



Role of Uncertainty Quantification in the Explainability of Large Language Models for the Nuclear Industry

June 2025

Changing the World's Energy Future

Nancy J Lybeck, Vaibhav Yadav, Ronald Laurids Boring PhD, Vivek Agarwal



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.

Role of Uncertainty Quantification in the Explainability of Large Language Models for the Nuclear Industry

Nancy J Lybeck, Vaibhav Yadav, Ronald Laurids Boring PhD, Vivek Agarwal

June 2025

**Idaho National Laboratory
Idaho Falls, Idaho 83415**

<http://www.inl.gov>

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

Role of Uncertainty Quantification in the Explainability of Large Language Models for the Nuclear Industry

Nancy J. Lybeck^{1*}, Vaibhav Yadav¹, Ronald L. Boring¹, Vivek Agarwal¹

¹Idaho National Laboratory, Idaho Falls, ID

[leave space for DOI, which will be inserted by ANS]

ABSTRACT

The meteoric rise of generative artificial intelligence (AI) large language models (LLMs) has created an opportunity to utilize them to increase efficiencies in a multitude of industries. While LLMs carry great potential to revolutionize the manner in which work is performed, numerous known deficiencies limit their utility, including the black box nature of the models, the stochastic nature of the response (i.e., presenting the same prompt multiple times results in different responses), and the potential for hallucination. Widespread adoption of LLMs in safety-critical industries such as nuclear will require some form of explainability to assure end users that the LLM's response to a given query is valid. Model uncertainty is inherently linked to the concepts of trust and explainability, and can be used to identify situations in which the model is insufficiently certain about its answer. Although uncertainty is not enough in and of itself to determine the suitability of an answer—a model can be very certain of an inaccurate answer—it still provides valuable supporting information. Practical methodologies for gauging or quantifying the uncertainty in LLM outputs are presented herein, along with examples based on nuclear-specific prompts.

Keywords: Large Language Models, Confidence, Uncertainty, Explainability, Trust

1. INTRODUCTION

The rapid advancement and ubiquity of generative artificial intelligence (AI) large language models (LLMs) has opened up numerous opportunities to enhance efficiency across various industries. Used in applications ranging from personalized recommendations on e-commerce sites to chatbots and virtual assistants [1], LLMs are now engaged by most people on a frequent basis. There are too many freely available LLMs today to list, but among the most well-known are Generative Pre-trained Transformer (GPT) 4, developed by OpenAI; Gemini, developed by Google AI; Large Language Model Meta AI (LlaMA), developed by Meta AI; and Falcon, developed by the Technology Innovation Institute. It can be quite informative to spend some time chatting with an LLM, especially when posing a query multiple times to see the different answers the model generates. Prompting an LLM also supports day-to-day work activities such as technical queries and problem solving. However, the variability of outputs from an LLM becomes a serious matter in cases where we need to assume the veracity of the generated responses.

Despite their potential to transform work processes, LLMs face several limitations, such as their opaque nature, inconsistent responses, and the risk of generating incorrect information. Indeed, the stochastic nature of their responses can be both fascinating and frustrating, depending on the purpose of the query. For LLMs to gain widespread acceptance in safety-critical sectors such as nuclear energy, they must offer

*Nancy.Lybeck@inl.gov

a degree of explainability to ensure that users can trust the validity of their responses. The concepts of trust and explainability are closely tied to model uncertainty, which can help identify instances in which the model’s confidence in its answer is insufficient. While model uncertainty alone cannot guarantee the accuracy of a response—since a model can be confidently wrong—it still provides essential contextual information. This work explores simple methodologies for assessing and quantifying the uncertainty in LLM outputs, with examples tailored to nuclear-specific queries.

2. Background

One useful way to think about uncertainty in machine learning models is lack of confidence in the output [2]. However, today’s increasingly sophisticated LLMs have been developed to produce responses that are highly polished and may convey more confidence in the answer than is warranted, and typically do not return a confidence score. Indeed, hallucinations—defined by IBM as “a phenomenon wherein an LLM perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate” [3]—are one of the biggest areas of concern in the use of LLMs.

Uncertainty quantification (UQ) has been widely studied for classical machine learning models. The traditional approach to quantifying total uncertainty in models includes both aleatoric (data) uncertainty and epistemic (model) uncertainty. In the context of LLMs, data uncertainty can be attributed to the prompts (e.g., an ambiguous prompt will result in high uncertainty) [4] and uncertainty in the training data, whereas model uncertainty arises when the model lacks the information necessary to answer the question correctly (i.e., the training data did not contain relevant information). UQ for LLMs presents many challenges, including huge dimensionality (GPT 4.0 has approximately 1.8 trillion parameters [5]), black-box model limitations, and the nature of the output (generated text vs. numerical or categorical results) [4].

