Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria

doi: 10.3850/ESREL2026061419_esrel26-p26141-cd

Human-AI Collaboration in Nuclear Fault Diagnosis: Operator Evaluation of LLM-Based Assistance

Jens-Patrick Langstrand
Applied Data Science, Institute for Energy Technology (IFE), Norway. E-mail: jens.patrick.langstrand@ife.no

Aleksander B. Gundersen
Human and Organizational Factors, IFE, Norway. E-mail: aleksander.gundersen@ife.no

Espen Nystad
Human and Organizational Factors, IFE, Norway. E-mail: espen.nystad@ife.no

Per Øivind Braarud
Human and Organizational Factors, IFE, Norway. E-mail: per.oivind.braarud@ife.no

Robert McDonald
Control Room and Interaction Design, IFE, Norway. E-mail: robert.mcdonald@ife.no

Daniel Odéen
Control Room and Interaction Design, IFE, Norway. E-mail: daniel.odeen@ife.no

Nuclear power plant operators face complex fault diagnosis challenges that require rapid interpretation of information across multiple systems and parameters. This study investigates the potential of large language models (LLMs), augmented with plant-specific documentation, to support operator diagnostic reasoning. A prototype LLM-based diagnostic assistance system was developed and evaluated in a nuclear control room simulator. Six experienced operators from two crews participated in a pilot study involving an intentionally difficult fault scenario. Once the fault had sufficiently manifested, the simulation was paused and the crew was granted access to the LLM. The model had access to operating procedures, technical manuals, and plant terminology, enabling it to generate diagnostic recommendations in response to operator prompts such as ”Steam pressure increasing, feedwater flow decreasing.” Operators provided qualitative feedback on system usefulness, presentation, and potential integration into operational practice. They viewed the LLM as a supplementary information source rather than an autonomous system. Identified use cases included diagnostic support, document retrieval, access to historical data, and operator training. This study represents an initial step in an iterative development process guided by operator feedback. The results demonstrate the feasibility and acceptance of LLM-based diagnostic assistance in nuclear control environments and outline a path toward transparent, trustworthy, and human-centered AI tools that enhance safety and efficiency while preserving human authority.

Keywords: Large language models, nuclear power plant operations, fault diagnosis, control room simulation, decision support systems, human-centered AI

1. Introduction

Nuclear power plant control room operators are responsible for maintaining safe and stable plant operation under a wide range of conditions, including abnormal and fault scenarios (IAEA, 2022). Effective fault diagnosis is critical in these situations, as operators must rapidly interpret complex and evolving plant information while adhering to strict safety principles. Although established procedures and decision-support tools provide structured guidance, diagnosing atypical or ambiguous faults remains challenging (Braarud and Kirwan, 2011; Roth et al., 1994). Locating and synthesizing relevant information from extensive documentation while an event unfolds can impose substantial cognitive load and slow diagnostic reasoning.

Recent advances in large language models (LLMs) have demonstrated capabilities in natural-language understanding and information synthesis that may offer new forms of supplementary operator support. While LLMs are trained on broad general-domain knowledge, their use in nuclear operations requires grounding in plant-specific documentation. When augmented with operating procedures, technical manuals, and domain terminology, LLM-based systems may support operators by retrieving relevant information and synthesizing responses that aid diagnostic reasoning.

2. Background and related works

This section reviews prior research on fault diagnosis and operator support in nuclear control rooms, including cognitive aspects, procedures, information systems, and diagnostic aids. It also summarizes emerging work on large language models (LLMs) in safety-critical domains and human-centered AI principles relevant to supplementary decision-support design. Together, this background establishes the context for the present exploratory study.

