Generative Artificial Intelligence Nuclear Evaluation (GAINE): A Structured Decision Framework to Support Large Language Model Evaluations in Nuclear Power Applications

Ronald Boring, Zach Spielman, Anna Hall, Torrey Mortenson, Rachael Hill, and Scott Bowman

Idaho National Laboratory, Idaho Falls, Idaho, U.S.A.

Abstract: As large language models (LLMs) are deployed for new applications, specifically safety-critical domains like nuclear power, scrutable evaluation methods for those LLMs must be used. This paper introduces the Generative Artificial Intelligence Nuclear Evaluation (GAINE) structured decision framework for applying verification, validation, and benchmarking across the LLM lifecycle phases of design, development, deployment, and monitoring. For each of these evaluation types and phases, there are both system and human methods, and this paper identifies when human-in-the-loop evaluations of LLMs are beneficial. Combined across evaluation phases and types, GAINE helps ensure the appropriateness and completeness of LLM evaluations.

1. INTRODUCTION

The advent of large language models (LLMs) and generative artificial intelligence (AI) in nuclear power has brought with it the need to evaluate the performance and efficacy of these applications. LLMs are finding uses within nuclear power, from licensing to operations, and from digital engineering to procedure development and deployment [1]. There is a push to improve accuracy, trustworthiness, generalizability, and appropriateness of the products of LLMs [2] while minimizing biases and hallucinations [3-4]. These qualities are established and confirmed through evaluations of the LLMs, such as verifying that the data used to fine-tune the models result in correct training, validating that the LLM performs tasks correctly, or benchmarking the performance of one LLM against another.

Nuclear applications introduce safety consequences and the explicit need to minimize the risk of poor LLM outputs. The extensive number of possible evaluations available across the lifecycle of general LLM development does not make this process easier. Evaluation methods are proliferating [5], and the best evaluations may be difficult to pick out amid a changing landscape. Even when evaluation methods are promising, these methods may not have been vetted for nuclear-specific applications.

Knowing how to select the right evaluations across the development lifecycle is a universal challenge of LLMs. Evaluating LLM models still has many gaps and lacks common standards and practices [6]. In this context, here we present a framework for considering evaluations across the development lifecycle for nuclear LLM applications. While this approach is tailored to nuclear energy, providing a systematic framework for selecting the right kinds of evaluations at different phases of development is generalizable to non-nuclear AI applications. Additionally, LLM evaluation tends to center on automated validation and comparison of models [7]. This paper helps LLM developers identify when human-in-the-loop evaluations may be necessary or beneficial.

2. A FRAMEWORK FOR LARGE LANGUAGE MODEL EVALUATIONS

2.1. GONUKE as a Starting Point

The Guideline for Operational Nuclear Usability and Knowledge Elicitation (GONUKE) [8] was originally developed as a simple structured decision framework for nuclear power plant evaluation. It

here serves as the foundation for developing an LLM evaluation framework. The purpose of GONUKE is to catalog different available human factors evaluation types across the development lifecycle of nuclear technologies, e.g., selecting appropriate human-centered evaluations across the development of a new reactor control room. The goal of such evaluations is to determine the optimal design based on human use of the system, which supports both assuring safety and streamlining regulatory approval.

Table 1: The GONUKE Structured Decision Framework for Selecting Evaluation Methods [8]

Evaluation Type	Evaluation Phase: Pre-Formative (*Planning and Analysis*[1])	Formative (*Design*[1])	Summative (*Verification and Validation*[1])	Post-Summative (*Implementation and Operation*[1])
Verification *(Expert Review)*	1 Design Requirements Review	2 Heuristic Evaluation	3 System Verification	4 Requalification against New Standards
Validation *(User Study)*	5 Baseline Evaluation	6 Usability Testing	7 Integrated System Validation	8 Operator Training and Operations Trending
Epistemiation *(Knowledge Elicitation)*	9 Cognitive Walkthrough (Task Analysis)	10 Operator Feedback on Design	11 Operator Feedback on Performance	12 Operator Experience Reviews

[1]Corresponds to the phases in NUREG-0711.