Many recent papers focus on aspects of UQ for LLMs, as evidenced by the curated list of Uncertainty, Reliability and Robustness in Large Language Models, which is maintained on GitHub by Jiaxin Zhang [6]. While a comprehensive review of UQ approaches for LLMs is outside the scope of this paper, the following works give an idea as to the breadth and complexity of the issue. Xiao et al. [7] explored the use of large-scale empirical analysis for UQ with pre-trained language models. Hou et al. [4] discussed gauging disagreements in model predictions under different prompt clarifications. Chen and Mueller [8] developed the BSDetector, a method that detects low-quality answers from LLMs and quantifies a numeric confidence score. Tanneru [9] dug deeper, looking at UQ specifically for natural language explanations of responses given by LLMs. Many of these approaches are complex, requiring knowledgeable users and significant computing resources.

The average LLM user is not an ML expert, does not necessarily care about the source of the uncertainty, and simply needs to assess the quality of an answer in a manner that does not require significant effort. Fortunately, certain common-sense techniques exist that any user can easily employ to assess the level of uncertainty in the model’s response. The same prompt—or paraphrased versions thereof—can be submitted multiple times, and the differences in response can be characterized, either based on user judgment or by using quantitative metrics to assess the similarity of the responses.

Cosine similarity is one metric commonly used to quantify the similarity between two text samples. At the most basic level, this metric calculates the cosine of the angle between two multi-dimensional spaces. This is easy to imagine in two-dimensional space, as two orthogonal vectors have zero similarity (the cosine of 90° is 0), and two vectors that overlap have perfect similarity (the cosine of 0° is 1). To calculate cosine similarity between two text samples [10], the samples must be converted into numerical vectors by removing insignificant terms such as “a,” “an,” or “the” (called stop words), then counting the frequency of specific words/subwords (tokens). A word of caution: cosine similarity does not account for context and semantics, both of which are important for determining the meaning behind words. For example, the sentences “The story is about a dog” and “The story is not about a dog” share a very high degree of similarity despite having opposite meanings. It is worth noting that there are a multitude of more sophisticated ways to measure the similarity between two text samples [11], including some that take into account semantics and context. Cosine similarity was selected for this paper based on its simplicity and the ready availability of online calculators.

3. EXAMPLES

The following examples were developed using Idaho National Laboratory's AI Virtual Assistant (AiVA), which was built on the Azure Government OpenAI platform and trained on the same information fed to the GPT-4o model. The information presented was not pulled in real-time; rather, it is based on the knowledge trained into AiVA. The knowledge cutoff date for AiVA was October 2023. Any updates, developments, or new data released after that date would not be included in the current knowledge base.

For the purposes of illustration, three questions were used as base prompts for the examples. These were then paraphrased to create a total of three different prompts that were submitted so as to judge the consistency of the responses. While rewording the prompts was not strictly necessary, it was implemented in an attempt to account for as many sources of variability as possible. The accurate conversation style (options are accurate, balanced, or creative) was used for all prompts. This setting affects the randomness of the responses, resulting in more deterministic and predictable outputs closely aligned with the training data. All related prompts were posed within the same conversation; this choice was made to better emulate average user interaction, however it is important to acknowledge the responses from latter prompts may be impacted by context retention.

3.1. Example 1

The first example question, which pertained to the purpose of trash racks in nuclear power plants, was chosen to show the consistency of responses to a question that should be well within the knowledge base of the LLM. The three prompts (P1.1–P1.3) and the responses they generated are shown below. In this case, it is fairly easy for the user to assess that these responses are all very similar to each other, and that while the responses are all different, there are no real inconsistencies.

P1.1 What are trash racks used for in nuclear power plants?

Trash racks are used in nuclear power plants to prevent debris and large objects from entering the water intake systems. They are essentially screens or grates placed at the intake points of cooling water systems to filter out items like leaves, branches, and other debris that could potentially damage pumps, clog pipes, or interfere with the cooling process. This helps ensure the safe and efficient operation of the plant by maintaining unobstructed water flow to the cooling systems.

P1.2 In the nuclear power industry, what purpose do trash racks serve?

In the nuclear power industry, trash racks serve the purpose of filtering out debris and large objects from the water entering the intake systems. They are essential in preventing potential damage to pumps, avoiding pipe clogging, and ensuring the continuous and efficient operation of the cooling systems. By keeping the intake water free from obstructions, trash racks help maintain the overall safety and reliability of the nuclear power plant.

P1.3 What is the purpose of a trash rack in a nuclear power plant?

The purpose of a trash rack in a nuclear power plant is to filter out debris and large objects from the water entering the plant's intake systems. This includes items such as leaves, branches, and other debris that could potentially damage pumps, clog pipes, and interfere with the cooling processes. By ensuring only clean water flows into the cooling and auxiliary systems, trash racks help maintain the plant's safety and operational efficiency.

For prompt P1.3, AiVA was also asked: "How confident is that answer?" The response was: "I am highly confident in this answer, as trash racks are a standard component in water intake systems for various types of power plants, including nuclear, to prevent debris-related issues."

While this example is simple enough that the user can readily determine that the three responses all convey the same general information, for the purposes of illustration the cosine similarities of the responses were calculated using an online tool hosted by Tilores [12], and these similarities are given in Table 1.

