. “On the Opportunities and Risks of Foundation Models.” Stanford Institute for Human-Centered Artificial Intelligence. *arXiv preprint arXiv:2108.07258*. <https://arxiv.org/abs/2108.07258>.
- Buchanan, B., and Shortliffe, E. 1984. “Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.” Addison-Wesley. <https://shortliffe.net/Buchanan-Shortliffe-1984/MYCIN%20Book.htm>.
- Bullwinkel, B., Minnich, A., Chawla, S., Lopez, G., Pouliot, M. et al. 2025. Lessons from red teaming 100 generative ai products. *arXiv preprint arXiv:2501.07238*. <https://doi.org/10.48550/arXiv.2501.07238>.
- Cacioppo J., and Petty R. 1982. “The Need for Cognition.” *Journal of Personality and Social Psychology*, 42, 116–131.
- Chancey, E., Bliss, J., Yamani, Y., and Handley, A. 2017. “Trust and the Compliance–Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence.” *Human Factors*, 59(3), 333–345. <https://doi.org/10.1177/0018720816682648> (Original work published 2017).
- Chung, H.-W., Hou, L., Longpre, S., Zoph, B., Tay, Y., et al. 2022. “Scaling Instruction-Finetuned language models.” *arXiv preprint arXiv:2210.11416*. <https://arxiv.org/abs/2210.11416>.

- Dallat, C., P. M. Salmon, and N. Goode. 2018. "Identifying and emergent risks across sociotechnical systems: the NETWORKED hazard analysis and risk management system (NET-HARMS)." *Theoretical Issues in Ergonomics Science*, 19(4): 456–482. <https://doi.org/10.1080/1463922X.2017.1381197>.
- Davis, F. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>.
- Davis, F., Bagozzi, R., and Warshaw, P. 1989. "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models." *Journal of Management Science*, 35(8), 982–1003. <https://doi.org/10.1287/mnsc.35.8.982>.
- de Cerqueira, J., Kemell, K., Rousi, R., Xi, N., Hamari, J., and Abrahamsson, P. 2025. "Mapping Trustworthiness in Large Language Models: A Bibliometric Analysis Bridging Theory to Practice." *arXiv preprint arXiv:2503.04785*. <https://doi.org/10.48550/arXiv.2503.04785>.
- de Hond, A., Leeuwenberg, T., Bartels, R., van Buchem, M., Kant, I., et al. 2024. "From text to treatment: the crucial role of validation for generative large language models in health care." *The Lancet Digital Health*, 6(7), 441–e443. [https://doi.org/10.1016/S2589-7500\(24\)00111-0](https://doi.org/10.1016/S2589-7500(24)00111-0).
- Diamant, N. 2025. "Prompt Engineering from Zero to Hero - Master the Art of AI Interaction." Nir Diamant.
- DOE. 2024. "Generative AI Reference Guide" (Version 2). Department of Energy. <https://www.energy.gov/sites/default/files/2024-12/Generative%20AI%20Reference%20Guide%20v2%206-14-24.pdf>.
- Dorton, S. L., Korman, J., Ralph, J. M., and Speaker, S. R. 2025. "Systemic Trust in Artificial Intelligence." *Ergonomics in Design*, 0(0). <https://doi.org/10.1177/10648046251366116>
- Dzindolet, M., Peterson, S., Pomranky, R., Pierce, L., and Beck, H. 2003. "The role of trust in automation reliance." *International Journal of Human-Computer Studies*, 58(6), 697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7).
- Endsley, M. 2023. "Ironies of artificial intelligence." *Ergonomics*, 66(11), 1656–1668. <https://doi.org/10.1080/00140139.2023.2243404>.
- Engber, C. 1995. "The relationship of lexical proficiency to the quality of ESL compositions." *Journal of Second Language Writing*, 4(2), 139–155. [https://doi.org/10.1016/1060-3743\(95\)90004-7](https://doi.org/10.1016/1060-3743(95)90004-7).
- EPRI. 2015. "Human Factors Guidance for Control Room and Digital Human-System Interface Design and Modification: Guidelines for Planning, Specification, Design, Licensing, Implementation, Training, Operation, and Maintenance for Operating Plants and New Builds." Report 3002004310, Electric Power Research Institute. <https://www.epri.com/research/products/000000003002004310>.
- Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. 2023. "Ragas: Automated Evaluation of Retrieval Augmented Generation." *arXiv preprint arXiv:2309.15217*. <https://doi.org/10.48550/arXiv.2309.15217>.
- Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. 2024. "Open LLM Leaderboard v2." Retrieved from https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

- Friedl, J. 2006. "Mastering Regular Expressions," Third Edition. O'Reilly Media.
- Gadde, S., and Kalli, V. 2020. "Applications of artificial intelligence in medical devices and healthcare." *International Journal of Computer Science Trends and Technology*, 8(2), 182–188.
https://www.researchgate.net/profile/Venkata-Kalli/publication/341204087_Applications_of_Artificial_Intelligence_in_Medical_Devices_and_Healthcare/links/5eb3b9ba92851cbf7fadc1ba/Applications-of-Artificial-Intelligence-in-Medical-Devices-and-Healthcare.pdf.
- Geroimenko, V. 2025. "The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks." Springer Nature.
- Gkikas, D., and Theodoridis, P. 2021. "AI in Consumer Behavior." *Advances in Artificial Intelligence-based Technologies*, 147–176. Cham: Springer International Publishing.
https://doi.org/10.1007/978-3-030-80571-5_10.
- Gururangan, S., Marasovifá, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. 2020. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." *Proceedings of ACL 2020*, 8342–8360. <https://doi.org/10.18653/v1/2020.acl-main.740>.
- Halko, N., Martinsson, P., and Tropp, J. 2011. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions." *SIAM Review*, 53(2), 217–288.
<https://doi.org/10.1137/090771806>.
- Hall, A., and Agarwal, V. 2024. "Barriers to the adoption of artificial intelligence and machine learning technologies in nuclear power." *Progress in Nuclear Energy*, 175, 105295.
<https://doi.org/10.1016/j.pnucene.2024.105295>.
- Hall, A., Murray, P., Boring, R., and Agarwal, V. 2024. "Human-Centered and Explainable Artificial Intelligence in Nuclear Operations." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68(1), 1563–1568. Sage CA: Los Angeles, CA: SAGE Publications.
<https://doi.org/10.1177/10711813241276463>.
- Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., et al. 2025. "Retrieval-Augmented Generation with Graphs (GraphRAG)." *arXiv preprint arXiv:2501.00309*. <https://arxiv.org/abs/2501.00309>.
- Hart, S. G., and Staveland, L. E. 1988. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." In *Advances in Psychology*, 52, 139–183.
[https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9).
- Hoerl, A., and Kennard, R. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1), 55–67. <https://doi.org/10.2307/1267351>.
- Hoff, K., and Bashir, M. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust." *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>.
- Hoffman, R., Mueller, S., Klein, G., and Litman, J. 2023. "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance." *Frontiers in Computer Science*, 5, 1096257. <https://doi.org/10.3389/fcomp.2023.1096257>.

- Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. 2020. “The Curious Case of Neural Text Degeneration.” In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rygGQyrFvH>.
- Hsu, F.-H. 1999. “IBM’s Deep Blue Chess grandmaster chips.” *IEEE micro*, 19(2), 70–81. <https://doi.org/10.1109/40.755469>.
- Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., et al. 2021. “LoRA: Low-Rank Adaptation of Large Language Models.” *arXiv preprint arXiv:2106.09685*. <https://arxiv.org/abs/2106.09685>.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., et al. 2025. “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.” *ACM Transactions on Information Systems*, 43(2), 1–55. <https://doi.org/10.1145/3703155>.
- Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., et al. 2024. “TrustLLM: Trustworthiness in Large Language Models.” *arXiv preprint arXiv:2401.05561*. <https://doi.org/10.48550/arXiv.2401.05561>.
- Hugging Face. “Embedding Leaderboard.” MTEB Leaderboard. Accessed April 28, 2025.
- IEC/ISO. 2022. “Trustworthiness — Vocabulary.” IEC/ISO TS 5723:2022
- IEEE. 2021. “IEEE Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution, and Management.” IEEE-2941-2021. <https://standards.ieee.org/ieee/2941/10363/>.
- ISO. 2018. “Ergonomics of Human-System Interaction – Part 11: Usability: Definitions and Concepts” ISO 9241-11:2018.
- Jamieson, G., Skraaning, G., and Joe, J. 2022. “The B737 MAX 8 Accidents as Operational Experiences with Automation Transparency.” *IEEE Transactions on Human-Machine Systems*, 52(4), 794–797. <https://doi.org/10.1109/THMS.2022.3164774>.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., et al. 2023. “Survey of Hallucination in Natural Language Generation.” *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>.
- Jian, J., Bisantz, A., and Drury, C. 2000. “Foundations for an Empirically Determined Scale of Trust in Automated Systems.” *International Journal of Cognitive Ergonomics*, 4(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04.
- Johnson, J., Douze, M., and Jégou, H. 2019. “Billion-Scale Similarity Search with GPUs.” *IEEE Transactions on Big Data*, 7(3), 535–547. <https://doi.org/10.1109/TBDATA.2019.2921572>.
- Juhasz, B. 2008. “The processing of compound words in English: Effects of word length on eye movements during reading.” *Language and Cognitive Processes*, 23(7-8), 1057–1088. <https://doi.org/10.1080/01690960802144434>.
- Kalodanis, K., Rizomiliotis, P., Feretzakis, G., Kalles, D., Verykios, V., & Anagnostopoulos, D. (2024, December). Enhancing Transparency in Large Language Models to Meet EU AI Act Requirements. In *Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics* (pp. 281–286). <https://doi.org/10.1145/3716554.3716597>.

- Kamran, S., Haleem, A., Bahl, S., Javaid, M., Prakash, C., and Budhhi, D. 2022. "Artificial intelligence and advanced materials in automotive industry: Potential applications and perspectives." In *Materials Today: Proceedings*, 62(6), 4207–4214. <https://doi.org/10.1016/j.matpr.2022.04.727>.
- Kaplan, A., Kessler, T., Brill, J., and Hancock, P. 2023. "Trust in Artificial Intelligence: Meta-Analytic Findings." *Human Factors*, 65(2), 337–359. <https://doi.org/10.1177/00187208211013988>.
- Karpinsky, N., Chancey, E., and Yamani, Y. 2016. "Modeling Relationships among Workload, Trust, and Visual Scanning in an Automated Flight Task." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1), 1550–1554. <https://doi.org/10.1177/1541931213601356>.
- Khalid, M. H., Ahmed, M. I., Roy, S., Bui, H., Reihani, S., Al Rashdan, A., and Mohaghegh, Z. 2023a, October. "Evaluation of Automation Trustworthiness in Nuclear Power Plants: A Risk-Informed Approach using Probabilistic Validation and Integrated Probabilistic Risk Assessment." In *Probabilistic Safety Assessment and Management (PSAM) International Topical Meeting on Artificial Intelligence (AI) and Risk Analysis*.
- Khalid, M. H., Bui, H., Farshadmanesh, P., Al Rashdan, A., and Mohaghegh, Z. 2023b, October. "Automation Trustworthiness in Nuclear Power Plants: A Literature Review." In *Probabilistic Safety Assessment and Management (PSAM) International Topical Meeting on Artificial Intelligence (AI) and Risk Analysis*.
- Kincaid, J., Fishburne Jr., R., Rogers, R., and Chissom, B. 1975. "Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel" (no. RBR875). Institute for Simulation and Training, 56. https://stars.library.ucf.edu/istlibrary/56/?utm_source=stars.library.ucf.edu%2Fistlibrary%2F56&utm_medium=PDF&utm_campaign=PDFCoverPages.
- Kosmyna, N., Hauptmann, E., Yuan, Y., Situ, J., Liao, X.-H., et al. 2025. "Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task." *arXiv preprint arXiv:2506.08872*. <https://arxiv.org/abs/2506.08872>.
- Kovesdi, C. 2021. "Examining the Use of the Technology Acceptance Model for Adoption of Advanced Digital Technologies in Nuclear Power Plants." In *International Conference on Applied Human Factors and Ergonomics*, 502–509. Cham: Springer International Publishing.
- Kovesdi, C., Spangler, R., Mohon, J., and Murray, P. 2024. "Development of Human and Technology Integration Guidance for Work Optimization and Effective Use of Information." Idaho National Laboratory, INL/RPT-24-77684. <https://doi.org/10.2172/2377343>.
- Kovesdi, C., Tylecote, A., Schadegg, M., Martin, L., and Al Rashdan, A. 2025. "Business Case Analysis for Artificial Intelligence-Large Language Model Technology Integration." Idaho National Laboratory, INL/RPT-25-85583.
- Kowald, D., Scher, S., Pammer-Schindler, V., Müllner, P., Waxnegger, K., et al. 2024. "Establishing and evaluating trustworthy AI: overview and research challenges." *Frontiers in Big Data*, 7, 1467222. <https://doi.org/10.3389/fdata.2024.1467222>.

- Kriman, N. E. 2024. “Measuring text summarization factuality using atomic facts entailment metrics in the context of retrieval augmented generation.” *arXiv preprint arXiv:2408.15171*. <https://doi.org/10.48550/arXiv.2408.15171>.
- Kukich, K. 1992. “Techniques for automatically correcting words in text.” *ACM Computing Surveys (CSUR)*, 24(4), 377–439. <https://doi.org/10.1145/146370.146380>.
- Lee, J., and Moray, N. 1992. “Trust, control strategies and allocation of function in human-machine systems.” *Ergonomics*, 35(10), 1243–1270. <https://doi.org/10.1080/00140139208967392>.
- Lee, J., and See, K. 2004. “Trust in Automation: Designing for Appropriate Reliance.” *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.
- Levenshtein, V. 1966. “Binary codes capable of correcting deletions, insertions, and reversals.” *Soviet Physics Doklady*, 10(8), 707–710.
- Leveson, N., and Thomas, J. 2018. “STPA Handbook.” http://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf.
- Lin, C. 2004. “ROUGE: A package for automatic evaluation of summaries.” In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81. Association for Computational Linguistics. <https://aclanthology.org/W04-1013.pdf>.
- Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z. et al. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*. <https://doi.org/10.48550/arXiv.2306.05499>
- Liu, Y., Yao, T., Ton, J.-F., Zhang, X., Guo, R. et al. 2023. “Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models’ Alignment.” *arXiv preprint arXiv:2308.05374*. <https://doi.org/10.48550/arXiv.2308.05374>.
- LLM Safety Leaderboard - a Hugging Face Space by AI-Secure. Huggingface.co. <https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard>. Accessed May 2025.
- Loft, S., Bhaskara, A., Lock, B., Skinner, M., Brooks, J., et al. 2023. “The Impact of Transparency and Decision Risk on Human–Automation Teaming Outcomes.” *Human Factors*, 65(5), 846–861. <https://doi.org/10.1177/00187208211033445>.
- Long, S., Lee, J., Yamani, Y., Unverricht, J., and Itoh, M. 2022. “Does automation trust evolve from a leap of faith? An analysis using a reprogrammed pasteurizer simulation task.” *Applied Ergonomics*, 100, 103674. <https://doi.org/10.1016/j.apergo.2021.103674>.
- Lu, C., Lyu, J., Zhang, L., Gong, A., Fan, Y., et al. 2020. “Nuclear Power Plants with Artificial Intelligence in Industry 4.0 era: Top-Level Design and Current Applications—A Systemic Review.” *IEEE Access*, 8, 194315–194332. <https://doi.org/10.1109/ACCESS.2020.3032529>.
- Lu, X. 2011. “A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers’ Language Development.” *TESOL Quarterly*, 45(1), 36–62. <https://doi.org/10.5054/tq.2011.240859>.
- Lyons, J., and Stokes, C. 2012. “Human–human reliance in the context of automation.” *Human Factors*, 54(1), 112–121. <https://doi.org/10.1177/0018720811427034>.