2.1. *Fault Diagnosis, Procedures, and Diagnostic Support in Nuclear Control Rooms*

Fault diagnosis in nuclear control rooms is widely recognized as a cognitively demanding activity requiring operators to integrate plant knowledge, procedural guidance, and real-time process information under uncertainty and time pressure. Operators engage in hypothesis-driven reasoning, continuously updating their mental models as plant conditions evolve (Roth et al., 1994). Abnormal plant behavior often manifests through subtle trends and indirect indicators, making it difficult to distinguish between normal transient fluctuations and underlying faults (Mumaw et al., 2000). These characteristics increase susceptibility to cognitive overload and misinterpretation, particularly in novel or unexpected situations.

Operating procedures and technical manuals form the knowledge base supporting operator actions and diagnostic reasoning (O'hara et al., 2000). While procedures provide structured guidance for anticipated conditions, strict adherence may be insufficient during complex or atypical events (Massaiu and Holmgren, 2013). Operators therefore rely on experience-based judgment and cross-reference multiple documents to interpret plant behavior. Although digital systems and electronic procedures have improved accessibility, navigating extensive documentation during time-critical situations remains challenging and can contribute to increased workload and delayed diagnosis.

Diagnostic support tools for nuclear power plant control rooms have evolved from early rule-based expert systems to model-based and data-driven approaches. Reviews of early applications indicate that expert systems developed in the 1980s primarily encoded operator knowledge using deterministic if–then rules for fault diagnosis and emergency response (Qi et al., 2023; Xiao et al., 2024). Subsequent research introduced probabilistic reasoning and explicit system models, including Bayesian networks and directed graphs, to address uncertainty (Qi et al., 2023).

Later work emphasized advanced alarm management and trend visualization to support operator decision-making in modernized control rooms (Boring et al., 2015). More recent studies have explored predictive and prognostic monitoring using data-driven and machine-learning techniques to provide earlier indications of emerging or previously unknown faults (Mendoza and Tsvetkov, 2024). Despite demonstrated benefits in reducing operator workload and improving situational awareness, challenges related to adaptability, validation, and workflow integration have limited widespread deployment, and diagnostic support systems continue to function primarily in an advisory role (Boring et al., 2015; Qi et al., 2023).

2.2. *LLMs and human-centered AI in safety-critical domains*

Recent peer-reviewed studies have explored the use of LLMs as operator support tools in safety-critical and complex control domains. Their capabilities in natural-language interaction, document retrieval, and information synthesis suggest potential value in augmenting human reasoning rather than replacing it. Reported applications span nuclear reactor control rooms, power system operations, robotic teleoperation, industrial drilling control rooms, spacecraft mission control, and laboratory automation (Lee et al., 2025; Choi et al., 2024; Yang et al., 2025; Fei et al., 2025; Zhao et al., 2024; Ferrigno et al., 2024; Carrasco et al., 2025; Xie et al., 2025).

However, concerns regarding hallucination, lack of grounding, and trust calibration remain significant in high-consequence domains (Anh-Hoang et al., 2025). Research on human-centered AI emphasizes transparency, interpretability, and appropriate trust calibration to mitigate automation bias and preserve operator authority (Endsley, 2023; Berretta et al., 2023). In nuclear operations, clear human responsibility and accountability are essential for regulatory compliance (CNSC et al., 2024).

As AI continues to expand the boundaries of traditional operator decision support, it is essential to improve our understanding of the challenges and best practices of human-AI interaction, particularly within LLM-based systems. Understanding how to design systems so both the control room team and the LLM are aligned in their understanding of current situations and the appropriate responses is essential (Kirwan, 2025).

2.3. *Summary and positioning of the present study*

Prior work highlights both the benefits and limitations of diagnostic support in nuclear control rooms. While procedures and diagnostic aids support operator performance, challenges related to information access, cognitive workload, adaptability, and workflow integration persist. LLM technologies offer new possibilities for document-centered information synthesis, yet their suitability for safety-critical nuclear environments remains uncertain.