Table 1: Cosine similarities (%) calculated for responses to prompts P1.1–P1.3

Prompt	P1.1	P1.2	P1.3
P1.1	—	84.9	87.6
P1.2	84.9	—	85.6
P1.3	87.6	85.6	—

3.2. Example 2

The second question regarded the Young’s modulus for Inconel 690, an alloy commonly used in nuclear fuel elements and fuel cladding. The Young’s modulus is a stiffness measure that indicates the extent to which the material deforms under high loads. It is important to note that the different prompts (P2.1–P2.3) used for this example do not specify temperature, and that this has a significant impact on the Young’s modulus. Reported Young’s modulus values for Inconel 690 range from 211 GPa at room temperature (68° F) to 125.5 GPa at 2000° F [13]. The results from AiVA are shown below.

P2.1 What is the Young’s modulus for Inconel 690?

The Young’s modulus for Inconel 690 is approximately 211 GPa (Gigapascals).

P2.2 Report the Young’s modulus used in modeling for Inconel 690?

The Young’s modulus commonly used in modeling for Inconel 690 is approximately 207 GPa (Gigapascals). However, it’s always best to refer to specific material property data sheets or modeling guidelines relevant to your project for the most accurate information.

P2.3 For Inconel 690 what is the Young’s Modulus?

The Young’s modulus for Inconel 690 is approximately 208 GPa (Gigapascals).

Again, it is fairly simple for the user to assess how similar these results are. The answers given vary between 207 and 211 GPa, and though this might be frustrating, all these values are in the reported range for the modulus at room temperature—a temperature that is not rigidly defined but generally accepted to be between 68-72° F. It is interesting to note that the generated responses in no way indicate that the prompt might be missing critical information such as temperature. The response to P2.2 comes closest, as it suggests the user look up specific materials property data sheets. The calculated cosine similarities are shown in Table 2. The results for P2.1 and P2.3 show 95% agreement, reflecting the fact that literally the only difference between those two responses is the value returned. P2.2 was more verbose, resulting in much lower similarity scores with P2.1 and P2.3. In general, similarity values will be less useful for questions that seek a quantitative answer.

In the absence of expertise in the subject matter being explored by the prompt, users can always ask the LLM what additional information should be considered. For this example, AiVA was asked, “What is the most important factor in determining the Young’s modulus for Inconel 690?” The response was: “The most important factor in determining the Young’s modulus for Inconel 690 is the temperature at which the measurement is taken. Inconel 690, like many materials, exhibits temperature-dependent mechanical properties. Ensuring accurate temperature control and documentation during the testing process is crucial for obtaining a reliable measurement of Young’s modulus.”

Table 2: Cosine similarities (%) calculated for the responses to prompts P2.1–P2.3

Prompt	P2.1	P2.2	P2.3
P2.1	—	63.0	95.0
P2.2	63.0	—	64.5
P2.3	95.0	64.5	—

3.3. Example 3

The final example is focused on the impact of stress on human performance in the control room. More specifically, we are interested in the impact on completion time for control room actions. The question is purposely vague, as there are many types of control room actions an operator might need to complete. Furthermore, this is a topic not fully covered in previous research [14], and so would be prone to generating spurious answers. The results obtained from the queries are given below.

P3.1 What is the effect of stress on time to complete control room actions in a nuclear power plant?

Stress can have a significant impact on the time required to complete control room actions in a nuclear power plant. Elevated stress levels can impair cognitive functions such as attention, memory, and decision-making, leading to slower reaction times and increased likelihood of errors. Operators under stress may take longer to process information, recognize alarms, and execute necessary procedures. This delay can affect the overall efficiency and safety of the plant operations. It is crucial to implement training and stress management techniques to help operators maintain performance under high-stress conditions.

P3.2 How do elevated stress levels change the time it takes a nuclear power plant operator to complete a control room action?

Elevated stress levels can increase the time it takes a nuclear power plant operator to complete a control room action due to several factors:

- 1. Decreased Cognitive Function: Stress can impair cognitive abilities such as attention, memory, and problem-solving skills, leading to slower information processing and decision-making.*
- 2. Reduced Situational Awareness: High stress can narrow an operator's focus, causing them to miss important cues or changes in the environment that are crucial for timely actions.*
- 3. Physical Symptoms: Stress can cause physical symptoms such as trembling hands, sweating, or increased heart rate, which can affect fine motor skills and coordination, leading to slower and less precise actions.*
- 4. Emotional Impact: Stress can lead to anxiety and frustration, which can further impair an operator's ability to perform tasks efficiently.*

Overall, elevated stress levels typically result in delayed response times and increased likelihood of errors in completing control room actions, which can have serious implications for the safety and efficiency of nuclear power plant operations.

P3.3 What is the impact of stress on how long it takes a nuclear plant operator to complete a control room action?

The impact of stress on the time it takes a nuclear plant operator to complete a control room action generally results in increased response times. Stress can lead to:

1. *Slower Decision-Making: Operators under stress may take longer to assess situations and make decisions due to impaired cognitive functions.*
2. *Delayed Reactions: High stress can cause operators to react more slowly to alarms and other critical indicators.*
3. *Increased Errors: Stress can increase the likelihood of mistakes, which may require additional time to correct.*
4. *Reduced Efficiency: Stress can affect the operator's overall efficiency, leading to longer times to complete procedures.*

In summary, stress typically prolongs the time required for operators to complete control room actions, and this can adversely affect the safety and smooth operation of nuclear power plants.