- Marino, M. "I, chatbot: The gender and race performativity of conversational agents." University of California, Riverside, 2006. http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqm&rft_dat=xri:pqdiss:3210419
- Martell, M., Baweja, J., and Dreslin, B. 2025. "Mitigative Strategies for Recovering From Large Language Model Trust Violations." *Journal of Cognitive Engineering and Decision Making*, 19(1), 76–95. <https://doi.org/10.1177/15553434241303577>.
- Mayer, R., Davis, J., and Schoorman, F. 1995. "An Integrative Model of Organizational Trust." *The Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.2307/258792>.
- McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. 2006. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." *AI Magazine*, 27(4), 12–14. <https://doi.org/10.1609/aimag.v27i4.1904>.
- McInnes, L., Healy, J., and Astels, S. 2017. "hdbscan: Hierarchical density based clustering." *Journal of Open Source Software*, 2(11), 205. <https://doi.org/10.21105/joss.00205>.
- McInnes, L., Healy, J., and Melville, J. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv preprint arXiv:1802.03426*. <https://arxiv.org/abs/1802.03426>.
- McLaughlin, G. 1969. "SMOG Grading: A New Readability Formula." *Journal of Reading*, 12(8), 639–646. <https://www.jstor.org/stable/40011226>.
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W. T., Koh, P. W., ... and Hajishirzi, H. 2023. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." *arXiv preprint arXiv:2305.14251*. <https://doi.org/10.48550/arXiv.2305.14251>.
- Muir, B. 1987. "Trust between humans and machines, and the design of decision aids." *International Journal of Man-Machine Studies*, 27(5-6), 527–539. [https://doi.org/10.1016/S0020-7373\(87\)80013-5](https://doi.org/10.1016/S0020-7373(87)80013-5).
- Muir, B., and Moray, N. 1996. "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation." *Ergonomics*, 39(3), 429–460. <https://doi.org/10.1080/00140139608964474>.
- Mylius, S. 2025. "Systematic Hazard Analysis for Frontier AI using STPA." *arXiv preprint arXiv:2506.01782*. <https://doi.org/10.48550/arXiv.2506.01782>.
- NIST. 2022. "Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations." Technical Report NIST SP 800-161r1-upd1. National Institute of Standards and Technology, Gaithersburg, MD: NIST AI 100–1. <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-161r1-upd1.pdf>.
- NIST. 2023. "Artificial Intelligence Risk Management Framework" (AI RMF 1.0). Technical Report NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg, MD: NIST AI 100–1. <https://doi.org/10.6028/NIST.AI.100-1>.
- NIST. 2024. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." National Institute of Standards and Technology, Gaithersburg, MD: NIST AI 600–1. <https://doi.org/10.6028/NIST.AI.600-1>.

- OpenAI. 2023. “GPT-4 system card.” <https://openai.com/research/gpt-4-system-card>.
- Parasuraman, R., and Manzey, D. 2010. “Complacency and bias in human use of automation: an attentional integration.” *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>.
- Parasuraman, R., and Riley, V. 1997. “Humans and Automation: Use, Misuse, Disuse, Abuse.” *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>.
- Perkins, L., Miller, J., Hashemi, A., and Burns, G. 2010. “Designing for Human-Centered Systems: Situational Risk as a Factor of Trust in Automation.” In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(25), 2130–2134. Sage CA: Los Angeles, CA: SAGE Publications.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., et al. 2020. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” *Journal of Machine Learning Research*, 21(140), 1–67. <https://jmlr.org/papers/v21/20-074.html>.
- Reimers, N., and Gurevych, I. 2019. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>.
- Rempel, J., Holmes, J., and Zanna, M. 1985. “Trust in close relationships.” *Journal of Personality and Social Psychology*, 49(1), 95–112. <https://doi.org/10.1037/0022-3514.49.1.95>.
- Rothman, D. 2024. “RAG-Driven Generative AI: Build custom retrieval augmented generation pipelines with LlamaIndex, Deep Lake, and Pinecone.” Packt Publishing.
- Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., et al. 2024. “RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation.” *arXiv preprint arXiv:2408.08067*. <https://doi.org/10.48550/arXiv.2408.08067>.
- Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M. et al. 2023. “ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems.” *arXiv preprint arXiv:2311.09476*. <https://doi.org/10.48550/arXiv.2311.09476>.
- Salmon, P., Stanton, N., Walker, G., Hulme, A., Goode, N., et al. 2023. “Handbook of Systems Thinking Methods.” CRC Press. <https://doi.org/10.1201/9780429281624>.
- Salton, G., and Buckley, C. 1988. “Term-weighting approaches in automatic text retrieval.” *Information Processing and Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0).
- Sato, T., Inman, J., Politowicz, M., Chancey, E., and Yamani, Y. 2023. “A Meta-Analytic Approach to Investigating the Relationship between Trust and Attention Allocation.” In *67th International Annual Meeting of the Human Factors and Ergonomics Society*, 20230014121.
- Schaller, R. 1997. “Moore’s law: past, present and future.” *IEEE spectrum*, 34(6), 52–59. <https://doi.org/10.1109/6.591665>.

- Sheridan, T. 2019. "Extending Three Existing Models to Analysis of Trust in Automation: Signal Detection, Statistical Parameter Estimation, and Model-Based Control." *Human Factors*, 61(7), 1162–1170. <https://doi.org/10.1177/0018720819829951>.
- Shieh, C., and Hosei, B. 2008. "Printed health information materials: evaluation of readability and suitability." *Journal of Community Health Nursing*, 25(2), 73–90. <https://doi.org/10.1080/07370010802017083>.
- Simon, M., Houghton, S. M., and Aquino, K. 2000. "Cognitive biases, risk perception, and venture formation: How individuals decide to start companies." *Journal of Business Venturing*, 15(2), 113–134. [https://doi.org/10.1016/S0883-9026\(98\)00003-2](https://doi.org/10.1016/S0883-9026(98)00003-2).
- Sitkin, S., and Pablo, A. 1992. "Reconceptualizing the determinants of risk behavior." *Academy of Management Review*, 17(1), 9–38. <https://doi.org/10.2307/258646>.
- Spatharioti, S., Rothschild, D., Goldstein, D., and Hofman, J. 2025. "Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–15. ACM. <https://doi.org/10.1145/3706598.3714082>.
- Stanislaw, H., and Todorov, N. 1999. "Calculation of signal detection theory measures." *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. <https://doi.org/10.3758/BF03207704>.
- Stanton, B., and Jensen, T. 2021. "Trust and artificial intelligence." preprint, 10.
- Stanton, N., Salmon, P., Rafferty, L., Walker, G., Baber, C., and Jenkins, D. 2017. "Human Factors Methods: A Practical Guide for Engineering and Design." CRC Press.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., et al. 2023. "Llama 2: Open Foundation and Fine-Tuned Chat Models." *arXiv preprint arXiv:2307.09288*. <https://doi.org/10.48550/arXiv.2307.09288>.
- Tully, S. M., Longoni, C., and Appel, G. 2025. "Lower artificial intelligence literacy predicts greater AI receptivity." *Journal of Marketing*. 89(5), 1–20. <https://doi.org/10.1177/00222429251314491>.
- Turing, A. 2009. "Computing Machinery and Intelligence." Springer Netherlands.
- van der Lee, C., Gatt, A., Van Miltenburg, E., and Krahmer, E. 2021. "Human evaluation of automatically generated text: Current trends and best practice guidelines." *Computer Speech & Language*, 67, 101151. <https://doi.org/10.1016/j.csl.2020.101151>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 5998–6008. https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dec91fbd053c1c4a845aa-Paper.pdf.
- Venkatesh, V., and Davis, F. 2000. "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies." *Management Science*, 46(2), 186–204. <https://www.jstor.org/stable/2634758>.

- Viveros-Muñoz, R., Carrasco-Sáez, J., Contreras-Saavedra, C., San-Martín-Quiroga, S., and Contreras-Saavedra, C. E. 2025. “Does the Grammatical Structure of Prompts Influence the Responses of Generative Artificial Intelligence? An Exploratory Analysis in Spanish.” *Applied Sciences*, 15(7), 3882. <https://doi.org/10.3390/app15073882>.
- Walker, C., Lin, L., Agarwal, V., Lybeck, N., Hall, A., et al. 2024. “Demonstration and evaluation of explainable and trustworthy predictive technology for condition-based maintenance.” Idaho National Laboratory. INL/RPT-24-80727. https://inldigitallibrary.inl.gov/sites/STI/STI/Sort_133534.pdf.
- Walker, C., et al. 2023. “Explainable Artificial Intelligence Technology for Predictive Maintenance.” Idaho National Laboratory. INL/RPT-23-74159. https://inldigitallibrary.inl.gov/sites/sti/sti/Sort_67216.pdf.
- Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., et al. 2024. “MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark.” *arXiv arXiv:2406.01574*. <http://arxiv.org/abs/2406.01574>.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., et al. 2022. “Taxonomy of Risks Posed by Language Models.” In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. <https://doi.org/10.1145/3531146.3533088>.
- Weizenbaum, J. 1976. “Computer Power and Human Reason: From Judgment to Calculation.” In *The Johns Hopkins University Press and the Society for the History of Technology*.
- Wirz, C., Demuth, J., Bostrom, A., Cains, M., Ebert-Uphoff, I., et al. 2025. “(Re)Conceptualizing trustworthy AI: A foundation for change.” *Artificial Intelligence*, 342, 104309. <https://doi.org/10.1016/j.artint.2025.104309>.
- Yamani, Y., Long, S., and Itoh, M. 2020. “Human–Automation Trust to Technologies for Naïve Users Amidst and Following the COVID-19 Pandemic.” *Human Factors*, 62(7), 1087–1094. <https://doi.org/10.1177/0018720820948981>.
- Yamani, Y., Long, S., Sato, T., Braitman, A., Politowicz, M., and Chancey, E. 2024. “Multilevel Confirmatory Factor Analysis Reveals Two Distinct Human–Automation Trust Constructs.” *Human Factors*, 67(2), 166–180. <https://doi.org/10.1177/00187208241263774>.
- Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., and Liu, Z. 2024. “Evaluation of Retrieval-Augmented Generation: A Survey.” *arXiv preprint arXiv:2405.07437*. https://doi.org/10.1007/978-981-96-1024-2_8.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., et al. 2022. “OPT: Open Pre-trained Transformer Language Models.” *arXiv preprint arXiv:2205.01068*. <https://doi.org/10.48550/arXiv.2205.01068>.
- Zhou, Z., Xiang, J., Chen, C., and Su, S. 2023. “Quantifying and Analyzing Entity-level Memorization in Large Language Models.” *arXiv arXiv:2308.15727*. <https://arxiv.org/abs/2308.15727>.

Page intentionally left blank

Appendix A

Benchmark Questions and Results for LLM trustworthiness Evaluations with RAG

Passage	Fact	Question	Options	Answer	Sources
These monitoring methods are further categorized into periodic testing, surveillance testing, continuous monitoring and inspections. This study focuses on the methods and procedures for performing the above activities to maintain the motors operationally ready in a nuclear facility.	Motor condition monitoring methods include periodic testing, surveillance testing, continuous monitoring, and inspections.	Which of the following is NOT one of the four categories of motor condition monitoring methods described?	['A) Periodic testing', 'B) Surveillance testing', 'C) Simulation modeling', 'D) Continuous monitoring']	C	https://www.nrc.gov/docs/ML0412/ML041280255.pdf
The overall preventive maintenance program is separated into two broad areas of activity aimed at mitigating the potential effects of equipment aging: Performance Evaluation and Equipment Maintenance. The latter involves actually maintaining the condition of the equipment while the former involves those activities undertaken to monitor degradation due to aging.	A preventive maintenance program is split into Performance Evaluation and Equipment Maintenance.	In the preventive maintenance program, which area focuses on monitoring degradation due to aging?	['A) Equipment Maintenance', 'B) Performance Evaluation', 'C) Corrective Maintenance', 'D) Predictive Planning']	B	https://www.nrc.gov/docs/ML0412/ML041280255.pdf
The intrusive-ness of test methods and the present state-of-the-art for using the test equipment in a plant environment are discussed. In conclusion, implementation of the information provided in this report will improve motor reliability in nuclear power plants.	Intrusive test methods and their applicability in-plant are key to improving motor reliability.	What aspect of test methods does the report emphasize as critical for improving motor reliability?	['A) Low cost', 'B) Intrusive-ness and in-plant applicability', 'C) Remote access capability', 'D) Automated data logging']	B	https://www.nrc.gov/docs/ML0412/ML041280255.pdf
The functional indicators evaluated in this study for each of the four categories are summarized in Table 1-1. By definition, some of these parameters can be included in more than one category. For example, the motor bearing vibration can be part of any of the four categories.	Motor bearing vibration can be used as a functional indicator in multiple monitoring categories.	Motor bearing vibration is a functional indicator that can be used in which of the following monitoring categories?	['A) Periodic testing only', 'B) Continuous monitoring only', 'C) Periodic testing, surveillance testing, continuous monitoring, and inspections', 'D) None of the above']	C	https://www.nrc.gov/docs/ML0412/ML041280255.pdf
Also included in these activities are failure modes and effect analysis (FMEA), root cause analysis, tests on failed motor components, and trends associated with these failures.	Failure modes and effects analysis, root cause analysis, and trend evaluation are integral to performance evaluation activities.	Which of the following activities is NOT listed as part of performance evaluation for motor maintenance?	['A) Failure modes and effects analysis', 'B) Root cause analysis', 'C) Trend evaluation', 'D) Finance cost-benefit analysis']	D	https://www.nrc.gov/docs/ML0412/ML041280255.pdf
Most fossil power plants contain an abundance of on-line operating data that is archived to provide detailed information concerning the operation of many plant components. The data for these components includes operating temperatures, pressures, flows, and vibration levels.	Condition-based monitoring uses on-line data such as temperatures, pressures, flows, and vibration levels.	Which type of data is NOT listed as part of the on-line operating information collected for condition-based monitoring?	['A) Operating temperatures', 'B) Pressures', 'C) Flow rates', 'D) Electrical voltages']	D	https://restservice.epri.com/publicdownload/0000000001008416/0/Product
Equipment condition assessment is a direct extension of on-line monitoring and refers to the process by which empirical modeling is used for predictive maintenance of plant components. The ultimate objective is to not only detect anomalies but also to determine the root cause of the problem based on the pattern of anomalies.	Equipment condition assessment uses empirical models for predictive maintenance to detect anomalies and identify root causes.	What is the primary objective of equipment condition assessment beyond detecting anomalies?	['A) Reducing data storage needs', 'B) Determining the root cause of problems', 'C) Automating equipment repairs', 'D) Scheduling staff shifts']	B	https://restservice.epri.com/publicdownload/0000000001008416/0/Product
Condition based calibration schedules present several advantages, including: reduced labor costs, reduced personnel radiation exposure, and increased instrument availability.	Condition-based calibration schedules reduce labor costs, reduce radiation	Which of the following is NOT listed as an advantage of condition-based calibration schedules?	['A) Reduced labor costs', 'B) Reduced personnel radiation exposure', 'C) Increased	D	https://restservice.epri.com/publicdownload/00

Passage	Fact	Question	Options	Answer	Sources
	exposure, and increase instrument availability.		instrument availability', 'D) Improved absolute measurement accuracy']		0000000001008416/0/Product
The development of an ECM model based on OLM technology requires the following general steps: 1. Select Component 2. Construct an Empirical Model 3. Define Failure Modes 4. Create Failure Signature Chart 5. Develop Anomaly Interpreter	Developing an ECM model involves selecting the component, constructing an empirical model, defining failure modes, creating a failure signature chart, and developing an anomaly interpreter.	Which of the following is NOT one of the general steps for developing an ECM model based on OLM technology?	['A) Selecting the component', 'B) Constructing an empirical model', 'C) Writing system-level optimization routines', 'D) Creating a failure signature chart']	C	https://restservice.epri.com/publicdownload/0000000001008416/0/Product
One weakness of this technique is that only relative changes in the performance can be detected, not absolute performance. If the loop is performing badly in the model training data set, then it will not be detected when the model is running online.	A limitation of empirical-model based online monitoring is that it cannot detect absolute performance deviations if poor performance is present in the training data.	What is a noted weakness of empirical-model based online monitoring?	['A) It requires manual data entry', 'B) It cannot detect absolute performance deviations', 'C) It only works offline', 'D) It replaces traditional calibration methods entirely']	B	https://restservice.epri.com/publicdownload/0000000001008416/0/Product
The inservice testing program for pumps requires that each safety-related pump be tested at intervals not to exceed 10 years under full-flow conditions to verify operability.	Safety-related pumps must be tested at least once every 10 years under full-flow conditions.	What is the maximum interval for inservice testing of safety-related pumps?	['A) 5 years', 'B) 10 years', 'C) 15 years', 'D) 20 years']	B	https://www.nrc.gov/docs/ML1021/ML102100236.pdf
Pump testing procedures include verification of flow rate, vibration level checks, and differential pressure measurements across the pump casing.	Inservice pump testing includes verifying flow rates, checking vibration levels, and measuring differential pressures.	Which activity is NOT part of the inservice pump testing procedures?	['A) Flow rate verification', 'B) Vibration level checks', 'C) Differential pressure measurement', 'D) Thermal imaging']	D	https://www.nrc.gov/docs/ML1021/ML102100236.pdf
An observed failure rate exceeding 2% per demand for a given pump model necessitates a root cause analysis to be conducted.	A pump failure rate over 2% per demand triggers a root cause analysis.	At what failure rate per demand must a root cause analysis be initiated for a pump model?	['A) 1%', 'B) 2%', 'C) 5%', 'D) 10%']	B	https://www.nrc.gov/docs/ML1021/ML102100236.pdf
The acceptance criteria for inservice pump tests are based on comparison to the baseline values established during initial testing.	Inservice pump test acceptance criteria are derived by comparing current results to initial baseline values.	Inservice pump test acceptance criteria are based on comparison with what?	['A) Manufacturer specifications', 'B) Industry averages', 'C) Initial baseline values', 'D) Regulatory limits']	C	https://www.nrc.gov/docs/ML1021/ML102100236.pdf
Any failures identified during inservice testing must undergo a documented root cause analysis within 30 days of the test.	Failures found in IST must have a root cause analysis completed within 30 days.	Within how many days must a root cause analysis be completed after finding a failure in IST?	['A) 7 days', 'B) 15 days', 'C) 30 days', 'D) 60 days']	C	https://www.nrc.gov/docs/ML1021/ML102100236.pdf
Guidelines indicate that pumps shall be inspected for seal and packing leakage during each IST interval.	Seal and packing leakage inspection is required during every pump inservice test interval.	What inspection is required for pumps during each IST interval?	['A) Electrical insulation resistance', 'B) Seal and packing leakage', 'C) Bearing alignment', 'D) Shaft runout']	B	https://www.nrc.gov/docs/ML1122/ML112231412.pdf
The report recommends trending of vibration data to detect gradual degradation of pump bearings over time.	Trending vibration data helps detect gradual pump bearing degradation.	Why is vibration data trended for pumps according to the guidelines?	['A) To calibrate sensors', 'B) To detect gradual bearing degradation', 'C) To validate pump design', 'D) To measure flow rate']	B	https://www.nrc.gov/docs/ML1122/ML112231412.pdf
Pumps exhibiting vibration amplitudes above 0.5 mils peak-to-peak are flagged for additional evaluation.	Vibration above 0.5 mil peak-to-peak triggers additional pump evaluation.	At what vibration amplitude are pumps flagged for further evaluation?	['A) 0.1 mil', 'B) 0.5 mil', 'C) 1.0 mil', 'D) 2.0 mil']	B	https://www.nrc.gov/docs/ML1122/ML112231412.pdf
Motor current signature analysis is cited as a useful non-intrusive technique for early fault detection in pump motors.	Motor current signature analysis is a non-intrusive early fault detection method for pump motors.	Which non-intrusive technique is recommended for early fault detection in pump motors?	['A) Thermal imaging', 'B) Motor current signature analysis', 'C) Ultrasonic testing', 'D) Oil analysis']	B	https://www.nrc.gov/docs/ML1122/ML112231412.pdf