The present study addresses this gap through an exploratory, simulator-based evaluation of a locally hosted LLM diagnostic support system grounded in plant-specific documentation, focusing on its potential role, interaction modes, and perceived value. It contributes qualitative insight into how licensed operators perceived and interacted with such a system during a challenging fault scenario, informing the design of human-centered AI support for nuclear operations.

3. Operator diagnostic assistance concept and knowledge architecture

In this initial prototype, the LLM was designed strictly as a diagnostic support tool rather than a decision-making or control system. Operators retained full authority, and the system remained idle until explicitly prompted, ensuring it functioned as a supplementary information source rather than an autonomous agent.

Within this role, the system supported diagnostic reasoning by facilitating hypothesis generation, information cross-checking, and navigation of relevant documentation during fault diagnosis. It was not intended to replace procedural execution or operator judgment, but to augment existing diagnostic practices through contextualized information access.

To ensure domain relevance, the LLM was augmented with plant-specific documentation commonly used in nuclear operations, including operating procedures for event-based scenarios (e.g., reactor trips and steam generator tube ruptures), alarm response procedures, and technical manuals describing major plant systems such as the reactor coolant system. Plant-specific terminology, including abbreviations and definitions, was also incorporated to improve response accuracy, as preliminary testing showed potential misinterpretation of domain-specific terms.

Source documents were processed using Docling to convert Word and PDF files into machine-readable Markdown for embedding and ingestion. Text embeddings were generated using the `BAAI/bge-m3` model. Documents were divided into 2000-token chunks with a 200-token overlap.

At runtime, the system employed a retrieval-augmented generation (RAG) approach. Operator prompts were embedded and used to retrieve relevant document chunks through a hybrid approach combining vector-based cosine similarity and BM25 keyword search. Retrieved passages were then re-ranked using the

`BAAI/bge-reranker-v2-m3` model to prioritize the most relevant sources. The top-ranked passages were inserted into the LLM context window together with a structured system instruction directing the model to ground its reasoning in the provided documentation and avoid unsupported or speculative claims. The model then generated a response conditioned on both the operator prompt and the retrieved context. No fine-tuning was performed; domain adaptation relied entirely on retrieval-based grounding.

When available, the documentation passages used during response generation were presented to operators as an expandable citation list at the bottom of the reply. This allowed operators to inspect the referenced source material directly and cross-check the basis for the generated output.

3.1. *Operator assistant frontend and model configuration*

OpenWebUI was used as the foundation for the operator assistance interface, providing a unified interaction point for the diagnostic support system (Timothy J. Baek, 2024). The platform was selected for its open-source nature, customization support, and built-in user management features enabling role-based access to information. Multiple locally hosted LLMs were evaluated via Ollama, including distilled DeepSeek models, Qwen, LLaMA, and Mistral variants. Based on preliminary testing, the Qwen3:8B model was selected due to its performance and response quality for this use case. The model was configured with a control-room-oriented persona and instructed to focus on identifying potential malfunctions and abnormalities grounded in the documentation.

4. Exploratory study methodology

This section outlines the exploratory methodology used to examine the feasibility and perceived usefulness of an LLM-based diagnostic assistance system in a simulated nuclear control room. Using a high-fidelity simulator, a challenging fault scenario, and qualitative data collection, the study focused on how experienced operators interacted with and evaluated the system rather than on quantitative performance.

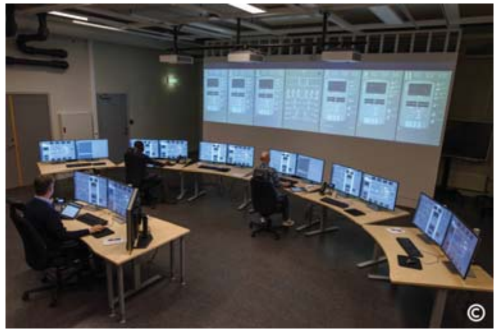

Fig. 1. Layout of the multi-unit SMR control room simulator and operator workstations.