The cosine similarities for responses P3.1–P3.3, as given in Table 3, indicate lower overall agreement among the responses. In fact, compared to the previous examples, it is less easy for the user to subjectively assess the similarity, as each response covers slightly different facets of stress on performance. The first response, P3.1, focuses on concrete performance outcomes such as the effects of stress on cognitive functions, including perceptual recognition tasks. P3.2, the second response, adds situation awareness, physical symptoms, and emotional impact. It also includes a comment about increased error rates, without explicitly linking error rates to task completion time. The third response, P3.3, talks about the time required for making decisions and responding appropriately, error rates, and efficiency—including aspects from P3.1 and P3.2, but pieced together in a different manner. Here, the effects of errors on increasing recovery time are explained.

Table 3: Cosine similarities (%) calculated for responses to prompts P3.1-P3.3

Prompt	P3.1	P3.2	P3.3
P3.1	—	76.6	78.7
P3.2	76.6	—	78.6
P3.3	78.7	78.6	—

The cumulative effect of multiple prompts within the same conversation is to add to the knowledge provided to the user in the previous responses—with each response building on the others, thus providing a fuller picture of the relationship between stress and the time needed to perform tasks in the control room. While each response is correct, the responses do not converge on a single set of answers. This variety of responses may be a reflection of the ambiguity of the topic and the lack of clear research literature behind the model training.

Interestingly, the responses fail to capture the nuances of stress and the task types (e.g., certain types of acute stress can actually speed up the time to complete tasks, though with a negative trade-off in terms of increased human error). Just as temperature matters when considering Young's modulus, the type of stress and type of task represent necessary context for determining the impact on speed to complete tasks. The responses are correct in common situations but provide insufficient context to differentiate between opposite effects of stress—sometimes stress slows down the operator, sometimes it actually speeds the operator up. Speeding up performance could lead to significant effects on operational safety, yet this is in no way conveyed to the user. Due to this omission, the responses must be considered incomplete, providing users no indication that the context matters. The cosine similarities hint at some degree of dissimilarity in the responses but do not give a measure of the correctness/completeness of those responses.

4. Hallucinations

The propensity of LLMs to return false information has been widely publicized, undermining trust in them and hindering their ultimate adoption. While every model is different, they all hallucinate. According to the Hallucination Leaderboard [15] maintained by Vectara on GitHub, hallucination rates for current LLMs range from 1.3% to nearly 30%. AiVA is based on GPT-4o, which has a relatively low hallucination rate of 1.5%. Even with that low rate, it is not difficult to present prompts that result in hallucinations. For example, the prompt “What are the top 5 publications on micro reactors?” was presented to AiVA. The response returned is given below.

As of my last update in October 2023, I cannot provide real-time data or the very latest publications. However, here are five highly regarded publications and resources on microreactors up to that time:

1. *“Microreactor Technology and Process Intensification” by Thomas Wirth - This book provides comprehensive coverage of microreactor technology and its applications in chemical processes.*
2. *“Micro Process Engineering: Fundamentals, Devices, Fabrication, and Applications” by Norbert Kockmann - This publication offers a detailed introduction to the fundamentals and applications of micro process engineering, including microreactors.*
3. *“Microreactors in Organic Chemistry and Catalysis” edited by Thomas Wirth - This edited volume includes various chapters on the use of microreactors in organic synthesis and catalytic processes.*
4. *“Microreactors: Design, Fabrication, and Applications” edited by Wolfgang Ehrfeld, Volker Hessel, and Holger Löwe - This book is a foundational text that covers the design, fabrication, and various applications of microreactors.*
5. *Research Articles in Journals: Look for specific articles in journals such as Chemical Engineering Journal, Lab on a Chip, and AIChE Journal. These journals often publish cutting-edge research on microreactor technology. For the most recent publications, I recommend searching databases like PubMed, IEEE Xplore, Scopus, or Google Scholar with keywords related to microreactors.*

Taken at face value, this appears to be a credible, albeit somewhat dated, list of references. However, digging into the details reveals two discrepancies. The first book listed was edited by Yong Wang and Jamelyn D. Holladay; Thomas Wirth does not appear to be an editor or contributing author. The fourth book, however, does not even appear to exist. This simple example points to the need for caution when relying on LLM output.

In safety-critical applications such as the nuclear industry, caution must be used when deploying LLMs for decision making purposes. More advanced models incorporate techniques such as Retrieval-Augmented Generation that reduce the incidence of hallucinations. In some cases, models can be fine-tuned by providing application specific information prior to deployment to improve overall results. Additionally, model settings such as temperature (a way to tune the level of creativity in the LLM response) are sometimes available for the user to adjust.

5. CONCLUSIONS

UQ for LLMs is a complex topic; however, users can use certain simple techniques to assess the quality of LLM output. Submitting multiple prompts within the same conversation—asking the same question in different words—generates responses that can be compared to each other. In some cases, the user can readily determine whether the responses contain the same general information or are substantially different. If the user is not comfortable assessing the differences, methodologies such as cosine similarity can be used to quantify the responses’ level of similarity to each other. Methods for quantifying similarity must be used carefully, as they may not account for context and syntax in the text.