The GONUKE framework includes two dimensions, as depicted in Table 1, covering evaluation phases and types. The framework allows system designers to identify suitable evaluation types depending on the phase of development. The evaluation *phases* are represented on the horizontal axis:

- Pre-formative
- Formative
- Summative
- Post-Summative

The formative-summative distinction stems from education [9] but is widely adopted in human factors and user experience contexts. In education, *formative* is the actual teaching and learning phase. Similarly, in system design, formative refers to the conceptual phases when the design is still occurring, meaning the concept and design of the system are still being formed. In contrast, *summative* is at the conclusion, no longer forming the knowledge in education or the design of the system. In education, summative may refer to the examination phase; in system design, summative refers to the deployment and operational phases of the product lifecycle such as final readiness reviews before deployment or, once deployed, ongoing performance monitoring.

As depicted in Table 1, the evaluative phases in GONUKE align with the four phases in NUREG-0711, *Human Factors Engineering Program Review Model* [10], the human factors process guide for the U.S. Nuclear Regulatory Commission. The four phases of NUREG-0711 include: Planning and Analysis, Design, Verification and Validation (V&V), and Implementation and Operation. The GONUKE framework has allowed better incorporation of formative evaluation than the typical summative metrics emphasized in NUREG-0711. For example, evaluation is identified almost exclusively as integrated system validation (ISV) in NUREG-0711, corresponding to Box 7 in GONUKE. GONUKE

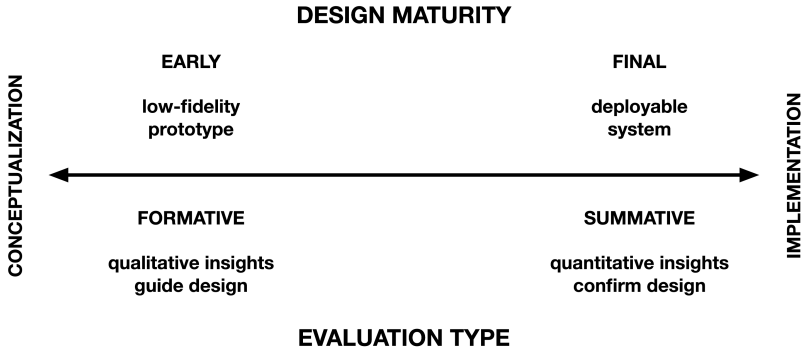
demonstrates 11 other types of human factors evaluations that may be used to inform and confirm the design of a system across its lifecycle.

The vertical axis of GONUKE covers three *types* of evaluations:

- Verification
- Validation
- Epistemiation

Verification seeks to ensure that the system is built right, while validation seeks to ensure the system works as intended for its users. In GONUKE, verification refers to comparison against a standard, while validation refers to demonstration of a performance result [11]. Practically speaking for human factors in nuclear power applications, *verification* means a subject matter expert evaluates the human-machine interface against an accepted reference like a usability heuristic (e.g., [12]) during formative evaluations or a human factors standard (e.g. NUREG-0700 [13]) during summative evaluations. *Validation* in this definition refers to human-in-the-loop studies, whereby the actual performance of users of the system is recorded. As depicted in Figure 1, earlier in the development lifecycle, qualitative validation measures such as operator thinkalouds during scenario runs may help inform the design [14]. Later in the lifecycle, objective human performance measures such as task completion rates and times confirm that the finalized design can be used successfully by human users.

Figure 1: Changes in Evaluation Type across the Design Lifecycle



GONUKE's final evaluation type, *epistemiation*, is a unique artefact of control room modernization studies. Epistemiation is about capturing knowledge of the existing user base needed for the new design [15]. For example, in an analog control room typical of many currently operating nuclear power plants, most operations are performed manually by humans, but the goal of modernization is to capture the human process knowledge to build a digital control system. Special methods for eliciting this knowledge differ from the methods used to verify or validate a system. However, epistemiation is not a universal evaluation construct applicable to all design or development applications.