Passage	Fact	Question	Options	Answer	Sources
The guidelines suggest conducting root cause analysis of pump failures using both failure modes and effects analysis (FMEA) and fishbone diagrams.	Root cause analysis of pump failures should use FMEA and fishbone diagrams.	Which methods are recommended for root cause analysis of pump failures?	['A) FMEA and fishbone diagrams', 'B) Pareto charts and SPC', 'C) Design of experiments', 'D) SWOT analysis']	A	https://www.nrc.gov/docs/ML1122/ML112231412.pdf
IEEE Std 43-2013 provides guidelines for insulation resistance testing of rotating machinery to detect winding deterioration.	Insulation resistance testing is used to detect winding deterioration in rotating machinery.	What is the primary purpose of insulation resistance testing in rotating machinery according to IEEE Std 43-2013?	['A) Measure vibration levels', 'B) Detect winding deterioration', 'C) Assess thermal performance', 'D) Calibrate speed sensors']	B	https://standards.ieee.org/standard/43-2013.html
The standard recommends testing at 500 V DC for machines rated up to 1 kV to establish baseline insulation resistance.	500 V DC testing is recommended for machines up to 1 kV to establish insulation resistance baseline.	At what voltage does IEEE Std 43-2013 recommend insulation resistance testing for machines rated up to 1 kV?	['A) 250 V DC', 'B) 500 V DC', 'C) 1 kV AC', 'D) 750 V DC']	B	https://standards.ieee.org/standard/43-2013.html
Insulation resistance values trending downward over time indicate progressive degradation and warrant further investigation.	Decreasing insulation resistance trends over time indicate progressive insulation degradation.	What does a downward trend in insulation resistance values indicate?	['A) Improved insulation', 'B) Progressive degradation', 'C) Stable performance', 'D) Sensor calibration error']	B	https://standards.ieee.org/standard/43-2013.html
Measurements should be taken at regular intervals, such as quarterly or semi-annually, depending on operating criticality.	Insulation resistance measurements should be taken quarterly or semi-annually based on criticality.	According to IEEE Std 43-2013, how often should insulation resistance measurements be performed for critical machinery?	['A) Daily', 'B) Quarterly or semi-annually', 'C) Every five years', 'D) On failure only']	B	https://standards.ieee.org/standard/43-2013.html
Polarization index (PI) and dielectric absorption ratio (DAR) tests are supplemented to the basic insulation resistance measurement for deeper analysis.	PI and DAR tests supplement basic insulation resistance measurements for deeper analysis.	Which supplementary tests are recommended by IEEE Std 43-2013 for deeper insulation analysis?	['A) PI and DAR', 'B) FFT analysis', 'C) Infrared thermography', 'D) Oil sampling']	A	https://standards.ieee.org/standard/43-2013.html
The report describes deriving failure probabilities on demand using beta distributions and failure rates using gamma distributions.	Failure-on-demand probabilities use beta distributions; failure rates use gamma distributions.	Which distribution models component failure rates in NUREG/CR-7037?	['A) Beta', 'B) Gamma', 'C) Normal', 'D) Exponential']	B	https://www.nrc.gov/docs/ML1109/ML110980205.pdf
Data for component performance estimates came from 1997â€“2007, while data for unplanned shutdown response covered 1987â€“2007.	Component performance uses 1997â€“2007 data; shutdown response uses 1987â€“2007 data.	What time span does NUREG/CR-7037 use for unplanned shutdown response data?	['A) 1980â€“1990', 'B) 1987â€“2007', 'C) 1997â€“2010', 'D) 2000â€“2015']	B	https://www.nrc.gov/docs/ML1109/ML110980205.pdf
Studies indicated that relief valve performance improved since the 1980s and early 1990s.	Relief valve performance improved since the 1980s and early 1990s.	Since when has relief valve performance generally improved according to NUREG/CR-7037?	['A) 1970s', 'B) 1980s and early 1990s', 'C) 2000s', 'D) 2010s']	B	https://www.nrc.gov/docs/ML1109/ML110980205.pdf
Performance data were obtained from the EPIX database and Licensee Event Reports (LERs).	Data sources include EPIX and LERs.	Which two data sources were used in NUREG/CR-7037 for relief valve performance?	['A) EPIX and LERs', 'B) FCFRDB and IRIS', 'C) Manufacturer tests', 'D) DOE reports']	A	https://www.nrc.gov/docs/ML1109/ML110980205.pdf
The methods employed are conventional estimation methods as documented in NUREG/CR-6823.	The report uses conventional estimation methods from NUREG/CR-6823.	Which document provides the estimation methods used in NUREG/CR-7037?	['A) NUREG/CR-6823', 'B) NUREG/CR-6928', 'C) NUREG/CR-5485', 'D) NUREG/CR-5751']	A	https://www.nrc.gov/docs/ML1109/ML110980205.pdf
This handbook provides methods for estimating parameters used in probabilistic risk assessment (PRA) models and for quantifying uncertainties in the estimates.	The handbook offers methods for estimating PRA model parameters and quantifying uncertainties.	What is the primary objective of NUREG/CR-6823?	['A) Define American Society of Mechanical Engineers (ASME) standards', 'B) Estimate PRA parameters and uncertainties', 'C) Model	B	https://www.nrc.gov/docs/ML0329/ML032900131.pdf

Passage	Fact	Question	Options	Answer	Sources
			fluid dynamics', 'D) Develop control algorithms']		
The data analysis portion provides estimates of initiating event frequencies, component failure rates, unavailabilities, and equipment non-recovery probabilities.	Parameter estimation covers initiating event frequencies, failure rates, unavailabilities, and non-recovery probabilities.	Which parameter is NOT covered by NUREG/CR-6823?	['A) Initiating event frequencies', 'B) Component failure rates', 'C) Equipment non-recovery probabilities', 'D) Thermal conductivity']	D	https://www.nrc.gov/docs/ML0329/ML032900131.pdf
Estimates of risk are obtained by propagating uncertainty distributions through PRA models.	Risk estimates are derived by propagating parameter uncertainty distributions through PRA models.	How does NUREG/CR-6823 suggest obtaining risk estimates?	['A) Monte Carlo heat transfer', 'B) Uncertainty propagation through PRA models', 'C) CFD simulations', 'D) Experimental testing']	B	https://www.nrc.gov/docs/ML0329/ML032900131.pdf
The handbook supplements procedural guides and standards like the ASME PRA standard by providing a compendium of good practices.	NUREG/CR-6823 supplements ASME PRA standards with good practices.	Which standard does NUREG/CR-6823 explicitly supplement?	['A) IEEE 43-2013', 'B) ASME PRA standard', 'C) EPRI TR-102290', 'D) ISO 9001']	B	https://www.nrc.gov/docs/ML0329/ML032900131.pdf
The handbook includes guidance for both plant-specific and generic parameter estimates.	Guidance covers plant-specific and generic parameter estimates.	NUREG/CR-6823 includes guidance for which type of parameter estimates?	['A) Plant-specific only', 'B) Generic only', 'C) Both plant-specific and generic', 'D) None of the above']	C	https://www.nrc.gov/docs/ML0329/ML032900131.pdf
EPRI NP-7502 suggested trending parameters like supply voltage, running current, motor speed, bearing temperature, winding temperature, and vibration.	Trending parameters include voltage, current, speed, bearing temp, winding temp, and vibration.	Which parameter is NOT listed for trending in EPRI NP-7502?	['A) Supply voltage', 'B) Winding resistance', 'C) Bearing temperature', 'D) Motor speed']	B	https://www.epri.com/research/products/1004384
The report influenced a shift from time-based to condition-based motor maintenance strategies.	NP-7502 influenced the shift to condition-based maintenance.	What maintenance strategy did NP-7502 promote?	['A) Time-based', 'B) Condition-based', 'C) Corrective-only', 'D) No maintenance']	B	https://www.epri.com/research/products/1004384
Widespread application of trending many electrical parameters was limited by cost-benefit and evaluation difficulties.	Adoption limited by cost-benefit and difficulty in parameter evaluation.	Why was trending of electrical parameters not widely adopted?	['A) Lack of technology', 'B) Cost-benefit and evaluation difficulty', 'C) Regulatory prohibition', 'D) Insufficient staff']	B	https://www.epri.com/research/products/1004384
The report recommended including insulation resistance and polarization index tests in trending.	Insulation resistance and polarization index tests are recommended for trending.	Which supplementary test did NP-7502 recommend for deeper analysis?	['A) FFT analysis', 'B) Polarization index', 'C) Oil analysis', 'D) Thermal imaging']	B	https://www.epri.com/research/products/1004384
DC hipot (step) testing was also identified as a valuable trending parameter.	DC hipot testing is valuable for trending in motor maintenance.	Which non-invasive test was included as a trending parameter?	['A) Acoustic emission', 'B) DC hipot', 'C) X-ray inspection', 'D) Dye penetrant']	B	https://www.epri.com/research/products/1004384
Corrective maintenance restores equipment that has failed, degraded, or does not conform to design or performance requirements.	Corrective maintenance restores failed or degraded equipment.	What does corrective maintenance involve?	['A) Upgrading software', 'B) Restoring failed or degraded equipment', 'C) Installing sensors', 'D) Reducing inspection frequency']	B	https://www.epri.com/research/products/000000000001022951/0/Product
A maintenance action is considered corrective if the component cannot meet its intended function or will before the next scheduled PM interval.	Corrective maintenance applies when a component fails or will fail before next PM interval.	When is maintenance classified as corrective?	['A) Equipment upgrades', 'B) Component will fail before next PM', 'C) Routine checks', 'D) Scheduled calibration']	B	https://www.epri.com/research/products/000000000001022951/0/Product

Passage	Fact	Question	Options	Answer	Sources
Corrective maintenance may include preventive orders invoked to correct anticipated degradation if criteria are met.	Corrective maintenance can use PM orders to correct anticipated degradation.	Can preventive maintenance orders be used in corrective maintenance?	['A) No, never', 'B) Yes, if degradation criteria met', 'C) Only for electrical systems', 'D) Only for mechanical systems']	B	https://www.epri.com/research/products/000000000001022951/0/Product
Preventive maintenance is scheduled to prevent failure of components before the next PM interval.	PM schedules are designed to prevent component failure before next interval.	What is the goal of preventive maintenance scheduling?	['A) Increase downtime', 'B) Prevent failure before next PM', 'C) Test new equipment', 'D) Reduce maintenance staff']	B	https://www.epri.com/research/products/000000000001022951/0/Product
Corrective and preventive maintenance definitions ensure consistency in maintenance programs.	Clear definitions of corrective and preventive maintenance ensure program consistency.	Why are maintenance definitions important?	['A) To comply with ISO 9001', 'B) Ensure program consistency', 'C) Increase costs', 'D) Avoid regulatory review']	B	https://www.epri.com/research/products/000000000001022951/0/Product

Page intentionally left blank

Appendix B

Example Formative Assessment of Trust

B-1. Introduction

Lack of user trust presents a barrier to adoption for AI/ML technologies in the nuclear industry (Hall & Agarwal, 2024). Trust (a property of the user) and trustworthiness (a property of the technology) must be explored further within the industry to better understand why users may be hesitant to use new technologies to carry out their responsibilities.

Trust is an important but difficult-to-measure concept. In a bid to develop sound, scientific and germane methods to examine user trust for LLM in nuclear, our first step was to conduct a formative assessment in which we explored how people use LLM in their personal and professional lives, factors that might facilitate trust, and any concerns that impacts user trust. For this step we used qualitative methods that produced rich, subjective data. Our second step was to pilot a user trust study of LLM used for predictive maintenance (i.e., the VIPER LLM chatbot introduced in Section 4.3). For this step we used a mixture of qualitative and quantitative methods.

The objectives of this pilot research were the following:

- 1) Provide useful first insights into everyday uses of LLM; namely large language models (LLMs), concerns, and considerations that may influence a user's trust when using LLM tools in the nuclear industry
- 2) Evaluate our methods to capture the effects of chatbot LLM accuracy and rationale (i.e., the model's explanation for its response) on user trust.

B-2. Method

The formative assessment was conducted virtually via Microsoft Teams. One member of the research team shared his/her screen to display a PowerPoint presentation. This was used to present trust questions to the participants as well as to conduct the LLM experiment. Although presented in PowerPoint, the responses and explanations were generated by the LLM itself. Before beginning, the study's purpose was communicated to the participants: to test methods that examine ways in which LLMs can be more trustworthy when supporting tasks for nuclear power plant work. Participants were told that it was an informal pilot study, and that their feedback would help strengthen our methods for future data collection. Team members read from a script to ensure standardization across participants. Testing sessions lasted approximately 1hr with one session taking 1.5 hours.

B-1.1 Participants

Participants were four individuals employed at INL who ranged in age (22-35 years), training, background, and operational experience. One subject was a nuclear engineering undergraduate summer intern with operator experience at a research reactor; one was a graduate intern with a background in nuclear navy; one a post-doc in nuclear engineering; and one was a human-computer interaction visualization expert with training in computer science. We intentionally gathered individuals with a wide range of nuclear backgrounds to help inform individual differences in the assessment of trust in LLM. This research was approved by the INL Institutional Review Board (IRB), and all respondents indicated consent to the research via email agreement (IRB approval INL000245).

B-1.2 Procedure

We began by defining an LLM with participants before going into the semi-structured interview and discussion. This was done to ensure clarity and consistency between the participants and the research team. The definition provided was:

- *“Large language models (or LLMs), a type of generative AI, is an advanced AI system that can understand and generate text like people do, making interactions conversational and natural. LLMs are trained on massive amounts of data to understand, generate, and process human-like language.”*

The team then provided an example to illustrate our definition and mental model of LLM. The example provided was:

- *“You are due for a trip to the grocery store and only have a few items in your refrigerator: chicken breasts, celery, mushroom soup and pepperjack cheese. You input these ingredients into ChatGPT (a type of LLM) and ask it for a recipe.”*

This recipe question was posed to INL’s AIVA software, a chatbot LLM introduced in November 2024 and available to all employees. A real-time AIVA response with the ingredients and recipe was displayed to the participants.

Following the LLM definition and example, we began by grounding the discussion through the participants’ own experiences and uses with LLMs for personal and work circumstances, to encourage them to begin thinking about their own interactions with LLMs. We asked about both personal and professional applications in an attempt to discern whether trust varies depending on the use-context (e.g., higher trust with work applications may seem more risky than personal). We then enquired about use cases, concerns, and trust the participants had (see Table 16). These questions were displayed one at a time.