UQ of LLM responses could provide insights for specific ways to reduce uncertainties. Researchers and users of LLM could explore various approaches such as providing additional data with prompts, prompt

engineering, or in some cases considering fine-tuning the LLM on a set of data specific to a domain application. UQ opens the opportunity for users to take actions towards more robust and consistent model performance. Additionally, UQ is an integral partner to explainability, trust, and reliability. In order for the user to trust model output, they need to know the reason for the response and how uncertain the answer is.

End users need to be aware that LLMs will present information in a polished and confident manner, even when there is ambiguity—perhaps unexpected—in the prompt. The user can attempt to remedy this by asking the LLM whether any additional information is needed to accurately answer the question of interest. The user can also ask the LLM how confident it is in its response to a specific prompt. While high confidence in the output does not guarantee its correctness, it is an important consideration when relying on LLM output.

LLMs are powerful tools that are fun to interact with but should be used in a thoughtful manner. At the most basic level, LLMs can help people improve their efficiency and productivity; however, users should always maintain a questioning attitude and take care to ensure the accuracy of the responses.

ACKNOWLEDGEMENTS

This work was made possible through funding from the U.S. Department of Energy’s Light Water Reactor Sustainability Program at Idaho National Laboratory (under contract no. DE-AC07-05ID14517), an agency of the U.S. Government. Neither the U.S. Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

References

- [1] S. Uspenskyi. “Large Language Model Statistics And Numbers (2025).” URL <https://springsapps.com/knowledge/large-language-model-statistics-and-numbers-2024>.
- [2] iMerit. “A Comprehensive Introduction to Uncertainty in Machine Learning.” Blog (2022). URL <https://imerit.net/blog/a-comprehensive-introduction-to-uncertainty-in-machine-learning-all-una/>.
- [3] IBM. “Hallucination.” URL <https://www.ibm.com/think/topics/ai-hallucinations>.
- [4] B. Hou, Y. Liu, K. Qian, J. Andreas, S. Chang, and Y. Zhang. “Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling.” (2024). URL <https://arxiv.org/abs/2311.08718>.
- [5] J. Howarth. “Number of Parameters in GPT-4 (Latest Data).” (2024). URL <https://explodingtopics.com/blog/gpt-parameters>.
- [6] J. Zhang. “Awesome-LLM-Uncertainty-Reliability-Robustness.” URL <https://github.com/jxzhangjhu/Awesome-LLM-Uncertainty-Reliability-Robustness?tab=readme-ov-file>.
- [7] Y. Xiao, P. P. Liang, U. Bhatt, W. Neiswanger, R. Salakhutdinov, and L.-P. Morency. “Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis.” (2022). URL <https://arxiv.org/abs/2210.04714>.
- [8] J. Chen and J. Mueller. “Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness.” (2023). URL <https://arxiv.org/abs/2308.16175>.
- [9] S. H. Tanneru, C. Agarwal, and H. Lakkaraju. “Quantifying Uncertainty in Natural Language Explanations of Large Language Models.” (2023). URL <https://arxiv.org/abs/2311.03533>.
- [10] P. Miesle. “What is Cosine Similarity: A Comprehensive Guide.” URL <https://www.datastax.com/guides/what-is-cosine-similarity>.
- [11] S. Kumar. “Understanding Text Similarity.” URL <https://www.linkedin.com/pulse/understanding-text-similarity-sanjay-kumar-mba-ms-phd-ughfc/>.

- [12] Tilores. “Cosine Similarity Calculator.” URL <https://tilores.io/cosine-similarity-online-tool>.
- [13] Special Metals. “Technical Bulletin on Inconel Alloy 690, Publication Number SMC-079.” URL <https://www.specialmetals.com/documents/technical-bulletins/inconel/inconel-alloy-690.pdf>.
- [14] R. L. Boring. “The Spatial Dimension in Human Reliability Analysis.” *Proceedings of the 33rd European Safety and Reliability Conference*, pp. 902–909 (2023).
- [15] Vectara. “Hallucination Leaderboard.” URL <https://github.com/vectara/hallucination-leaderboard>.



Nancy J. Lybeck

Vaibhav Yadav

Ronald L. Boring

Vivek Agarwal

June 18, 2025



Role of Uncertainty Quantification in the Explainability of Large Language Models for the Nuclear Industry