Verification and validation in GONUKE are about selecting evaluation options whenever they are needed, not prescribing a particular evaluation workflow. The purpose of GONUKE is to guide the human evaluator in determining what evaluations are needed and classifying what purpose those evaluations serve. In some cases (e.g., [16]), GONUKE may be used to provide recommended methods as a shortlist for specific evaluation contexts. A common approach is to use verification at the conceptual phase and validation at the implementation phase. However, GONUKE identifies V&V activities that can occur at *any* stage of development.

GONUKE has been extended to encompass graded approaches to evaluation [17]. In graded approaches, the progression of the lifecycle does not consider every possible type of evaluation. Instead, the types of evaluations are modular, driven by the system development needs. GONUKE helps to downselect which specific evaluations are actually needed and which can be skipped in particular

contexts, whereby skipping unnecessary evaluation steps saves time and costs. In a graded approach, the level of effort expended for evaluations is commensurate with the level of risk for the system being implemented.

GONUKE has also been used for multi-stage evaluation [18]. Multi-stage validation entails validating specific systems more than once, typically by adding new features as the product matures across the lifecycle. This approach is common in iterative design-evaluation cycles in user-centered design [19]. GONUKE uniquely goes beyond the typical focus of multi-stage validation to include multi-stage *evaluation*, including both multi-stage validation and verification. This unique combination allows the evaluation type to pivot across the lifecycle as needed, such as the different human factors evaluations prescribed in the ANSI/HFES-400 standard, *Human Readiness Level Scale in the System Development Process,* [20] as the design of the system matures from conceptual, to demonstration, to deployment.

A graded approach may be used in a multi-stage manner, in which different types of evaluations are used in each iteration. For example, a complete cross-section of GONUKE evaluations might be used in the first design iteration of a system, but more selective evaluations are employed as the design is refined, because only changes need to be tested. Alternately, a subset of evaluations may inform the formative stages of design, while a complete set of evaluations may be used for safety thoroughness as the design reaches maturity and deployment.

2.2. GAINE for LLM Evaluations

GONUKE has proven a successful way to guide the selection of human factors evaluations for nuclear power applications. The same general approach of classifying the phases and types of evaluations provides a useful framework for selecting evaluations during LLM development. To prevent possible confusion with the control room evaluations that are the staple of GONUKE, this variant is given a new name to reflect its focus on AI: Generative Artificial Intelligence Nuclear Evaluation (GAINE). The GAINE structured decision framework identifies types of evaluations that can be deployed across the lifecycle of LLM development. Table 2 provides a list of questions that drive each phase and type of evaluation in GAINE.

Table 2: GAINE Objectives by Evaluation Phase and Type

Evaluation Type	Evaluation Phase: Data	Development	Deployment	Monitoring
Verification *(Built right?)*	Are the data correct and processed correctly?	Are the model, training, and alignment correct?	Does the system work as intended and within guardrails?	Does the system continue to operate as intended?
Validation *(Works as intended?)*	Are the data representative of real-world use?	Does the model perform the task correctly?	Does the system work for users for actual tasks?	Does the system continue to work for users?
Benchmarking *(How does it compare?)*	Are the data quality and coverage comparable?	How does model compare to prior versions and competitors?	How does system compare to previous releases?	Has system drifted vs. historical performance?

The horizontal axis of GAINE encompasses four phases of LLM development:

- Data
- Development
- Deployment
- Monitoring

These four stages are simplified but readily cross-walk to other more expansive lifecycle models like ISO/IEC Standard 5338 [21]. While individual LLM development efforts may feature additional pipeline steps, GAINE phases broadly cover the fundamental aspects of preparing the data inputs, building the LLM, turning the model into a usable system, and keeping the model working and improving it. Some phases may have multiple iterations within a single phase, e.g., Development of an LLM may involve separate activities for initial training and fine-tuning, both of which benefit from evaluations to ensure the model is meeting its performance objectives.