Table 16. LLM question set

No.	QUESTION
1	Are there any activities you use LLM (LLMs) in your personal life or work?
2	Would you consider yourself a ‘novice’ or ‘power user’ of LLMs?
3	Where are areas that the LLM is not currently applied at work but could be helpful?
4	What concerns do you have using the LLM for these uses?
5	Describe how these concerns impact your trust in the technology

A key objective of this set of questions was to identify which characteristics of the LLM-produced response are most important for user trust and general technology acceptance; this step explored how different LLM uses (e.g., based on the different roles of participants) may create different user needs regarding the qualities of an LLM response when establishing trust and technology acceptance.

In step 2 of the pilot study, we performed a more formalized assessment of LLM trust using an LLM developed as a capability of INL’s VIPER tool for predictive maintenance. The LLM is designed for maintenance operators to interactively engage with the tool by asking technical questions about predictive maintenance, work orders, recommendations and their confidence levels, the kinds of data being used, and the types of ML algorithms employed. An example of the LLM display integrated into the VIPER dashboard can be seen in Figure 29.

As a complement to step 1 in which qualitative data was collected about user trust for an AI tool with which participants were familiar, step 2 captured both qualitative and quantitative measures to evaluate aspects of user trust using experimental methods. Specifically, we examined how convincing the LLM was to the participants as a function of LLM response accuracy and rationale.

A strength of this empirical design is that because none of the participants had interacted with VIPER’s LLM before the experiment, level of engagement was controlled across experimental conditions. This reduced potential bias in the findings such that any observed differences across experimental conditions

could reasonably be interpreted as a function of trust, and not familiarity. The findings from this formative assessment will be used to enhance methods to assess user trustworthiness and trust for human-LLM interactions in the nuclear industry.

For the purposes of this study, we interrogated the LLM's accuracy using a truth-seeking question set developed by a member of the research team. Multiple-choice questions are an effective and efficient way to assess LLM accuracy (or 'truthfulness') because a forced-choice answer can only be right or wrong. An open-ended response might be partially right or wrong to varying degrees. Appendix C documents the full question set from where the questions used in this study were sampled. An example of a question and response displayed to the participants is shown in Figure 41.

LLM Trivia

EXAMPLE LLM Concept 1

Do you already know the answer, or have a hunch? Yes or No

If "Yes", which letter do you think the answer is?

How convinced are you by this explanation?

1 - Not at all convinced
2 - Slightly convinced
3 - Moderately convinced
4 - Very convinced
5 - Extremely convinced

Question

What is the primary objective of equipment condition assessment beyond detecting anomalies?

A. Reducing data storage needs
B. Determining the root cause of problems
C. Automating equipment repairs
D. Scheduling staff shifts

LLM Response

B

LLM Explanation

While anomaly detection is part of equipment condition assessment, the primary objective goes further. Understanding why an anomaly exists and the root cause is crucial for preventing recurrence, optimizing maintenance strategies, and extending equipment life.

Here's why the other options are incorrect:

- A) Reducing data storage needs: Condition assessment often increases data storage needs due to the collection of sensor data and analysis results.
- C) Automating equipment repairs: Condition assessment informs repair decisions but doesn't automatically perform them.
- D) Scheduling staff shifts: Staff scheduling is unrelated to the core purpose of understanding equipment health.

8

Figure 41. Example of multiple choice question, LLM response and explanation.

Four different answers were furnished to the LLM, and it produced two pieces of information:

- 1) The answer it deemed correct (LLM Response)
- 2) Its explanation or rationale (LLM Explanation)

We employed a semi-structured interview format to collect participant ratings about how convincing the LLM's explanations were (using a Likert scale 1-5; see Figure 41) and their thought process in making their decision. If participant's did not think out loud, we followed up with probing questions to capture the reasons that the ratings were given. The purpose of step 2 was to identify the different contextual and experiential factors that participants use to draw conclusions about truthfulness, as well as to gain a better understanding of their mental model when engaging with the LLM tools in professional activities.

The chatbot LLM currently possesses an 87% accuracy score on average. Therefore, we designed each question block so that 4 / 5 LLM responses were accurate for an accuracy score of 80%. We did this to more faithfully represent the LLM's current capabilities, and to gain some preliminary insights into whether ratings would vary as a function of perceived accuracy. The participants were not made aware of the LLM's accuracy score, or that any of the responses would be inaccurate.

The independent variable in this experiment was rationale, which was operationalized as the reasoning or explanation the LLM provided to support its conclusion, presented as information to the user. As noted in Section 7, the way in which an LLM rationalizes its response can vary by many different technical factors (e.g., the way it was fine-tuned, different prompt engineering techniques, or using different types of RAG).

For step 2, we focused primarily on the manipulation of candidate types of RAG approaches in their effect on the rationale or explanation produced by the LLM.

LLM responses that do not use RAG-generated responses are based solely on the internal knowledge of the model (i.e., acquired in its training). The LLM response and explanation shown in Figure xx is an example of a trial that used a non-RAG response (no RAG) – i.e., there is no additional information available beyond the produced explanation. RAG however, is a technique used in AI development that incorporates large bodies of information from an external corpus that is beyond what the LLM learned through training data (Rothman, 2024). A primary advantage of RAG is that the corpus of information used for it is non-parametric, meaning that it is available and can be used directly when producing a response from the LLM, allowing for a more accurate and specific response. This contrasts with parametric knowledge from fine-tuning where the model’s knowledge is stored internally (e.g., through its weights), but the data itself is not readily accessible.

This can include state-of-the-art information from external documents and sources, as well as the internet which delivers more updated knowledge to the LLM. The main benefits of using RAG are LLM response with a higher degree of accuracy and relevancy with respect to the specific context in which the question was asked.

Different database architectures may be used when configuring a RAG pipeline. We were specifically interested in evaluating methods to test the effectiveness of RAG used with vector database (Vector RAG), a common architecture, compared to RAG used with a knowledge graph database (Graph RAG). Figure 42 is an example of a trial in which the LLM used Vector RAG. Compared to the baseline no RAG trials (i.e. Figure 41), source data was available to the participants (pink bar at the bottom of the screen). When requested, the Show Source Data option brought up additional information in a pop-up screen. See sections 7.3.1.1 and 7.3.1.2 for discussions of Vector RAG and Graph RAG, respectively.

The screenshot displays the 'LLM Trivia' interface. On the left, a sidebar contains an 'EXAMPLE LLM Concept 2' section with a question: 'What is the maximum interval for in-service testing of safety-related pumps?' and four multiple-choice options (A: 5 years, B: 10 years, C: 15 years, D: 20 years). Below this is a section for 'How convinced are you by this explanation?' with five radio button options ranging from 'Not at all convinced' to 'Extremely convinced'. The main area shows the 'LLM Response' as 'B' and a detailed 'LLM Explanation' that references a specific document (IEEE Std 43-2013). A pink bar at the bottom of the main area contains a 'Show Source Data' button. A blue arrow points from this button to a pop-up window on the right titled 'Relevant Data Source(s):'. This window displays a large block of text from the IEEE Std 43-2013 document, which discusses insulation resistance testing procedures and the effects of temperature and moisture on insulation resistance measurements.

Figure 42. Example of source data the LLM deemed relevant for the Vector RAG condition.

Graph RAG builds on Vector RAG by incorporating graph structures into the retrieval and generation process and responses typically generate a great deal more detail than Vector RAG or non-RAG responses. Figure 43 demonstrates an example of a Graph RAG trial – each node of the graph on the right hand side pulls up a pop up (black box) that contains several individual explanations that inform the overall explanation.

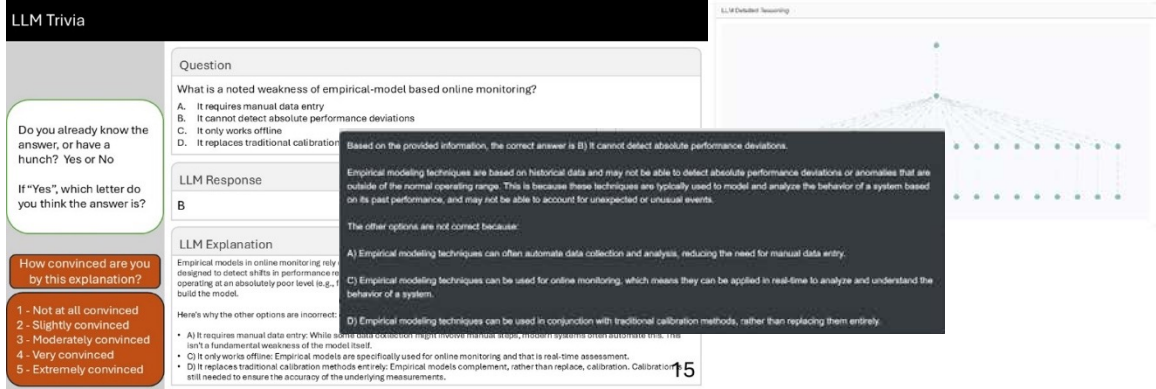


Figure 43. Example of Graph RAG rationale whereby each node represents data sources used by the LLM.

Trials were grouped into three rationale blocks of 5 questions for a total of 15 trials per participant – accuracy remained constant across trials for each block. There were no duplicate questions across blocks. The order of the blocks was: No RAG → Vector RAG → Graph RAG.

Prior to the experiment, participants were briefly introduced to the three experimental RAG conditions and instructed that their task was to rate the LLM explanation. It was strongly stressed to the participants that we did not expect them to know or provide the answers to the multiple-choice questions. However, for each trial, after we showed them the question but before we showed the LLM's response, we asked participants whether they already knew the answer, or thought they might know. We did this to understand whether the participants were using information other than that given by the LLMs in formulating their ratings.

Participants were then presented with the no RAG, Vector RAG and Graph RAG blocks of trials, respectively. At the end of each block, a direct measure of trust was collected in the form of a self-report rating. This was captured on a Likert scale 1 (no trust) – 10 (very high trust).

B-3. Preliminary Findings

Explanation Ratings

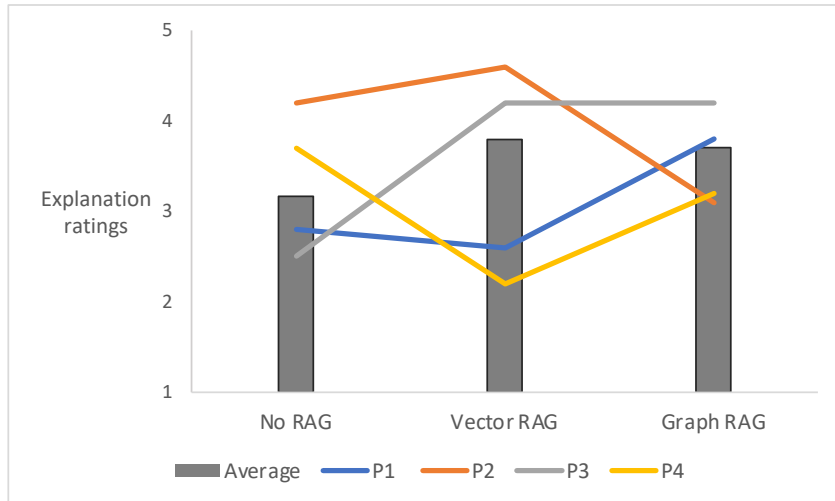


Figure 44. Explanation ratings across RAG conditions.

Figure 44 shows the results to the question “How convinced are you by this explanation?” (using a Likert scale 1-5) across the three experimental conditions. Each of the colored lines represents one of the participants (i.e., P1-P4) and the bar chart represents averaged ratings. The No RAG condition attracted the lowest ratings on average, but scores were more similar than different across conditions. This is because user profiles were distinctly different. Looking at the pattern of results individually, there was little agreement among participants. For example, P2 and P3 rated Vector RAG the highest or joint highest, and P1 and P4 rated it the lowest. In addition, all three RAG conditions were rated highest, and simultaneously lowest, by at least one participant.

Trust Ratings

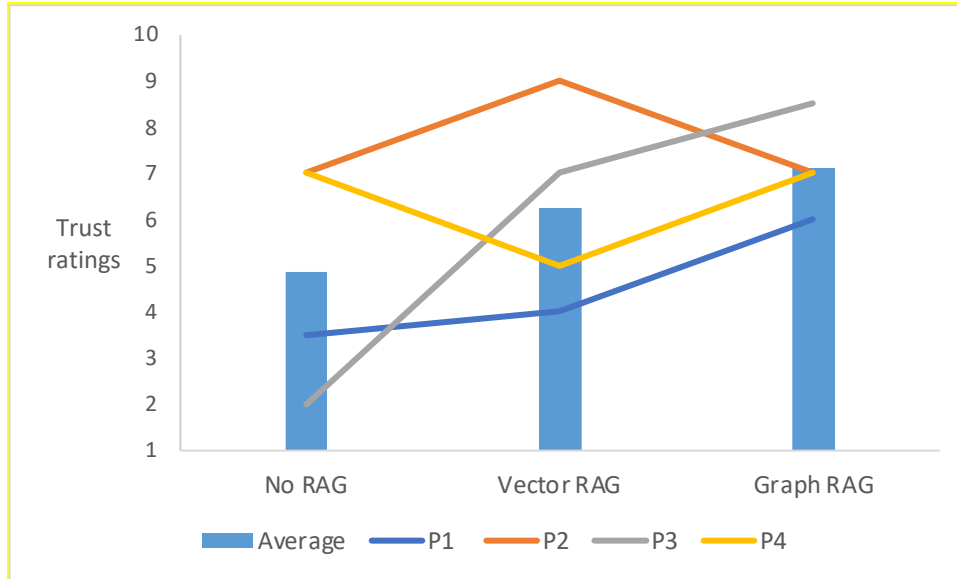


Figure 45. Trust ratings across RAG conditions.

Figure 45 shows the results to the question “How would you rate the trustworthiness of the LLM explanations?” (using a Likert scale 1[no trust] – 10[very high trust]) across the three experimental conditions. For each participant, his/her pattern of results was very similar to the pattern for LLM explanations. As with the explanation ratings, No RAG attracted the lowest scores. Graph RAG attracted the highest scores, although there was little overall difference in trust scores between Vector RAG and Graph RAG. Again, we observed disparate individual user profiles across conditions, especially for the No RAG condition which ranged from 2 to 7. Similarly, while P1 and P3 followed a general ratings increase with each block, P2 and P4 rated No RAG and Graph RAG equally. However, they were misaligned on their perceptions of Vector RAG, with P2 rating it the highest and P4 rating it the lowest.

Time to complete blocks

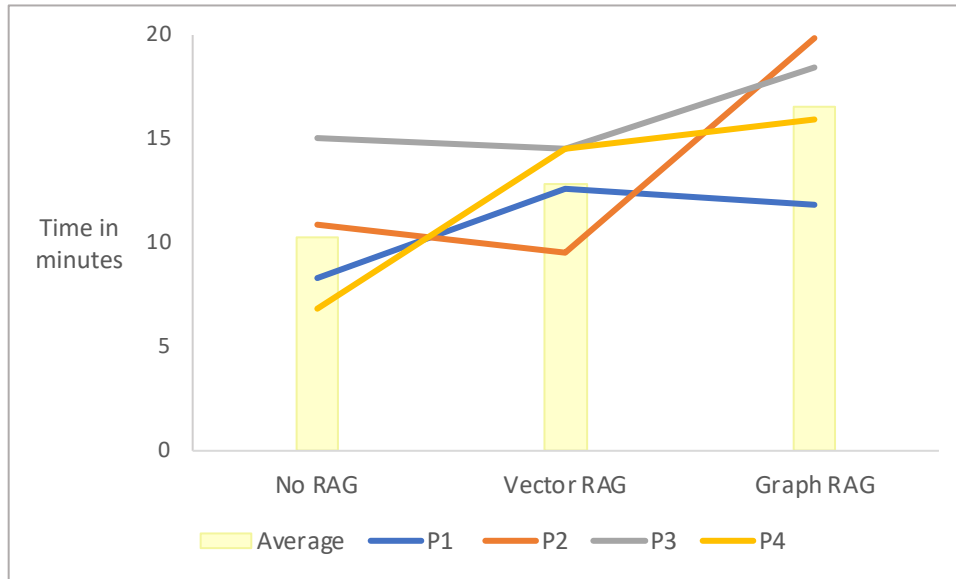


Figure 46. Time in minutes to complete each block.

Figure 46 shows the results of time taken in minutes to complete each block. It must be noted that these times were generated with approximate timestamps and caution is advised when interpreting these results. On average, there was a linear relationship between time taken to complete the blocks and RAG condition. Graph RAG took the longest to complete, which took longer than Vector RAG, which took longer than No RAG. Given the level of additional information to consume with each progressive RAG condition, this finding is unsurprising.

We were interested to know whether LLM accuracy had a bearing on ratings. For one trial in every block, the LLM produced an inaccurate response. Figure 47(a) shows that there were some sensitivities to incorrect responses shown by the participants in that ratings for trials in which the LLM generated the correct responses were higher than when it generated incorrect responses.