Large Language Models (LLMs) are ubiquitous



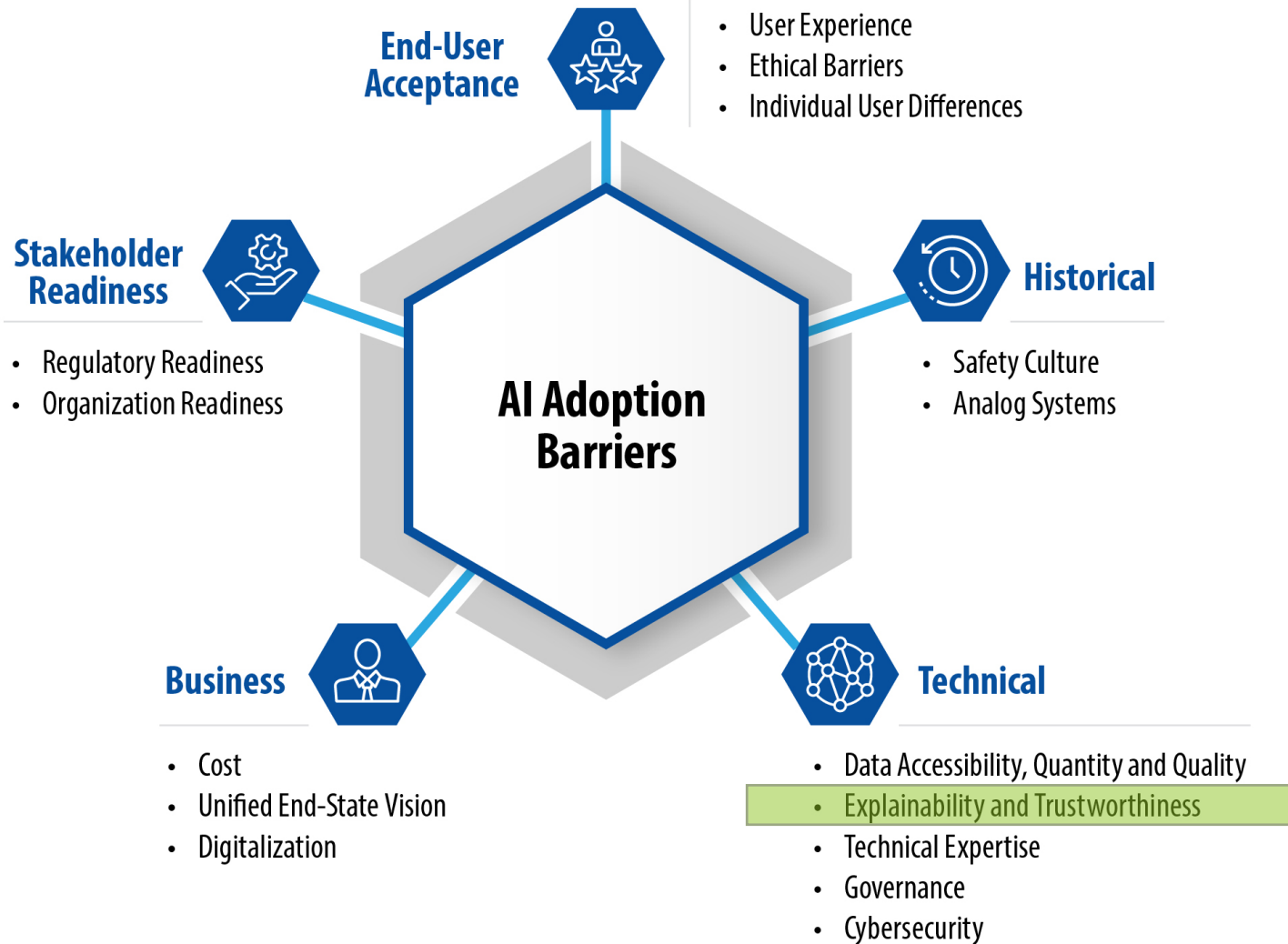
Generative AI has a lot to offer ...

“Generative AI has the potential to change the world in ways that we can’t even imagine. It has the power to create new ideas, products, and services that will make our lives easier, more productive, and more creative. It also has the potential to solve some of the world’s biggest problems, such as climate change, poverty, and disease. The future of generative AI is bright, and I’m excited to see what it will bring.” ~Bill Gates

with some caveats

“The top performers understand the limitations of the tool. Just like you don’t use a hammer to turn a screw, you don’t ask gen AI questions that are best answered in other ways.” ~ Matt Banhozler, McKinsey & Company

Adoption of AI in Nuclear



23-50623

Anna Hall, Vivek Agarwal. "Barriers to Adopting Artificial Intelligence and Machine Learning Technologies in Nuclear Power." *Progress in Nuclear Energy*, Volume 175, 2024, <https://doi.org/10.1016/j.pnucene.2024.105295>.



Trust, Explainability, and Uncertainty

- Uncertainty is inherently linked to trust and explainability
 - Uncertainty can be thought of as lack of confidence in model output
 - High uncertainty indicates model outputs are not trustworthy
 - Low uncertainty is desirable but not sufficient for trust
 - A model can be confidently wrong
- Uncertainty Quantification (UQ) for classical Machine Learning has been widely studied
- Many recent papers focus on aspects of UQ for LLMs
 - Data uncertainty
 - Prompts
 - Uncertainty in the training data
 - Model uncertainty
 - Inadequate training data
 - Many proposed methods are complex, requiring knowledgeable users and significant computing resources

Low-key Approach

- LLMs are inherently stochastic—identical prompts submitted multiple times generate different results
- Additional variability can be incorporated by paraphrasing the prompt
- Methods for comparing results:
 - User judgement
 - Subject matter expert (SME) assessment
 - Quantitative metrics like cosine similarity