The types of evaluation along the vertical axis consist of:

- Verification
- Validation
- Benchmarking

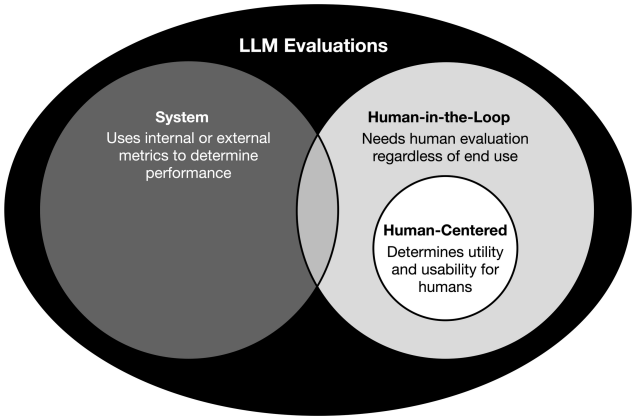
where *verification* confirms the model is built right, *validation* ensures that the model works as intended, and *benchmarking* compares performance relative to other models. Note that benchmarking is a comparative form of V&V and may use many of the same underlying evaluation methods as V&V.

Each evaluation type can be subdivided into two categories:

- System
- Human

System here refers to the ability to evaluate automatically without the need for human intervention, e.g., applying semantic similarity comparisons between different prompt responses. *Human* refers to human checking of the LLM, e.g., when subject matter experts rate the completeness, truthfulness, or accuracy of LLMs across the four lifecycle phases.

Figure 2: System vs. Human-in-the-Loop vs. Human-Centered Evaluations



Human-in-the-loop evaluations as discussed in GAINE differ from human-centered evaluations, whose focus is primarily on ensuring the LLM is usable by humans (see Figure 2). Human-in-the-loop evaluations are a process requiring human input to ensure the completeness of evaluations, whereas human-centered evaluations represent a usability goal to ensure satisfactory end use by humans. GAINE's human-in-the-loop methods may be used to support the development of human-centered LLMs, but that is not its main purview.

While the focus of this paper is on evaluation, human or otherwise, realistically no amount of evaluation can compensate for a poorly understood or modeled workflow. Any human task that an LLM is enlisted to enhance, simplify, or automate will benefit from the application of human factors methods such as task and work analyses, which systematically map the activities that are being performed. This human factors precursor to model development precedes other stages

Table 3: The GAINE Structured Decision Framework for Selecting LLM Evaluation Methods

Evaluation Type		Evaluation Phase: Data	Development	Deployment	Monitoring
Verification *(Is the LLM built right?)*	**System**	**1a** Automated Verification Against Data Requirements	**2a** Automated Verification of Model Implementation	**3a** Automated Conformance Testing	**4a** Automated Requalification Testing
	Human	**1b** SME* Data Requirements/ Traceability Review	**2b** SME* Design Requirements Review	**3b** SME* Operational Readiness Review	**4b** SME* Periodic Requalification/ Compliance Review
Validation *(Does the LLM work as intended?)*	**System**	**5a** Automated Data Quality Checks	**6a** Automated Model Build Checks	**7a** Automated Deployment Readiness Testing	**8a** Continuous Automated Monitoring
	Human	**5b** User Data Readiness Review	**6b** User Model Tuning Review	**7b** User Acceptance/ Workflow Testing	**8b** Periodic User Workflow Audit
Benchmarking *(How does the LLM perform relatively?)*	**System**	**9a** Automated Data Comparison	**10a** Automated Model Benchmarking	**11a** Automated Benchmark in Realistic Workflow	**12a** Longitudinal Benchmark Tracking
	Human	**9b** SME* Comparative Corpus Review	**10b** SME* Pairwise Model Comparison	**11b** Human-in-the-Loop Comparative Study	**12b** Longitudinal Organizational Impact Assessment

*Subject Matter Expert

and is essential for creating the scaffold upon which the LLM is built across the lifecyle. Only after the LLM elements are assembled can they be meaningfully evaluated.

Table 3 outlines methods across GAINE's evaluation phases and types. Note that, like GONUKE, the methods presented are high level families of evaluations and do not provide the specific evaluations or metrics. For example, Validation during Development (i.e., Boxes 6a and 6b) may suggest specific methods like perplexity tests for the system and rating of prompt-response sets by humans. However, the evaluator will find the GAINE framework most useful for identifying *when* and *what* evaluations are needed rather than *which* specific evaluation methods are best suited or *how* to carry out those evaluations. Future guidance will address the selection of specific methods once the category of evaluation is determined.