We were also interested in whether congruency influenced ratings. That is, when the LLM response (whether correct or not) matched that of the participants'. We hypothesized that congruent trials would increase ratings, because the LLM would match the mental model of the answer provided by the participant. Figure 47(b) shows that there was little difference for the No RAG condition, but that the participants were more apt to rate the LLM explanations higher for the Vector RAG and Graph RAG conditions, if the response matched that of the participant. From a methodological standpoint, this is informative because it may be the case that the participants were basing their ratings on something other than the LLM's explanation. Note - P4 responded "I don't know" for every trial and so was excluded from this analysis.

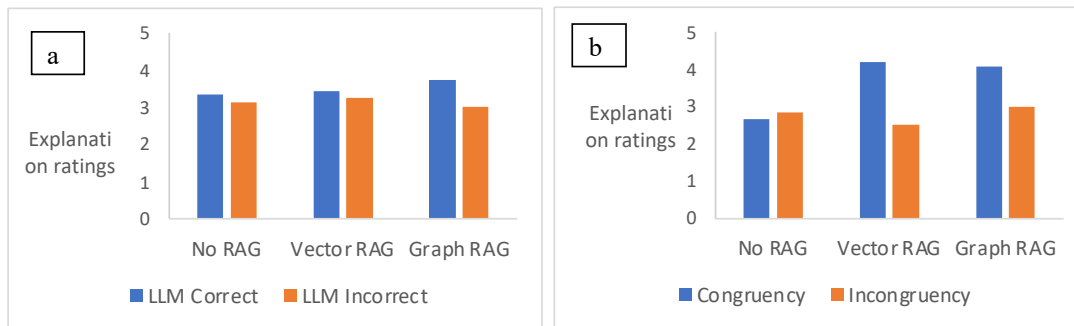


Figure 47. Explanation ratings across LLM correctness (a) and congruency (b).

No RAG feedback

Comments about this LLM centered largely around lacking explanations. Specifically, not enough justification, not countering arguments formed by the participants, being too vague, unsatisfactory, making assumptions or being illogical. A lack of sources, links or citations was perceived negatively and reduced trustworthiness. The participant from nuclear navy mentioned that source citations are the standard in nuclear power operations and gave a trust rating of 2/10.

There was also positive feedback with two participants rating 7/10. Comments included that the explanations for some trials made sense or were intuitive. Additionally, confidence in the LLM increased when citing standards, when definitions were provided or when responses matched question content.

Vector RAG feedback

The addition of source data provided both positive and negatives benefits. While pointing to source documents generally increased convictions, especially if the precise answer could be located, detection of unreliable, incorrect or irrelevant source documents had the opposite effect. In addition, there were perceived disconnects between the explanations and source data – for example, reference to a page number or text that was not apparent in the source data.

Interestingly, there were several instances when participants did not access the source data. Broadly speaking, if participants found what they were looking for in the explanation alone, this was deemed reliable enough and the source data unnecessary. For example, seeing direct quotes in the explanation that matched the question exactly was perceived favorably. There were several comments that the source data was difficult to read and it was hard to locate relevant information, with one participant largely dismissing it for this reason.

GraphRAG feedback

Positive comments included liking the tree structure to track the summary statement from a source, the nodes giving more understanding, being well-defined, believable, and presented in a more digestible format (compared to the source data from the Vector RAG condition). For the most part, the nodes increased ratings from the explanation alone. For the incorrect trial, the nodes communicated uncertainty and a lack of confidence which generally increased trust, because it was perceived as ‘honest’. In addition, converging nodes increased confidence (i.e., when they communicated the same or similar information). Similar issues to Vector RAG included questions about the quality of the source information. Additionally, the nodes often used the terminology “intermediate answers,” and this confused participants.

It is interesting to note that despite attracting the highest trust rating, the GraphRAG condition took the longest to complete. This reflects the accuracy/ time trade-off – i.e., one purpose of LLMs is to provide time efficiencies but use time may necessarily have to increase if trustworthiness increases. GraphRAG

may reduce labor necessary to find the correct answer, but it adds a verification component which adds time.

Results Summary

All four participants engaged with the LLM in either their professional and / or personal lives. Two considered themselves LLM “novices”, one a “power user” and one in the middle. Most reported they used the LLM to support professional writing tasks: e.g., grammar checks and requesting edits, and as a search engine to provide a starting point for an unfamiliar topic, or if stuck. Other uses included help with coding e.g., Python or C++. There were calls for the LLM to be used at work as a search engine and to support querying the content of visualizations.

There were several concerns about the LLM that participants were using in their personal and professional lives that impacted trust. Interestingly, each participant had different concerns, however, one converging theme was problems verifying the responses either because reference links are not provided (e.g., as with INL’s AIVA) or they are incorrect or misaligned. Other concerns included becoming reliant on Gen AI, hallucinations, and not knowing the quality of the data sets used to train the models. Last, there was a concern about inadvertently giving ones intellectual property over to the the LLM via enquiries that the model would then train on.

Overall, there was a healthy skepticism of the LLM and most pilot subjects indicated that they take part in additional verification steps of the LLM’s response. Two participants indicated that it’s simply a tool and tools can be inaccurate – i.e., the LLM is not (yet) trustworthy to the point of human replacement. One participant indicated very low trust in Gen AI.

B-4. Discussion: Methodological Lessons Learned

The main goals of this formative assessment were to provide useful first insights into trust concerns about LLMs and evaluate our methods. Here we provide lessons learned that can be carried into future research.

Notably, our sample size was too small to draw meaningful conclusions about the RAG conditions, and correspondingly we could not perform inferential statistics. Future studies should aim for at least 30 participants. However, a strength of our formative assessment was the diversity of the sample in terms of background, training and experience. As can be observed in our findings, no two participant profiles were alike, and we learned that what individuals are looking for in terms of trustworthiness varies widely.

Other recommendations are that the source data for the Vector RAG be put in a more digestible format. There were complaints that there was not enough time to read the “data dump”. Using software to capture button clicks and timestamps would be hugely beneficial to data collection, as would allowing for more interaction with the LLM.

Notwithstanding, participants appreciated the trials where the process of elimination was reflected in the explanation, and the LLM automatically ruled out the other options (this was the case for most of the trials across blocks). In addition, when the explanation directly referenced language in the question, this provided reassurance and ratings tended to be higher. Another positive marker for trust was when the LLM arrived at its decision in a similar way to the participant (i.e. shared mental model). These are all considerations that can be used for future trust studies.

Employing multiple-choice questions was an experimental design choice that allowed us to control for LLM truthfulness, which is an integral component of user trust. Notwithstanding, it’s possible that this question format lowered internal validity (i.e., our findings may have been a function of personal beliefs and not LLM RAG condition). This is because the question format invited deductive reasoning, and for most trials participants would attempt to answer the questions themselves. Answers were formulated based on logic, recognition memory, and some prior knowledge such that participants may have benchmarked the

LLM's response and explanation against their own logic, which may have influenced ratings. Figure 47b suggests this may have been the case. Using multiple-choice questions also lowers ecological validity because it is unlikely nuclear power plants operators would query the LLM using this format. It is more likely they would ask an open-ended question. However, for experiments that employ open-ended questions, this makes interpretation of truthfulness more difficult to concern.

One strength of the current body of research is that we used a mixed-methods approach analyzing AI trust at the usability and experimental level, with both forced-choice and open-ended questions. Together, this provided a comprehensive analysis that allowed the research team to understand how users of AI in nuclear power experience the technology in their own words, and through experimental methods, draw conclusions about factors that facilitate trust in AI across individuals.

Page intentionally left blank

Appendix C

Example Summative Assessment of Trust

C-1. Introduction

There has been extraordinary growth in the development and use of AI. LLM, namely LLMs, is one of the most recent advancements that offers significant potential in automating everyday tasks that demand text generation, translation, summarization, code generation, and writing, to name a few. LLMs are being used across many industries, with some being safety-significant, such as nuclear electricity generation (Al Rashdan et al., 2024). While such technology has significant potential in transforming the way that businesses, organizations, and industries perform work, LLMs are not without their limitations. Currently available LLMs can hallucinate, where the response (i.e., their output) is neither factually accurate nor logically sound.

Hallucinations are a major threat to an LLM's trustworthiness. An untrustworthy response may appear logically sound to the user at face value, and traceability of a response to an original data source may be impossible to obtain. The potential consequences of inappropriately trusting an LLM can have deleterious consequences, such as misusing unreliable LLMs and disusing reliable LLMs (Parasuraman & Riley, 1997), especially when it is being used for critical infrastructure, such as in operating or maintaining a nuclear power plant.

Further, it remains relatively unexplored what psychological mechanisms underlie the process of trust development, maintenance, and recovery through dynamic interactions between human operators and LLMs in nuclear power plant operation tasks. In the context of LLMs, trustworthiness may be better characterized by various evaluation criteria being developed for LLMs such as relevance, hallucination, and question-answering accuracy, beyond the attributes identified in the TrustLLM study (Huang et al., 2024). To improve LLM performance, RAG allows "crosschecking" answers an LLM is generating internally with an external knowledge base (e.g., accident and condition reports) to provide accuracy verification and avoid hallucinations and privacy violations. Additionally, with RAG, human operators may access levels of "confidence" for each output from an LLM.

To examine the impact of representing confidence of RAG-powered AI on human trust and performance in the nuclear power plant context, we designed and executed a human-subject experiment. In the experiment, participants recruited from a community of Norfolk, Virginia, were asked to determine whether each statement is true or false based on LLM's response in a condition where its confidence, according to the fact score, was represented by color-coding each sentence compared to the control condition with no highlights. We predicted that participants would perform more accurately and quickly on the decision-making task with the highlights than without, because they would integrate RAG-based AI outputs into their decision-making more efficiently by selecting relevant information more rapidly.

C-2. Method

C-2.1 Participants

Eighteen participants ($M = 28.06$ years old, $SD = 6.58$ years; 11 female, 7 male, 1 other) were recruited from the community of Old Dominion University, Norfolk, Virginia. All the participants reported normal or corrected-to-normal visual acuity (near acuity = 20/26.67; far acuity = 20/23.89) and normal color perception. Table * presents the participant demographics information in each condition.

Table 17. Participant demographics information for each condition.

	No Color-Coding (Control Condition)	Color-Coding/ Highlight (Experimental Condition)	<i>t</i> -value	<i>p</i> -value
Age	26.55	29.55	.96	.34
Gender	4 Females, 4 Males, 1 Other	7 Females, 2 Males	N/A	N/A
Visual acuity (near)	28.88	24.44	2.13	.048
Visual acuity (far)	26.11	21.66	1.80	.09
AI Literacy	13.77	13.77	0	1
Need for Cognition	87.88	90.33	.34	.73

C-2.2 Apparatus

A Samsung T24C550 23.6” LED monitor (1920 x 1080 pixel) with a refresh rate of 60 Hz was used to present the stimuli. A Dell Precision 3460 computer was used to run a custom program in PsychoPy. Participants’ eye movements were recorded using a Tobii Nano eye tracker with a sampling rate of 60 Hz. The experiment was conducted in a quiet room with dimmed light.

C-2.3 Stimuli

A total of 52 stimulus images were rendered by systematically combining the bot’s response with or without RAG-based information references and target questions to be answered by each participant (Figure 48).

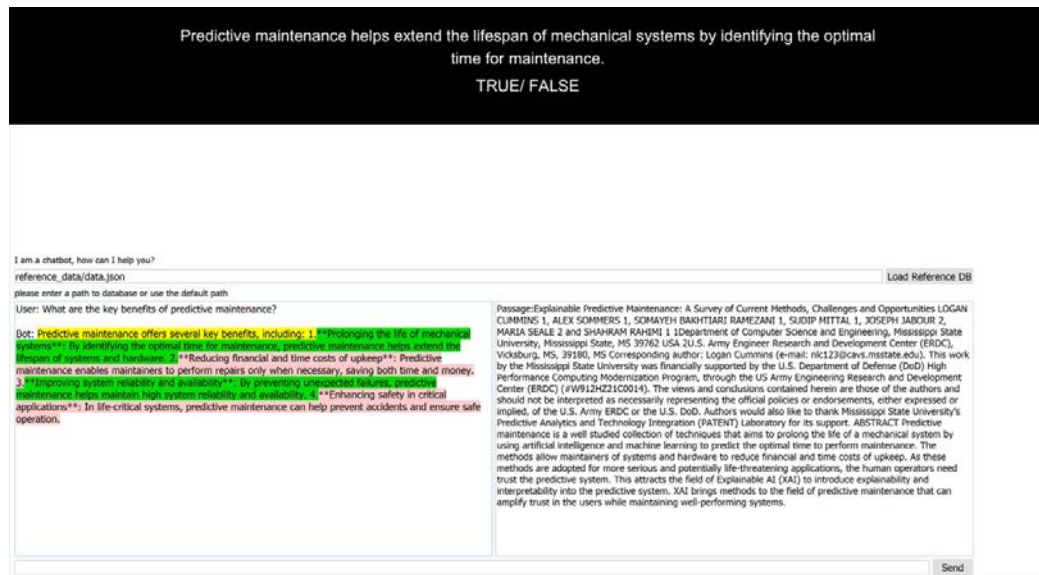


Figure 48. Example stimuli of color-coding/ highlight (experimental condition).

A total of 13 prompts regarding the circulatory water system were entered into VIPER, and its response to each prompt was recorded. From these prompts, two atomic facts were randomly sampled, and an equal number of true/ false statements were generated for the experiment (26 per condition). Each sentence generated by the bot was cross-checked for accuracy and precision, objectively quantified using the

FActSCORE⁵, ($M = .414$, $SD = 0.323$). Each sentence generated in the display was color-coded according to the respective fact score for the highlight condition while the color-coding was not available for the no-highlight condition.

C-2.4 Procedure

Participants first reviewed the notification statement and verbally and voluntarily agreed to participate in the study, followed by visual acuity using the Snellen chart and color perception using the Ishihara color blindness test. The participants then completed the need for cognition questionnaire (Appendix C-1) and the AI literacy questionnaire (Appendix C-2). The participants received brief introduction to the mechanism of the Circulatory Water System (CWS) and the basics of regulations on the operation of the system on PowerPoint slides. After each participant was randomly assigned to either the highlight or no-highlight condition, the experimenter calibrated the eye tracker. The participants read responses by a simulated LLM and were asked to decide whether each target statement is correct or incorrect in reference to its atomic facts provided by the LLM. Each participant completed 10 practice trials followed by 26 experimental trials. Upon the completion of the experimental trials, the participants were asked to complete questionnaires -- Jian et al. (2000) trust scale, Chancey et al. (2017) trust scale, National Aeronautics and Space Administration-Task Load Index (NASA-TLX; Hart & Staveland, 1988), perceived risk (Simon, Houghton, & Aquino, 2000), and technology acceptance and behaviors. The participants were debriefed, thanked, and dismissed.

C-2.5 Methodological Design

The present experiment employed a between-subject design with Highlight as the between-subject factor.

C-2.5.1 Dependent Measures

The dependent variables were the two signal detection performance measures – sensitivity and response bias from signal detection theory (SDT) to assess trust, self-reported trust, percent dwell time, perceived risk, and technology acceptance and behaviors.

In addition to using SDT, trust was measured using a modified version of Jian et al.'s (2000) trust questionnaire (Appendix C-3) and Chancey et al.'s (2017) trust questionnaire (Appendix C-4). Jian et al. (2000) trust questionnaire (Cronbach's $\alpha = .93$; Gutzwiller et al., 2019) consists of 12 items on a 7-point Likert scale (1 = not at all, 7 = extremely) which measure trust and distrust (reverse-coded) to automation. Chancey et al. (2017) trust questionnaire (Cronbach's $\alpha = .97$; Sato, 2020) comprises of 13 items on a 12-point Likert scale (1 = not descriptive, 12 = descriptive) grouped into three informational bases of trust – performance-, process-, and purpose-based trust. A recent work indicates that these two questionnaires measure different facets of human-automation trust (Yamani et al., 2024).

A basis for using SDT is discussed next.

C-2.5.2 Background – Challenges in Evaluating Trust in LLMs

There are notable challenges in evaluating trust with generative AI like LLMs. For instance, trust is a multifaceted construct, which can be considered a belief, an attitude, an intention, or a behavior. In general, trust is a psychological construct that is known to influence behavior (Chancey et al., 2017), human performance (Sato et al., 2020), attitude toward technology acceptance (Kovesdi, 2021), and human-automation teaming.

⁵ A discussion of the use of FActSCOREs are provided in the Methodological Design section, presented later in this paper.

An in-depth review of the literature pertaining to all of the characteristics that comprise trust in automation-based systems (i.e., including AI) is beyond the scope here. Though, a pertinent concept is trust calibration (Lee & See, 2004), which signifies a degree of appropriate user trust given the trustworthiness of the AI system (i.e., LLM). The relationship between user trustworthiness and trust of the AI system, such as an LLM, is represented in Figure 49(A), with the diagonal dotted line representing appropriate trust calibration.