Cosine Similarity

- Calculates cosine of the angle between two multi-dimensional spaces
- In 2-dimensional space
 - Orthogonal vectors have similarity 0%
 - Parallel vectors have similarity 100%
- Text samples are converted to numerical vectors
 - Removing insignificant terms
 - Counting frequency of words or subwords
- Caution: cosine similarity does not account for context and semantics
- Cosine similarity only looks at consistency of the response – not accuracy

<https://tilores.io/cosine-similarity-online-tool>

Examples

- Idaho National Laboratory's AI Virtual Assistant (AiVA)
 - Built on the Azure Government OpenAI platform
 - Trained on the same information fed to the GPT-4o model
 - Knowledge cutoff date for AiVA was October 2023
- 3 questions were used as base prompts
 - P1: What are trash racks used for in nuclear power plants?
 - P2: What is the Young's modulus for Inconel 690?
 - P3: What is the effect of stress on time to complete control room actions in a nuclear power plant?
- Each question was paraphrased twice to create 3 unique prompts
- Accurate conversation style was selected
- All prompts were posed within the same "conversation"

Prompt P1

P1.1 What are trash racks used for in nuclear power plants?

Trash racks are used in nuclear power plants to prevent debris and large objects from entering the water intake systems. They are essentially screens or grates placed at the intake points of cooling water systems to filter out items like leaves, branches, and other debris that could potentially damage pumps, clog pipes, or interfere with the cooling process. This helps ensure the safe and efficient operation of the plant by maintaining unobstructed water flow to the cooling systems.

P1.2 In the nuclear power industry, what purpose do trash racks serve?

In the nuclear power industry, trash racks serve the purpose of filtering out debris and large objects from the water entering the intake systems. They are essential in preventing potential damage to pumps, avoiding pipe clogging, and ensuring the continuous and efficient operation of the cooling systems. By keeping the intake water free from obstructions, trash racks help maintain the overall safety and reliability of the nuclear power plant.

P1.3 What is the purpose of a trash rack in a nuclear power plant?

The purpose of a trash rack in a nuclear power plant is to filter out debris and large objects from the water entering the plant's intake systems. This includes items such as leaves, branches, and other debris that could potentially damage pumps, clog pipes, and interfere with the cooling processes. By ensuring only clean water flows into the cooling and auxiliary systems, trash racks help maintain the plant's safety and operational efficiency.

Evaluation of P1 results

- All 3 responses convey the same general information

Prompt	Cosine Similarity		
	P1.1	P1.2	P1.3
P1.1	—	84.9%	87.6%
P1.2	84.9%	—	85.6%
P1.3	87.6%	85.6%	—

Prompt P2

P2.1 What is the Young's modulus for Inconel 690?

The Young's modulus for Inconel 690 is approximately 211 GPa (Gigapascals).

P2.2 Report the Young's modulus used in modeling for Inconel 690?

The Young's modulus commonly used in modeling for Inconel 690 is approximately 207 GPa (Gigapascals). However, it's always best to refer to specific material property data sheets or modeling guidelines relevant to your project for the most accurate information.

P2.3 For Inconel 690 what is the Young's Modulus?

The Young's modulus for Inconel 690 is approximately 208 GPa (Gigapascals).

Evaluation of P2 results

- Answers range from 207-211 Gpa
- All within range of reported modulus at Room Temperature
- Results in no way indicate that the prompt is missing critical context: temperature

Prompt	Cosine Similarity		
	P2.1	P2.2	P2.3
P2.1	—	63.0%	95.0%
P2.2	63.0%	—	64.5%
P2.3	95.0%	64.5%	—

Follow-up question: **What is the most important factor in determining the Young's modulus for Inconel 690?**

The most important factor in determining the Young's modulus for Inconel 690 is the temperature at which the measurement is taken. Inconel 690, like many materials, exhibits temperature-dependent mechanical properties. Ensuring accurate temperature control and documentation during the testing process is crucial for obtaining a reliable measurement of Young's modulus.

P3.1

P3.1 What is the effect of stress on time to complete control room actions in a nuclear power plant?

Stress can have a significant impact on the time required to complete control room actions in a nuclear power plant. Elevated stress levels can impair cognitive functions such as attention, memory, and decision-making, leading to slower reaction times and increased likelihood of errors. Operators under stress may take longer to process information, recognize alarms, and execute necessary procedures. This delay can affect the overall efficiency and safety of the plant operations. It is crucial to implement training and stress management techniques to help operators maintain performance under high-stress conditions.

P3.2

P3.2 How do elevated stress levels change the time it takes a nuclear power plant operator to complete a control room action?

Elevated stress levels can increase the time it takes a nuclear power plant operator to complete a control room action due to several factors:

- 1. Decreased Cognitive Function: Stress can impair cognitive abilities such as attention, memory, and problem-solving skills, leading to slower information processing and decision-making.*
- 2. Reduced Situational Awareness: High stress can narrow an operator's focus, causing them to miss important cues or changes in the environment that are crucial for timely actions.*
- 3. Physical Symptoms: Stress can cause physical symptoms such as trembling hands, sweating, or increased heart rate, which can affect fine motor skills and coordination, leading to slower and less precise actions.*
- 4. Emotional Impact: Stress can lead to anxiety and frustration, which can further impair an operator's ability to perform tasks efficiently.*

Overall, elevated stress levels typically result in delayed response times and increased likelihood of errors in completing control room actions, which can have serious implications for the safety and efficiency of nuclear power plant operations.