GAINE may be used as a step-by-step pipeline for each phase of LLM development, identifying appropriate evaluations to use along the way. In some safety-critical contexts, it may be desirable or necessary to consider the full cross-section of evaluation types across the lifecycle phases, ensuring completeness and minimizing risk. In other cases, a graded approach may be appropriate. As noted, GAINE does not prescribe a single pipeline for all applications. Like GONUKE, GAINE supports a modular, graded approach in which only subsets of evaluations are needed. For example, a small language model application with a narrow, non-safety range of use may not require multi-stage evaluations. Alternately, a well vetted and established LLM that is receiving a slight refinement as a minor version update may only warrant a limited set of evaluations to ensure the model has not degraded. In such cases, the developer can use GAINE to select the most appropriate evaluations for that use case.

3. EXAMPLE NUCLEAR USE CASES

Following are two simplified examples to demonstrate how GAINE can be applied to nuclear LLM applications.

3.1. Legacy Digitization of Nuclear Data as Input into LLMs

Under a project called Pantheon [22], Idaho National Laboratory is currently in the process of digitizing records related to designing, building, and operating 52 historic research reactors [23]. These records can serve as an invaluable resource embedded in an LLM for lessons learned to inform future research and commercial reactor deployment. There are numerous challenges related to making these records accessible, from ensuring legacy paper and digital records are of good quality as inputs, to determining any restrictions on the use of such data, to properly categorizing the data. Such considerations related to the Data phase questions in Table 2 include: Are the data correctly processed? Are the data representative and generalizable? Are the data sets across reactor generations comparable?

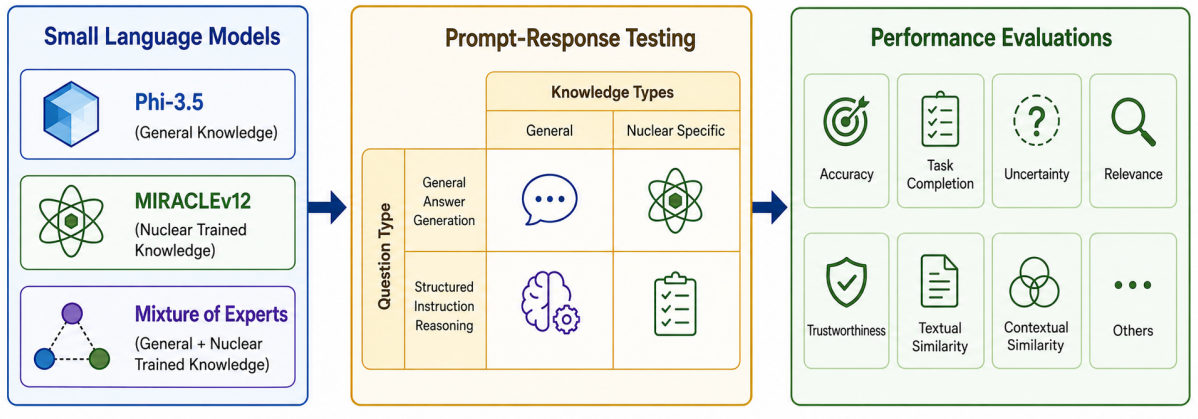
Because the data sources are in many cases first of a kind, predating standardized taxonomies and ontologies, the ability of the system to automatically evaluate the data inputs is limited. In this case, the GAINE framework provides pointers for counterpart human evaluation activities when an automated system evaluation cannot proceed. Before automated validation of the data (GAINE Box 5a in Table 3) can proceed, it first requires human subject matter experts to review and classify the data (GAINE Box 1b in Table 3). Especially in novel applications of LLMs like nuclear reactor data curation, default evaluation schemes more typically used in LLM development may not be available. GAINE helps the LLM developer determine alternatives when more common evaluations are not feasible.