Within Figure 49, Points (A) and (D) represent a poor calibration of trust given the trustworthiness of the LLM. In the context of information produced by an LLM, Point A would characterize a user who trusts the information produced by the LLM when the information is in fact not trustworthy (i.e., neither factual nor logical). Depending on the situation, a consequence of this unjustified trust could result in a threat to safety or economic loss due to the incorrect use of the information (e.g., a misdiagnosis of a plant fault). Unjustified distrust (Point D) could also result in similar consequences. Ideally, the user should appropriately trust the LLM when it is trustworthy and distrust it when it is not trustworthy.

Another notable challenge, particularly relevant to this work, is that evaluating aspects of an LLM's trustworthiness is not a straightforward endeavor. As highlighted by Martell, Baweja, and Dreslin (2024), the correctness of responses generated by the LLM are not always a simple binary correct or incorrect response. Given the rich text-based responses that are generated by LLMs, the "correctness" of a response may reside on a continuum. That is, a given response may include several factual statements that are each dependent on its context surrounding a user's prompt.

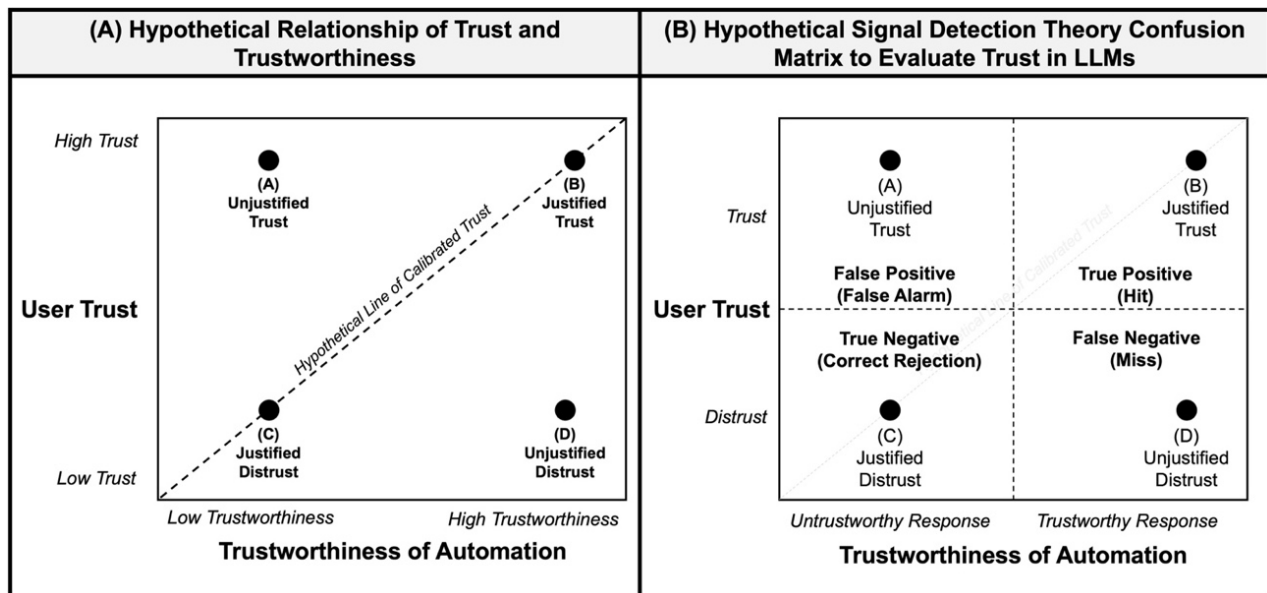


Figure 49. Trust and Trustworthiness Characterization for Evaluating LLMs.

The second challenge in evaluating trust is that evaluating LLM trustworthiness is not a straightforward endeavor. As highlighted by Martell, Baweja, and Dreslin (2024), the correctness of responses generated by the LLM are not always a simple binary correct or incorrect response. Given the rich text-based responses that are generated by LLMs, the "correctness" of a response may reside on a continuum. That is, a given response may include several factual statements that are each dependent on its context surrounding a user's prompt and the relevant knowledge source and differentially contribute to the trustworthiness of the response or the LLM itself.

Evaluating trustworthiness in the entirety of a given response produced by an LLM is not always possible. One approach to address this challenge is to decompose an LLM's response into the smallest unit of factual information that can be extracted: an atomic fact (Kriman, 2024). Atomic facts provide a level of

granularity that enables a binary comparison of its factual correctness based on a given corpus of text information. Measures of the correctness of an LLM’s response can thus be generated using aggregate measures of the atomic facts scored within a given response. For example, the FactSCORE is one such aggregate measure of factual precision that accounts for the percentage of atomic facts supported by a given knowledge source (Min et al., 2023).

Here, we propose that this same approach may be applicable when evaluating a user’s trust in a given response, thus allowing for a more direct comparison of human trust in an LLM and the trustworthiness of the LLM to determine whether trust is justified or not. Specifically, we explore the application of signal detection theory (SDT) for this evaluation, which is described next.

C-2.5.3 Background – Prospective Approaches to Address Challenges

Evaluating the factual correctness of an LLM’s response, an important attribute of its trustworthiness, in its entirety for a given response is not always possible. One approach to address this challenge is to decompose an LLM’s response into the smallest unit of factual information that can be extracted: an atomic fact (Kriman, 2024). Atomic facts provide a level of granularity that enables a binary comparison of its factual correctness based on a given corpus of text information. Measures of the correctness of an LLM’s response can thus be generated using aggregate measures of the atomic facts scored within a given response. For example, the FactSCORE is one such aggregate measure of factual precision that accounts for the percentage of atomic facts supported by a given knowledge source (Min et al., 2023).

Here, we propose that this same approach may be applicable when evaluating a user’s trust for a given response, thus allowing for a more direct comparison of human trust and the trustworthiness of the LLM to determine whether trust is justified or not. Specifically, we explore the application of SDT, as discussed in Sheridan (2019) for this evaluation of trust in LLM use.

C-2.5.4 Use of Signal Detection Theory for Evaluating Trust

Here, we discuss how SDT can be used to evaluate trust in LLMs. An overview of SDT and its metrics is first provided, followed by a discussion of how these metrics may be used to assess trustworthiness and trust in human-LLM interactions.

C-2.5.4.1 Background to SDT

The characterization of signal + noise and noise is represented in a theoretical sense by two respective distribution pairs, such as those in Figure 50.

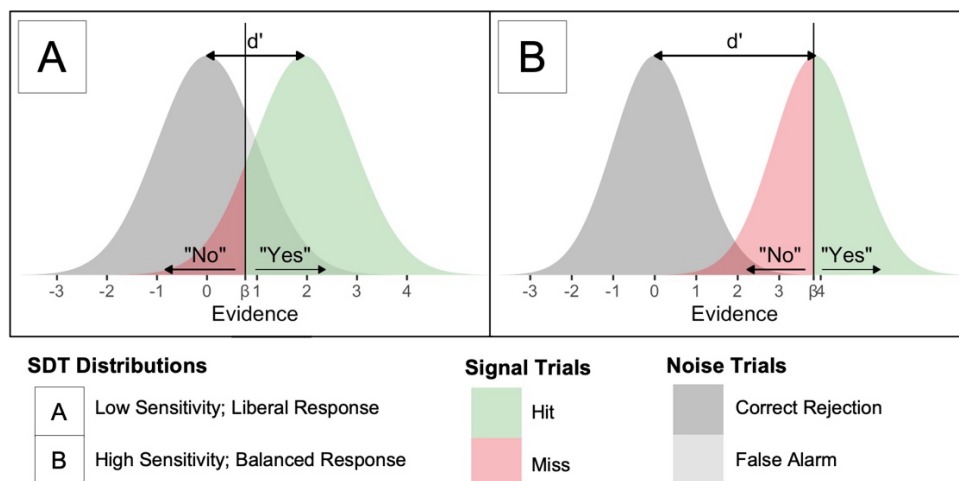


Figure 50. Hypothetical signal and signal + noise distributions using SDT as a framework.

The ability of a human to discern the signal from noise may vary by the human's ability to detect the signal resulting from the signal's salience (sensitivity) or by the human's threshold level for evidence accumulation to render either yes or no response (response bias). The noise (dark gray + gray) and signal (red + green) distributions in Figure 50(A) overlap more, indicating the sensitivity of the human is lower than that in Figure 50(B).

Thus, the human should be able to discern a signal more easily from (B) than (A), with all other variables fixed. The response bias that there is a signal is indicated by the black vertical line in Figure 50. Here, (A) illustrates a more liberal judgement in responding to a signal when compared to (B). The tendency of responding liberally or conservatively is indicated by the response bias.

Importantly, for (A), the portion of the noise distribution that is left of the response criterion (dark gray area) represents the rate of correct rejections (CRs). The portion of the noise distribution right of the response criterion (light gray area) represents the rate of false alarms (FAs). The portion of the signal distribution that is left of the response criterion (red area) represents the rate of misses (Ms). Finally, the portion of the signal distribution that is right of the response criterion (green area) represents the rate of hits (Hs). These response characteristics are used to comprise SDT metrics for sensitivity and response bias, such as those in Stanislaw and Todorov (1999).

Sensitivity. Sensitivity, as indicated by d' , is expressed by the following:

$$d' = \Phi^{-1}(Hs) - \Phi^{-1}(FA) \quad (3)$$

Here, the difference between the one-tailed z score for hits, $\Phi^{-1}(Hs)$, to the one-tailed z score for false alarms, $\Phi^{-1}(FA)$, denotes d' . This measure is often considered parametric in nature, indicating that it assumes a normal distribution of signal + noise and noise. When data deviate from a particular probability distribution (e.g., normal distribution), nonparametric measures could be used such as A' , as expressed as the following:

$$A' = .5 + \left[\frac{(Hs-FA)(1+Hs-FA)}{4Hs(1-FA)} \right] \quad (4A)$$

A' is expressed as the following when there are fewer hits than false alarms:

$$A' = .5 - \left[\frac{(FA-Hs)(1+FA-Hs)}{4FA(1-Hs)} \right] \quad (4B)$$

Response Bias. The parametric measure of response bias is denoted as β , which is expressed as:

$$\beta = e^{\left\{ \frac{[\Phi^{-1}(FA)]^2 - [\Phi^{-1}(Hs)]^2}{2} \right\}} \quad (5)$$

One nonparametric measure of response bias is denoted as B'' , which is expressed as the following when there are at least as many hits as false alarms:

$$B'' = \left[\frac{H(1-H)-F(1-FA)}{H(1-H)+F(1-FA)} \right] \quad (6A)$$

B'' is expressed as the following when there are fewer hits than false alarms:

$$B'' = \left[\frac{FA(1-FA)-H(1-H)}{FA(1-FA)+H(1-H)} \right] \quad (6B)$$

The application of these measures can be performed in common statistical software and more details can be found in Stanislaw and Todorov (1999).

C-2.5.4.2 Use of SDT for Evaluating User Trust

We position that the extension of SDT discussed by Sheridan (2019) to evaluate trust is fitting for evaluating trust with LLMs. The trustworthiness of the LLM may be indicative of the characteristics of the signal + noise distribution relative to noise distribution within the SDT framework. At the most ideal state, the LLM would produce all factual statements (i.e., no hallucinations) to which there is only signal, absent of noise. More likely, however, the LLM may produce seemingly factual statements that are actually non-factual, or hallucinate. In such cases, the user must discern factual information (signal) from non-factual information (noise).

Further, a user may be more apt to trusting the LLM regardless of its grounding in truth for a variety of reasons (see Yamani et al., 2024), including internal factors (e.g., their knowledge of the domain or former AI experience) or external factors (e.g., time pressure). Response bias in this sense would measure a user's predisposition and/or shift in their decision criterion during the interaction with the LLM to trusting the LLM independent of the LLM's conspicuousness in communicating its confidence to the user. A user who is apt to trusting the LLM regardless of whether the information provided is factual would be characterized by Figure 50(A). A user who is more conservative in trusting the LLM would be characterized by Figure 50(B). Next, we discuss specifically how SDT may be used in assessing user trust following an LLM's response.

C-2.5.4.3 Application of Signal Detection Theory in Evaluating Trust

Sheridan (2019) proposes three models that can be applied to analyze trust in automation, with one being SDT. For this work, we propose using SDT as an evaluative framework for comparing trustworthiness and trust when interacting with LLMs. In doing so, the first step is to parse an LLM's response into the atomic facts that are represented. For example, Figure 51 illustrates a user prompt followed by a response provided from an LLM. Here, the response provided by the LLM contains two atomic facts, which are colored in green and in yellow to differentiate the two, respectively. The review of these two atomic facts in the example represents the smallest useful unit of factual information that can be used to evaluate the factual correctness of a response generated by an LLM.

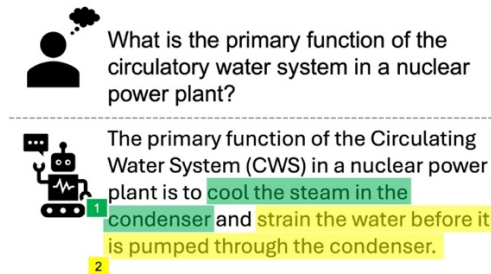


Figure 51. Illustration of atomic facts in an LLM response.

Next, a given atomic fact is compared in terms of its factual correctness to its given knowledge source. While an aggregation of these comparisons of atomic facts can be made (e.g., developing FActSCOREs), the application of SDT focuses on the individual atomic facts. Using SDT, each atomic fact is paired with a corresponding statement of the fact, allowing a user to respond as to whether the user believes the atomic fact generated by the LLM to be true (factual) or false (not factual). Figure 52 illustrates this comparison using the first atomic fact from Figure 51.

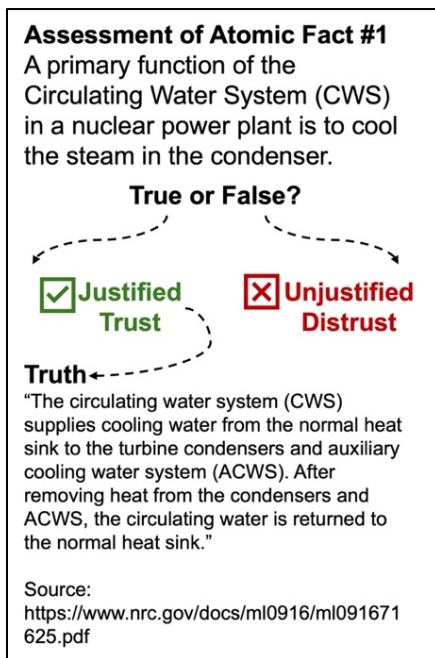


Figure 52. Evaluation of a single atomic fact.

The result is a direct comparison of discrete outcomes of the factual correctness (an important aspect of trustworthiness) and user's trust as indicated by their response, as seen in Figure 49(B). In the SDT framework, therefore, users make binary decisions to either trust or not trust a response by LLMs, allowing researchers to characterize the degree of the users' trust calibration by constructing the SDT confusion matrix.

Looking at Figure 51, the assessment of the atomic fact ("A primary function of the Circulating Water System (CWS) in a nuclear power plant is to cool the steam in the condenser") produced from the LLM could have one of two outcomes:

- **Outcome 1 | True Atomic Fact.** The user believes the atomic fact produced by the LLM is true; in which case, the atomic fact is factual. This results in justified trust (i.e., hit) as seen in the top right portion of Figure 49(B).
- **Outcome 2 | True Atomic Fact.** The user believes the atomic fact produced by the LLM is false; however, the atomic fact is indeed factual. This results in unjustified distrust (i.e., miss) as seen in the bottom right portion of Figure 49 (B)

Conversely, if the atomic fact of interest was not factual (false), the possible outcomes of user trust would entail:

- **Outcome 1 | False Atomic Fact.** The user believes the atomic fact produced by the LLM is false where the atomic fact is not factual. This results in justified distrust (i.e., correct rejection) as seen in the bottom left portion of Figure 49(B).

- **Outcome 2 | False Atomic Fact.** The user believes the atomic fact produced by the LLM is true, but the atomic fact is actually not factual. This results in unjustified trust (i.e., false alarm) as seen in the top left portion of Figure 49 (B).

An evaluation of a user's trust and LLM's trustworthiness using SDT as a framework entails collecting a sufficient sample of individual atomic fact comparisons comprised by multiple prompt-LLM response exchanges, using such protocol described here.

C-3. Results

Sensitivity

Participants who received RAG-based output without color-coding showed numerically greater levels of sensitivity ($M = 3.55$) than those with color-coding ($M = 2.22$) but the difference did not reach the statistical significance [independent-samples $t(16) = 1.504$, $p = .152$, $95\%CI = [-3.22, .547]$, $B_{10} = .874$]. Figure 53 presents the mean sensitivity in each group.