P3.3

P3.3 What is the impact of stress on how long it takes a nuclear plant operator to complete a control room action?

The impact of stress on the time it takes a nuclear plant operator to complete a control room action generally results in increased response times. Stress can lead to:

- 1. Slower Decision-Making: Operators under stress may take longer to assess situations and make decisions due to impaired cognitive functions.*
- 2. Delayed Reactions: High stress can cause operators to react more slowly to alarms and other critical indicators.*
- 3. Increased Errors: Stress can increase the likelihood of mistakes, which may require additional time to correct.*
- 4. Reduced Efficiency: Stress can affect the operator's overall efficiency, leading to longer times to complete procedures.*

In summary, stress typically prolongs the time required for operators to complete control room actions, and this can adversely affect the safety and smooth operation of nuclear power plants.

Evaluation of P3 results

- In this case, it is harder for the user to subjectively assess similarity
- SME evaluation noted:
 - P3.1 focuses on concrete performance outcomes
 - P3.2 adds situation awareness, physical symptoms, and emotional impact
 - P3.3 talks about time required for making decisions and responding appropriately, error rates, and efficiency
- More importantly, the SME noted that the responses fail to capture the nuances of stress and task types

Prompt	Cosine Similarity		
	P3.1	P3.2	P3.3
P3.1	—	76.6%	78.7%
P3.2	76.6%	—	78.6%
P3.3	78.7%	78.6%	—

Hallucinations

IBM defines hallucination as “a phenomenon wherein an LLM perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate.”¹

- Hallucinations are one of the biggest areas of concern in use of LLMs
- Techniques such as Retrieval-Augmented Generation (RAG) can reduce the incidence of hallucination
- Hallucination rates for current LLMs range from 0.7% to 29.9%²
- All models hallucinate
- It is not difficult to present prompts that result in hallucinations
- AiVA is based on GPT-4o, which has a 1.5% hallucination rate

¹<https://www.ibm.com/think/topics/ai-hallucinations>

²<https://github.com/vectara/hallucination-leaderboard>



Hallucination Example

P: What are the top 5 publications on micro reactors?

As of my last update in October 2023, I cannot provide real-time data or the very latest publications. However, here are five highly regarded publications and resources on microreactors up to that time:

- 1. “Microreactor Technology and Process Intensification” by Thomas Wirth - This book provides comprehensive coverage of microreactor technology and its applications in chemical processes.*
- 2. “Micro Process Engineering: Fundamentals, Devices, Fabrication, and Applications” by Norbert Kockmann - This publication offers a detailed introduction to the fundamentals and applications of micro process engineering, including microreactors.*
- 3. “Microreactors in Organic Chemistry and Catalysis” edited by Thomas Wirth - This edited volume includes various chapters on the use of microreactors in organic synthesis and catalytic processes.*
- 4. “Microreactors: Design, Fabrication, and Applications” edited by Wolfgang Ehrfeld, Volker Hessel, and Holger Löwe - This book is a foundational text that covers the design, fabrication, and various applications of microreactors.*
- 5. Research Articles in Journals: Look for specific articles in journals such as Chemical Engineering Journal, Lab on a Chip, and AIChE Journal. These journals often publish cutting-edge research on microreactor technology. For the most recent publications, I recommend searching databases like PubMed, IEEE Xplore, Scopus, or Google Scholar with keywords related to microreactors.*

Hallucination: Check the Details

1. *“Microreactor Technology and Process Intensification”* by **Thomas Wirth** - This book provides comprehensive coverage of microreactor technology and its applications in chemical processes.
 - This book was edited by Yong Wang and Jamelyn D. Holladay. Thomas Wirth does not appear to be an editor or contributing author.
4. **“Microreactors: Design, Fabrication, and Applications”** edited by Wolfgang Ehrfeld, Volker Hessel, and Holger Löwe - This book is a foundational text that covers the design, fabrication, and various applications of microreactors.
 - This book does not appear to exist.

Conclusions

- LLMs are powerful tools
- As with any tool, LLMs should be used in a thoughtful manner
 - Use LLMs specifically designed to reduce hallucinations
 - Submit multiple similar prompts to gauge answer consistency
 - Ask the LLM how confident the response is
 - Avoid prompt ambiguity
 - Ask the LLM if any additional information is needed to accurately answer the question of interest
 - Fine-tuning a LLM may be helpful/necessary in specific domains
 - Maintain a questioning attitude: Confidence does not imply accuracy!

WITH GREAT POWER
(WHICH INCLUDES TINKERING WITH LARGE LANGUAGE MODELS)

COMES GREAT RESPONSIBILITY



Sustaining National Nuclear Assets

Vivek.Agarwal@inl.gov

lwrs.inl.gov