3.2. Nuclear Small Language Model Evaluations

Within the U.S. Department of Energy's Light Water Reactor Sustainability Program, there is a project to investigate nuclear-trained small language models (SLMs) like Machine Intelligence for Review and Analysis of Condition Logs and Entries (MIRACLE) [24]. MIRACLE has been adapted to find and summarize information in the U.S. Nuclear Regulatory Commission's Agencywide Documents Access and Management System (ADAMS) public document repository. Recent efforts have evaluated different versions of MIRACLE across representative use cases. These evaluations considered both human-in-the-loop evaluations related to user trust [25] and automated evaluations that can be performed by the system. These evaluations have covered (see Figure 3):

- *Verification of the Data:* This encompasses the proper categorization of information in ADAMS to ensure it is representative of the way human users classify information, e.g., ensuring that information related to corrective actions is not intermixed with design basis documentation. These represent GAINE Boxes 1a (Automated Data Verification) and 1b (Subject Matter Expert Data Reviews).
- *Validation during Development:* The development of the SLM models was validated against different use cases, e.g., general knowledge vs. nuclear-specific queries, to determine the SLM was able to complete the tasks correctly. These evaluations represent GAINE Boxes 6a (Automated Model Checks) and 6b (User Model Tuning Review).
- *Benchmarking for Deployment:* Finally, the actual deployment to the models was compared for their utility in answering user queries and in performing objective accuracy measures of model performance. In this case, different versions of the model were compared (e.g., a fine-tuned nuclear model that replaced general foundation model knowledge vs. a version that retained general foundation model knowledge but added nuclear-specific knowledge using low-rank adapters). These evaluations represent GAINE Boxes 11a (Automated Benchmark) and 11b (Human-in-the-Loop Comparative Study).

Figure 3: Evaluating Small Language Models for General and Nuclear Knowledge



This study provided a useful demonstration of how GAINE aligns and adapts across the lifecycle of the SLM. The findings, briefly mentioned, demonstrated that system metrics may not always align with human metrics. In other words, it's possible for a system to perform well on standard system evaluations but still reveal shortcomings in the actual human use and trust in that system. The two dimensions should not be treated interchangeably; moreover, success

in a system metric does not guarantee suitability to human use or acceptance and trust by the user.

Indeed, research on trust transfer in automated systems [26] suggests that failures on behalf of an automated system will translate into long-term distrust of the system. This research may suggest a failure to calibrate trust in SLM or LLM systems could result in long-term low adoption of these technologies and missed opportunities to deploy reasonable AI systems to support nuclear plant operations. GAINE corroborates the importance of human-in-the-loop evaluations and identifies counterpart human evaluations to system evaluations. Evaluating AI systems fully ensures not only that they are performing as needed but also that they will actually be accepted and used by humans.

4. WHEN HUMAN-IN-THE-LOOP EVALUATION IS NEEDED

As LLMs increasingly approximate human psychology, the measures used to evaluate LLMs must likewise parallel the measures used in psychology. The field of psychology understands that not all processes are cognitively accessible, meaning human consciousness cannot introspect insights on the underlying mechanisms [27]. External measures must be brought to understand those processes. Likewise, with LLMs, external measures must be brought to the evaluation process. External measures most often take the form of human subject matter experts evaluating the outputs of the LLMs.

Human-in-the-loop evaluation is necessary when the evaluation question depends on human judgment, domain expertise, or user acceptance, rather than only measurable system behavior. Human evaluations are indispensable for:

- Categorization of novel or poorly fit data that do not readily align with an existing system ontology
- Determination of correctness or completeness of model outputs when no objective measures are available for system evaluations
- Signing off on domain-specific model knowledge by subject matter experts to ensure over-generalization and hallucinations are not present
- Relatedly, assurance that models do not produce invisible failures [28] that result when models inadvertently hide lack of knowledge through plausible but incorrect outputs
- Certification of model outputs that will be used in a safety-critical or operational context with potential negative consequences if there are erroneous outputs
- Resolution and troubleshooting of intermittent faults where models exhibit poor performance only in certain contexts but show strong performance otherwise
- Reasonableness checks for model reuse in new contexts to prevent application beyond the scope or ability of the model
- Operational equivalence testing when models are used to automate previously human tasks
- Alignment of model outputs to fit human-centered use (as depicted in Figure 2).