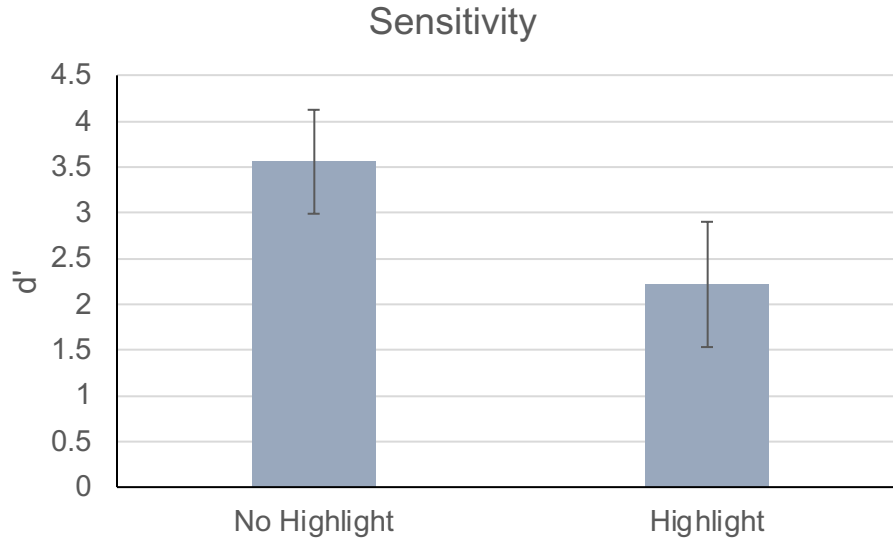


Figure 53. Mean sensitivity in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

Response Bias

Both the No-Highlight and Highlight groups showed significantly liberal response bias settings [$M = -1.779$, one-sample $t(8) = -6.280$, $p < .001$, $B_{10} = 135.878$ for the No-Highlight condition; $M = -1.111$, one-sample $t(8) = -3.242$, $p = .011$, $B_{10} = 5.564$ for the Highlight condition], suggesting that they are more likely to decide that each prompt is factually correct. The response bias setting, however, was statistically comparable between the two conditions [independent-samples $t(16) = 1.504$, $p = .152$, $95\%CI = [-.273, 1.611]$, $B_{10} = .874$].

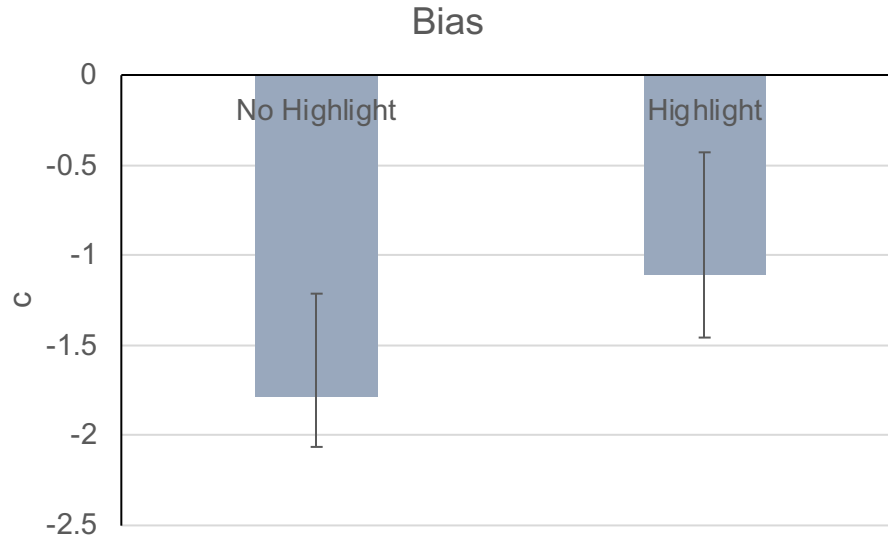


Figure 54. Mean bias in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

Jian et al. Trust

The data indicate no statistically different levels of trust in LLMs between the two conditions [$M = 28.000$ vs. 27.666 for the No-Highlight and Highlight conditions, respectively; independent-samples $t(16) = -.093$, $p = .926$, $95\%CI = [-7.869, 7.203]$, $B_{10} = .412$].

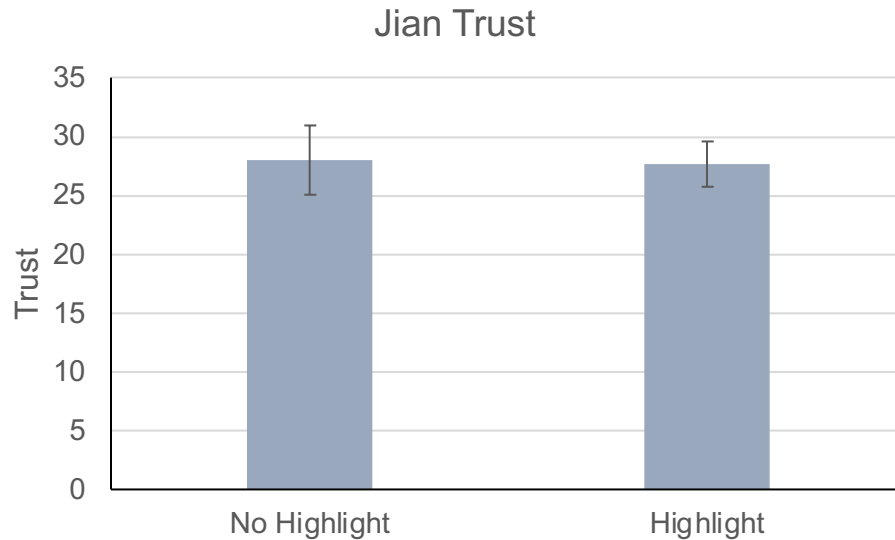


Figure 55. Mean trust measured on Jian et al. (2000) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

On the distrust dimension, participants in the Highlight condition reported numerically greater levels of distrust toward the LLM than those in the No-Highlight condition ($M = 18.777$ vs. 15.777), but the difference was not statistically reliable [independent-samples $t(16) = 1.047$, $p = .310$, $95\%CI = [-3.074, 9.074]$, $B_{10} = .598$].

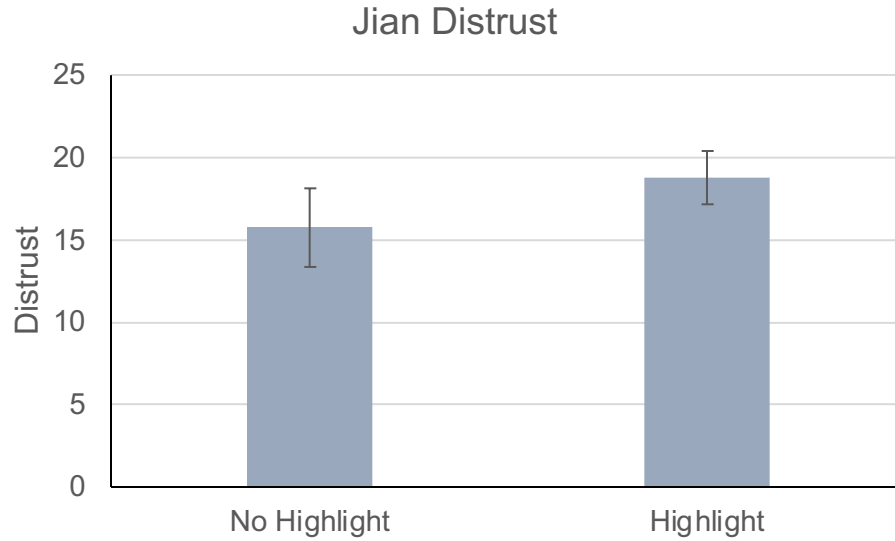


Figure 56. Mean distrust measured on Jian et al. (2000) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

Chancey et al. Trust

Performance-based trust. Participants in the Highlight condition reported numerically lower levels of performance-based trust than those in the No-Highlight condition [$M = 35.333$ vs. 28.555], but again this difference was not statistically significant [independent-samples $t(16) = 1.213$, $p = .242$, $95\%CI = [-18.618, 5.062]$, $B_{10} = .677$].

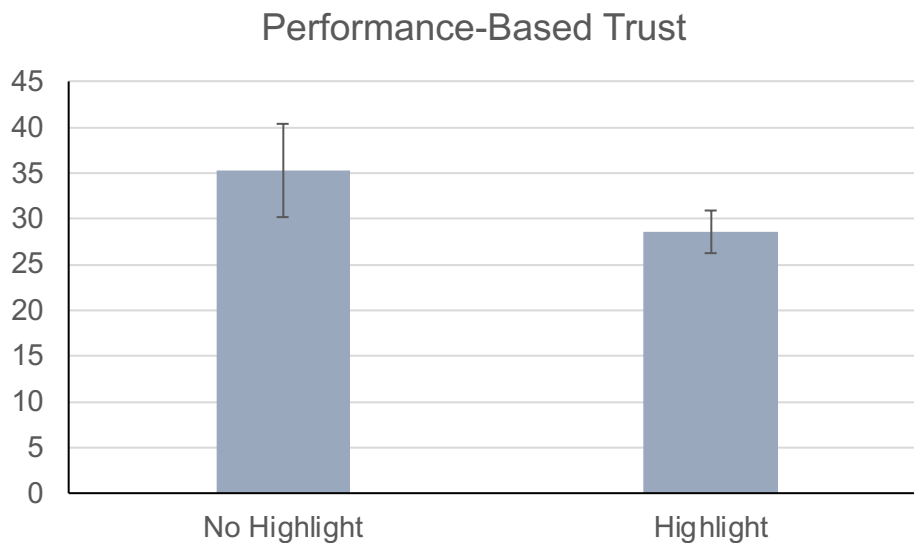


Figure 57. Mean performance-based trust measured on Chancey et al. (2017) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

Process-based trust. Participants reported statistically comparable levels of process-based trust [$M = 33.888$ vs. 36.444 for the Highlight and No-Highlight conditions, respectively; independent-samples $t(16) = .430$, $p = .672$, $95\%CI = [-15.149, 10.038]$, $B_{10} = .438$].

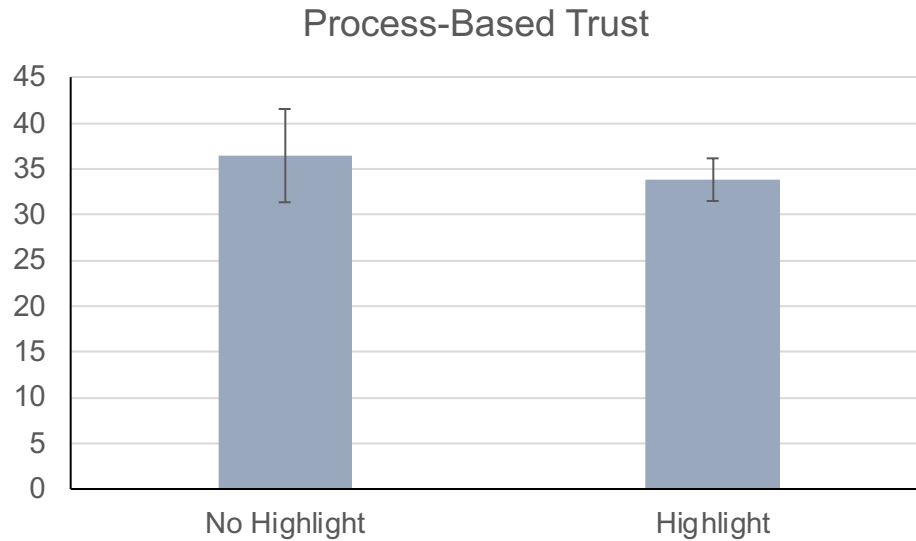


Figure 58. Mean process-based trust measured on Chancey et al. (2017) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

Purpose-based trust. Again, participants in the Highlight condition tended to show lower levels of purpose-based trust than those in the No-Highlight condition [$M = 16.444$ vs. 15.111 ; independent-samples $t(16) = .423$, $p = .677$, $95\%CI = [-8.004, 5.338]$, $B_{10} = .438$].

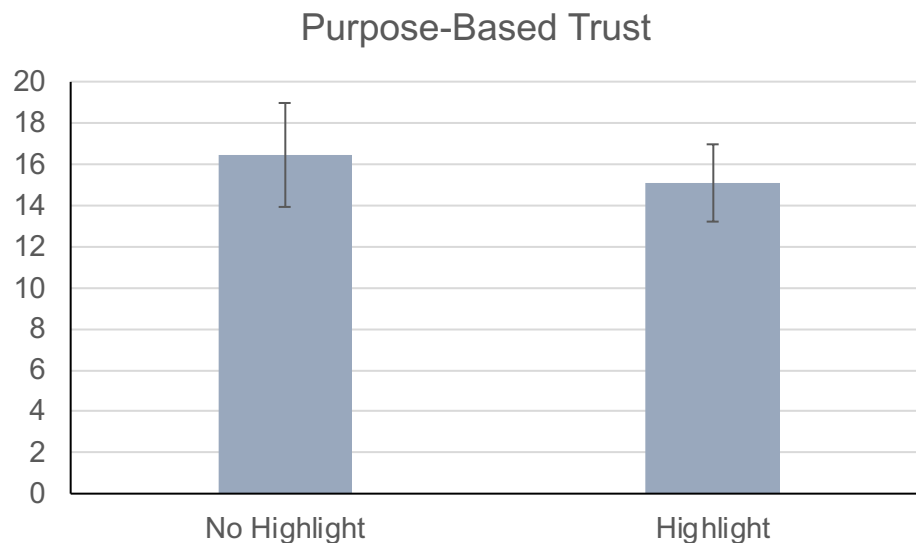


Figure 59. Mean purpose-based trust measured on Chancey et al. (2017) trust scale in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

Subjective Workload

Participants who were presented color-coded outputs from the LLM reported greater levels of subjective workload ($M = 58.888$) than those without color-coded outputs ($M = 50.555$), but the difference did not reach the statistical difference [independent-samples $t(16) = 1.003$, $p = .330$, $95\%CI = [-9.267, 25.933]$, $B_{10} = .580$].

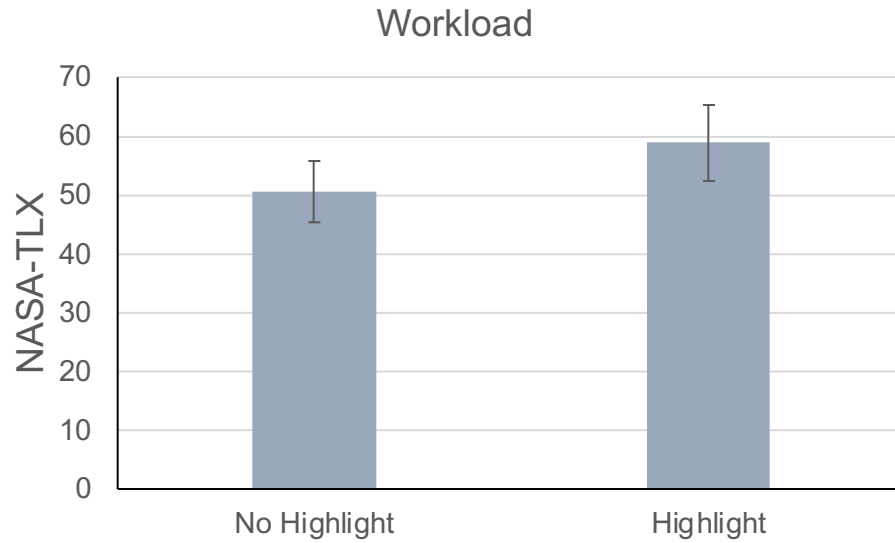


Figure 60. Mean workload measured on NASA-TLX in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

Perceived Risk

Participants reported statistically comparable levels of perceived risk [$M = 27.555$ vs. 25.000 ; independent-samples $t(16) = .386$, $p = .704$, $95\%CI = [-11.475, 16.586]$, $B_{10} = .433$].

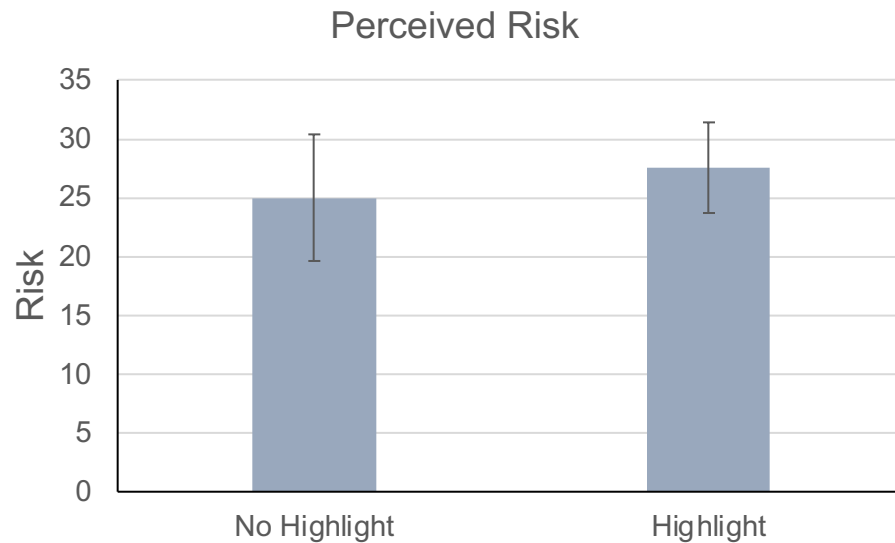


Figure 61. Mean perceived risk scores in the No-Highlight and Highlight conditions. Error bars represent 1 standard error of the mean.