Human-in-the-loop evaluations should be used whenever model correctness and completeness, fit and domain-specificity, safety, or human usability are required.

5. CONCLUSION

This paper has expanded the evaluation type and phase framework originally used for human factors in GONUKE to a broader framework to guide LLM development. GAINE identifies

phases in the LLM development lifecycle that require or benefit from evaluations. Additionally, GAINE differentiates separate goals and processes for verification, validation, and benchmarking. Finally, GAINE suggests complementary system and human evaluations. System evaluations tell whether the LLM behaves well against a measurable technical target, but human evaluations tell whether the model outputs are adequate, meaningful, and safe in the intended work context.

Several concepts originally explored in GONUKE may benefit from more detailed treatment and adaptation in GAINE. For example, as noted in the discussion around Figure 2 in this paper, human-centered AI is essentially focused on usability (i.e., the *U* in GONUKE), and ensuring the usability of LLMs likely requires standalone guidance beyond the current general GAINE overview. Additionally, the concept of knowledge acquisition is a critical component of creating LLMs that emulate human expertise [8], and aspects of knowledge elicitation (i.e., the *KE* in GONUKE) should be adapted and expanded for use in LLMs.

Future work to refine GAINE will explore more use cases for LLM development in nuclear power applications. Such use cases will allow the authors to develop nuanced guidance for evaluation measures, including illustrating how such measures can be part of an overall evaluation harness in LLMs.

Future GAINE guidance should also more precisely define when it is appropriate and necessary to use human experts to judge fitness of outputs. As part of this refinement, guidance on specific evaluation measures should be provided to assist not only with identifying the classes of evaluations that are needed but also pinpointing specific measures that meet quality criteria for nuclear power.

As the GAINE decision support framework demonstrates, it's important to consider the limitations of system evaluations and adequately include human evaluations across the lifecycle of LLM development. GAINE provides a useful starting point to prioritize the right evaluations during LLM development, particularly when automated metrics must be supplemented by human expertise and judgment.

Acknowledgements

This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory, an agency of the United States Government. Neither the United States 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. Idaho National Laboratory is a multi-program laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517. This research was funded through the Laboratory Directed Research and Development program at Idaho National Laboratory.

The authors acknowledge the use of generative artificial intelligence tools during preparation of this manuscript for refinement of exploratory ideas. These tools were not used as authoritative sources. All AI-assisted content was critically reviewed, fact-checked, and edited by the authors, who take full responsibility for the manuscript.

References

[1] U.S. Nuclear Regulatory Commission, "Advancing the use of artificial intelligence at the U.S. Nuclear Regulatory Commission," SECY-24-0035, 2024.

[2] Y. Yamani, A. Jackson, C. Kovesdi, J. C. Joe and J. Mohon, "Trustworthiness and trust: Identifying factors that drive successful human-AI interaction in nuclear power plant applications," Nuclear Plant Instrumentation and Control & Human-Machine Interface Technology (NPIC&HMIT 2025), pp. 330-339, 2025.

[3] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang and N. K. Ahmed, "Bias and fairness in large language models: A survey," Computational Linguistics, vol. 50, no. 3, pp. 1097-1179, 2024.

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto and P. Fung, "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, article 248, pp. 1-38, 2023.

[5] M. Mohammadi, Y. Li, J. Lo and W. Yip, "Evaluation and benchmarking of LLM agents: A survey," in Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, pp. 6129-6139, 2025.

[6] A. Reuel, A. Hardy, C. Smith, M. Lamparth, M. Hardy and M. J. Kochenderfer, "BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices," Advances in Neural Information Processing Systems, vol. 37, pp. 21763-21813, 2024.

[7] National Institute of Standards and Technology, "Practices for automated benchmark evaluations of language models," NIST AI 800-2, Initial Public Draft, U.S. Department of Commerce, 2026.

[8] R. L. Boring, T. A. Ulrich, J. C. Joe and R. T. Lew, "Guideline for operational nuclear usability and knowledge elicitation (GONUKE)," Procedia Manufacturing, vol. 3, pp. 1327-1334, 2015.

[9] M. Scriven, "The methodology of evaluation," in Curriculum Evaluation, R. E. Stake, Ed. Chicago, IL, USA: Rand McNally, 1967.

[10] J. M. O'Hara, J. C. Higgins, S. A. Fleger and P. A. Pieringer, "Human factors engineering program review model," NUREG-0711, Rev. 3, U.S. Nuclear Regulatory Commission, 2012.

[11] R. B. Fuld, "Verification and validation: What's the difference?" Ergonomics in Design, vol. 5, no. 3, pp. 28-33, 1997.

[12] J. Nielsen, "Enhancing the explanatory power of usability heuristics," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 152-158, 1994.

[13] J. M. O'Hara, S. A. Fleger, N. Hughes Green and S. Morrow, "Human-system interface design review guidelines," NUREG-0700, Rev. 4, U.S. Nuclear Regulatory Commission, 2026.

[14] K. A. Ericsson and H. A. Simon, Protocol Analysis: Verbal Reports as Data, Rev. ed. Cambridge, MA, USA: MIT Press, 1993.

[15] R. L. Boring, R. Lew and T. A. Ulrich, "Epistemiation: An approach for knowledge elicitation of expert users during product design," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 60, pp. 1699-1703, 2016.

[16] C. Kovesdi, J. Joe and R. Boring, "A guide for selecting appropriate human factors methods and measures in control room modernization efforts in nuclear power plants," in Advances in Intelligent Systems and Computing, vol. 787, pp. 441-452, 2018.

[17] R. L. Boring, T. A. Ulrich and R. Lew, "Putting GONUKE into practice: Considerations for human factors evaluations," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 65, pp. 618-622, 2021.

[18] R. L. Boring, T. A. Ulrich and R. Lew, "Multi-stage evaluation using GONUKE," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 68, pp. 822-828, 2024.
[19] J. D. Gould and C. Lewis, "Designing for usability: Key principles and what designers think," Communications of the ACM, vol. 28, no. 3, pp. 300-311, 1985.
[20] Human Factors and Ergonomics Society, Human Readiness Level Scale in the System Development Process, ANSI/HFES 400-2021, Washington, DC, USA: Human Factors and Ergonomics Society, 2021.
[21] International Organization for Standardization and International Electrotechnical Commission, Information Technology—Artificial Intelligence—AI System Life Cycle Processes, ISO/IEC 5338:2023(E), Geneva, Switzerland: ISO/IEC, 2023.
[22] J. Brownlee, N. Hergesheimer, A. Morin, J. Rogers, M. Ross Kunz, J. Lee, and J. Browning. "Standards and Data Architectures for Interoperable Digital Twins: An Ontology-Driven Approach with DeepLynx." Submitted May 2026.
[23] S. M. Stacy, Proving the Principle: A History of the Idaho National Engineering and Environmental Laboratory, 75th Anniversary Edition, 1949-2024. Idaho Falls, ID, USA: Idaho Operations Office, U.S. Department of Energy, 2024.
[24] C. Krome, A. Al Rashdan, J. Wadsworth and K. Giraud, "MIRACLE: Machine learning for screening nuclear plant conditions," Idaho National Laboratory, copyrighted software, 2020.
[25] C. Kovesdi, Y. Yamani, A. Jackson and L. Lin, "Application of signal detection theory in evaluating trust of information produced by large language models," Proceedings of the Human Factors and Ergonomics Society, vol. 69, no. 1, pp. 582-587, 2025.
[26] M. Wischnewski, N. Krämer and E. Müller, "Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions," in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023.
[27] N. Block, "Perceptual consciousness overflows cognitive access," Trends in Cognitive Sciences, vol. 15, no. 12, pp. 567-575, 2011.
[28] C. Potts and M. Sudhof, "Invisible failures in human-AI interactions," arXiv:2603.15423, 2026.