C-4. Discussion

We conducted a human-subjects experiment to examine the impact of color-coding outputs of a mock LLM on decision-making performance and trust toward the LLM. Participants evaluated a series of statements regarding the circulatory water system based on LLM outputs generated with RAG, either with or without color-coded indicators reflecting sentence-level confidence based on fact scores. Overall, experiment participants reported a significantly liberal response criterion setting, tending to answer “yes” more often than “no” in the current task. Contrary to our initial prediction, however, preliminary results revealed that participants exposed to color-coded outputs tended to report lower levels of decision-making performance and trust in the LLM and higher levels of workload. This pattern of data from human subjects suggests that, while LLMs have the potential to enhance operational efficiency when designed and deployed thoughtfully, adding an extra layer of information – such as confidence-based color-coding – may hinder rather than help human users. One possible explanation is that the additional confidence cues using the color-coding mechanism increased cognitive load by requiring participants to mentally integrate both the content and its associated confidence levels during decision-making. Elevated workload levels may reflect this point. This observation aligns with a recent finding that LLM use can reduce neural connectivity as measured on alpha and beta-band neural oscillations, indicating potential cognitive, neural, and behavioral costs (Kosmyna et al., 2025).

Interestingly, despite their interaction with the identical LLM, participants tended to distrust more when the output was color-coded than not color-coded on the Jian et al.’s (2000) trust scale and trusted LLM less on all three dimensions of Chancey et al.’s (2017) trust scale. This “dampening” effect of trust, despite the identical LLM outputs, indicates that trust in LLM may be amenable to changes in characteristics of AI defining its trustworthiness such as transparency, reliability, and explainability as well as compatibility and display explicitness on visual interfaces. These factors are often categorized as external variables in the TAM framework proposed to predict trust in technology.

An alternative explanation for the current findings is a familiarity effect wherein participants may have been unfamiliar with the color-coded LLM outputs and thus potentially struggled to strategically incorporate the new set of information. However, this account alone is unlikely to fully explain the current results because all participants received the identical RAG-based content, allowing them to cross-check the accuracy and reliability of the output regardless of the presence of color-coding. If participants found the color cues unhelpful or confusing, they could have simply ignored them and relied on the same RAG-based content. Therefore, the potentially negative impact of color-coding in this experiment may stem not only from unfamiliarity but also from the increased cognitive demands associated with integrating confidence cues into the decision-making process.

Two approaches may help elucidate the underlying mechanisms in further research. First, participants could be more explicitly trained on the color-coding scheme and its intended function, providing them with a clearer understanding on how specific colors map onto the LLM’s confidence levels. For example, a recent study by Spatharioti and colleagues (2025) demonstrated that a color-coded highlighting system improved decision accuracy in an information search task supported by an LLM with variable reliability along with its color mapping (green for “high confidence” and red for “low confidence”) visible on the display. A clearer understanding of the color-coded highlighting mechanism may alleviate the cognitive demands associated with interpreting multiple information streams to guide their decision-making process.

Second, the present study cannot disentangle whether the observed effects were due to the RAG mechanism, the color-coding, or their interaction. It is possible that participants largely did not pay attention to the RAG-based references in the No-Highlight condition while in the Highlight condition while the presence of salient color-coding may have prompted participants to actively scrutinize the source information. This may have increased cognitive workload and decreased sensitivity. Future research should include additional control conditions in which participants receive LLM outputs without the RAG-based

reference with or without the color-coding. These two control conditions would help clarify the specific processes that participants employ when evaluating atomic facts from LLM-generated outputs.

The current results should be taken with caution, because, due to the logistical and recruitment issues, the sample size in the current report was smaller ($N = 18$) than originally anticipated ($N = 48$), preventing from making strong statistical inferences from the data. In the new partnership with Old Dominion University, data collection will continue to collect data to allow more robust statistical comparisons between the conditions in this emerging area of research.

In conclusion, the current human-subject experiment offers new insights into how human users incorporate LLM-generated outputs of varying accuracy when rendering decisional choices in the nuclear power plant-relevant task. Overall, participants adopted more liberal response criteria, increasing their probability to say that each statement is accurate, and showed a decrease in performance and trust, and an increase in workload. Recent studies have begun to explore the potential benefits and costs of LLMs on human performance in especially cognitively demanding tasks, in the goal of alleviating cognitive load of human users in high-stakes tasks, with mixed findings (Kosmyna et al., 2025; Spatharioti et al., 2025). Identifying specific LLM characteristics that influence human performance in such tasks is an urgent research priority. A series of carefully controlled human-subject experiments are essential for modeling and potentially controlling human-AI interactions under cognitive load. Such research can inform the effective deployment of AI/LLM technologies in nuclear power plant-specific contexts and contribute to the broader goal of optimizing human-systems integration in AI-augmented decision environments.

Appendix C-1

Caccioppo et al. (1982) Need for Cognition Scale

Need for Cognition Scale

For each statement listed below, circle the number that indicates the extent to which you feel it is characteristic of you. For example, if the statement is not at all like you, circle number 1 under “Extremely Uncharacteristic,” or if you really can’t decide if the statement is or is not characteristic of you, circle number 3 under “Uncertain.”

1. I would prefer complex to simple problems.

1	2	3	4	5	6	7

2. I like to have the responsibility of handling a situation that requires a lot of thinking.

1	2	3	4	5	6	7

3. Thinking is not my idea of fun.*

1	2	3	4	5	6	7

4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.*

1	2	3	4	5	6	7

5. I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.*

1	2	3	4	5	6	7

6. I find satisfaction in deliberating hard and for long hours.

1	2	3	4	5	6	7

7. I only think as hard as I have to.*

1	2	3	4	5	6	7
---	---	---	---	---	---	---

8. I prefer to think about small, daily projects to long-term ones.*

1	2	3	4	5	6	7
---	---	---	---	---	---	---

9. I like tasks that require little thought once I've learned them.*

1	2	3	4	5	6	7
---	---	---	---	---	---	---

10. The idea of relying on thought to make my way to the top appeals to me.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

11. I really enjoy a task that involves coming up with new solutions to problems.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

12. Learning new ways to think doesn't excite me very much.*

1	2	3	4	5	6	7
---	---	---	---	---	---	---

13. I prefer my life to be filled with puzzles that I must solve.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

14. The notion of thinking abstractly is appealing to me.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

15. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

1	2	3	4	5	6	7

16. I feel relief rather than satisfaction after completing a task that required a lot of mental effort.*

1	2	3	4	5	6	7

17. It's enough for me that something gets the job done; I don't care how or why it works.*

1	2	3	4	5	6	7

18. I usually end up deliberating about issues even when they do not affect me personally.

1	2	3	4	5	6	7

Appendix C-2

Modified Tully et al. (2025) AI Literacy Survey

1. Which of the languages below are used to code a robot?
 - a. Python
 - b. HTML
 - c. Computer Vision
 - d. Hypertext
2. Which technology is the primary enabler of Artificial Intelligence?
 - a. Machine Learning
 - b. Electric Battery
 - c. Robotics
 - d. Engineering
3. Which of these is a notable algorithmic problem?
 - a. Counting
 - b. Sorting
 - c. Multiplication
 - d. Division
4. Which of these is not a best practice when trying to protect your privacy online?
 - a. Using a two-factor authentication to access account and devices
 - b. Storing your passwords in the notes section of your smartphone
 - c. Updating your operating system promptly when an update becomes available
 - d. Using different login information
5. If using an algorithm, can the results be biased against specific groups of people?
 - a. No, unlike humans, algorithms are free of emotions and cannot be wrong
 - b. Yes, but only if you train them on privileged populations
 - c. Yes, but only if you use them with incomplete data
 - d. Yes, even with correct data and training, algorithms can provide biased results
6. How algorithmic decisions best explained?
 - a. Only for very basic and simple algorithms
 - b. They can be explained by sharing the algorithm with everyone
 - c. As complex code, they cannot be explained
 - d. Algorithms can be explained by sharing the contributing factors to decisions
7. AI can take images and short sound clips of people and create long, convincing impersonation of these people saying anything. What is this technology called?
 - a. Deepmind
 - b. Deepfake
 - c. Deeplearn
 - d. Deepvoice
8. Google automatically completes the text you type in the search bar...

- a. By looking at the most common search terms used by others and your search history
 - b. Based only on your search history
 - c. Based only on what people in your zip code are searching for
 - d. By using your phone camera when you type and reading your lips
9. Algorithms may improve website performance by doing all of the following except for....
- a. Showing us content that might interest us
 - b. Showing us relevant ads
 - c. Monitoring which content we engage in
 - d. Changing our screen brightness
10. What information do social media algorithms not typically use when deciding what information to display to a person?
- a. Articles people have previously clicked on
 - b. Videos people have previously watched online
 - c. Information people have shared on social media
 - d. What would benefit society
11. When you use a web-based service which stores its data in "the cloud", where is the data actually stored?
- a. In a hidden folder in your computer
 - b. At the company's headquarters
 - c. At a large server farm
 - d. In your browser's cookies
12. If you turn off personalized ads, you are most likely to see ads...
- a. Targeting people similar to you
 - b. That are relevant for you
 - c. That are interesting to you
 - d. For commonly purchased products
13. When do algorithms tend to perform better than humans?
- a. When lots of structured data is involved and the risk from a wrong answer is minimal
 - b. When there is no structured data available and the risk from a wrong answer is minimal
 - c. When lots of structured data is involved and the risk from a wrong answer is significant
 - d. When there is no structured data available and risk from a wrong answer is significant
14. What is training data and why is it important?
- a. Training data is used when the model is fully calibrated, to train new users with how to work with the model
 - b. Training data is used to calibrate the model during creation, creating a model that performs best according to that data

- c. Training data is used to test the model after it was created, to see how it works with new data
 - d. Training data is used to minimize bias in older models and train them to new population types
15. Indicate if the following is primarily the result of an algorithm or not: What I write in a post on social media
- a. Yes
 - b. No
16. Indicate if the following is primarily the result of an algorithm or not: What I type in a text message to a friend
- a. Yes
 - b. No
17. Indicate if the following is primarily the result of an algorithm or not: When I read a direct message on a social media platform
- a. Yes
 - b. No
18. Indicate if the following is primarily the result of an algorithm or not: Returning the results of a search query on a search engine (e.g., Google)
- a. Yes
 - b. No
19. Indicate if the following is primarily the result of an algorithm or not: Dual authentication to access one of my accounts
- a. Yes
 - b. No
20. What is the process of a machine learning model making a prediction based on new data called?
- a. Coding
 - b. Training
 - c. Testing
 - d. Inference

Appendix C-3

Jian et al. (2000) Trust Questionnaire

Checklist for Trust between People and Automation

Below is a list of statement for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please mark an “x” on each line at the point which best describes your feeling or your impression.

(Note: not at all = 1: extremely = 7)

1. The artificial intelligence is deceptive

1	2	3	4	5	6	7
---	---	---	---	---	---	---

2. The artificial intelligence behaves in an underhanded manner

1	2	3	4	5	6	7
---	---	---	---	---	---	---

3. I am suspicious of the artificial intelligence’s intent, action, or outputs

1	2	3	4	5	6	7
---	---	---	---	---	---	---

4. I am wary of the artificial intelligence

1	2	3	4	5	6	7
---	---	---	---	---	---	---

5. The artificial intelligence’s actions will have a harmful or injurious outcome

1	2	3	4	5	6	7
---	---	---	---	---	---	---

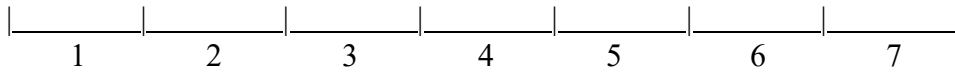
6. I am confident in the artificial intelligence

1	2	3	4	5	6	7
---	---	---	---	---	---	---

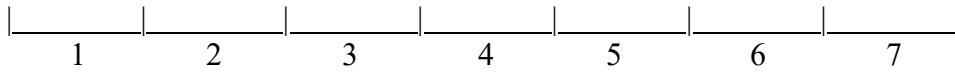
7. The artificial intelligence provides security

--	--	--	--	--	--	--

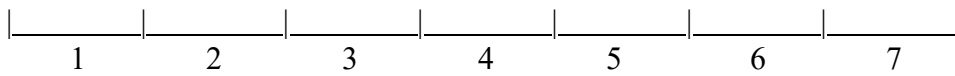
1 2 3 4 5 6 7
8. The artificial intelligence has integrity



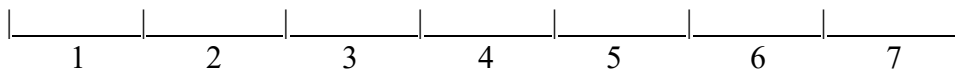
9. The artificial intelligence is dependable



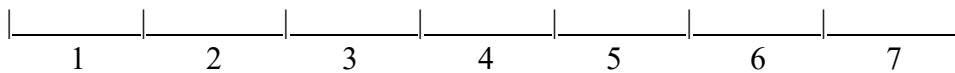
10. The artificial intelligence is reliable



11. I can trust the artificial intelligence



12. I am familiar with the artificial intelligence



Appendix C-4

Chancey et al. (2017) Trust Questionnaire

TRUST QUESTIONNAIRE

Below is a list of statements for evaluating trust between people and artificial intelligence. Please circle the number that best describes your feeling or your impression of the artificial intelligence you used during the task.

1. Even when the artificial intelligence gives me unusual advice, I am certain that the aid's advice will help me to perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

2. For me to perform well, I can rely on the artificial intelligence to function properly

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

3. It is easy to follow what the artificial intelligence does to help me perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

4. The artificial intelligence's advice reliably helps me perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

5. The artificial intelligence's advice consistently helps me perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

6. I understand how the artificial intelligence will help me perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

7. Even if I have no reason to expect that the artificial intelligence will function properly, I still feel certain that it will help me to perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

8. Although I may not know exactly how the artificial intelligence works, I know how to use it to perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

9. To help me perform well, I believe advice from the artificial intelligence even when I don't know for certain that it is correct.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

10. To help me perform well, I recognize what I should do to get the advice I need from the artificial intelligence the next time I use it.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

11. I will be able to perform well the next time I use the artificial intelligence because I understand how it behaves.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

12. The artificial intelligence always provides the advice I require to help me perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

13. The artificial intelligence adequately analyzes the system consistently, to help me perform well.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

Appendix C-6

Chancey et al. (2017) Risk Questionnaire

PERCEIVED RISK QUESTIONNAIRE

The following questions are about how you perceive the level of risk associated with maintaining a high level of performance during the experiment.

I believe that...

1. The consequences for performing poorly on these tasks are substantial.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

2. The overall risk of performing poorly on these tasks is high.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

3. Overall I would label the consequences of performing poorly on these tasks as something negative.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

4. I would label the consequences of performing poorly on these tasks as a significant loss.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

5. Performing poorly on the experimental tasks could have negative ramifications.

Not Descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 :Very Descriptive

Appendix C-7

Modified Technology Acceptance Questionnaire

Modified Technology Acceptance Questionnaire

The following questions are about your behavioral intention, perceived usefulness, and perceived ease of the artificial intelligence. Please circle the number that best describes your feeling or your impression of the artificial intelligence. All items will be measured on a 7-point Likert scale, where 1 = strongly disagree, 2 = moderately disagree, 3 = somewhat disagree, 4 = neutral (neither disagree nor agree), 5 = somewhat agree, 6 = moderately agree, and 7 = strongly agree.

Behavioral Intention

1. If the artificial intelligence is available in the market at an affordable price, I intend to purchase the artificial intelligence.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

2. If my artificial intelligence is equipped with similar capabilities, I predict that I would use my artificial intelligence when working.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

3. Assuming that the artificial intelligence is available, I intend to use the artificial intelligence regularly.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

Perceived Usefulness

1. Using the artificial intelligence would improve my performance.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

2. Using the artificial intelligence increases my safety.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

3. Using the artificial intelligence enhances effectiveness in my work.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

4. I would find the artificial intelligence useful.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

Perceived Ease of Use

1. My interaction with the artificial intelligence would be clear and understandable.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

2. I would find the artificial intelligence difficult to use (reverse-scaled item).

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

3. Interacting with the artificial intelligence would not require a lot of mental effort.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree

4. I would find it easy to get the artificial intelligence to do what I want it to do.

Strongly Disagree: 1 2 3 4 5 6 7: Strongly Agree