4.1. *Nuclear control room simulator*

The study was conducted in a high-fidelity, full-scope nuclear control room simulator representative of operational environments. A multi-unit small modular reactor (SMR) simulator modeled six simultaneously operating units, providing a complex and realistic diagnostic context.

An intentionally difficult fault scenario was selected to elicit diagnostic uncertainty. The scenario involved a reduction in feedwater heating in one SMR unit, producing gradually developing symptoms distributed across multiple systems. This design required integration of information from multiple sources and reflected conditions in which fault identification is non-trivial.

4.2. *Participants and crew composition*

Six licensed nuclear power plant operators participated in the study. Data were collected between September 15 and 17, 2025, at the Halden Man–Machine Laboratory (HAMMLAB) IFE (2026). Participants were employed at U.S. commercial nuclear power plants and included four reactor operators, one unit supervisor, and one shift manager, with one to twelve years of operational experience. All participants received an overview of study procedures and provided written informed consent.

Participants were organized into two three-person crews with differing levels of prior simulator familiarity. Training was conducted by simulator experts and included hands-on scenarios covering plant operation, control interfaces, computer-based procedures, and multi-unit operation.

During the simulation, the shift supervisor was positioned behind two operators at a workstation with access to all six units and alarm systems. Each operator supervised three units using dedicated reactor and turbine displays, supported by a shared overview display, see figure 1. The LLM-based diagnostic assistance system was hosted on a separate laptop introduced into the control room by the experiment lead.

4.3. *Study procedure and data collection*

Each crew progressed through the scripted fault scenario until a predefined manifestation threshold was reached, at which point the simulation was paused and access to the LLM-based diagnostic assistance system was provided. The system was accessed via the dedicated laptop, with output mirrored on a large monitor to support shared use. Operators interacted with the system using natural-language prompts describing observed plant conditions and trends. The LLM was instructed to respond in a structured format, highlighting probable causes and potential immediate actions grounded in the available documentation. Figure 2 shows an example operator prompt and corresponding system response.

Immediately after the LLM interaction concluded, each crew took part in a separate semi-structured group interview lasting approximately 45 minutes. The interview was conducted by the experimental lead, who guided the discussion using predefined open-ended questions covering system usefulness and usability, trust and reliability, decision support and human judgment, safety and risk perceptions, system limitations, and potential improvements.

5. Operator interview findings

This section presents qualitative findings from operator interactions with the LLM-based diagnostic assistance system, focusing on system use, perceived value and limitations, and workflow integration. Themes observed across both crews inform system design rather than provide quantitative performance evaluation.

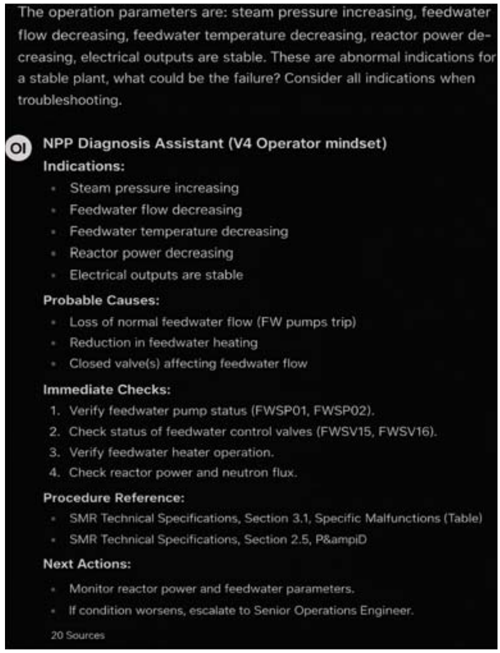

Fig. 2. Example operator prompt of current plant indicators and the corresponding LLM-generated response.

5.1. *Operator use, trust, and perceived role of the LLM during diagnosis*

During the diagnostic task, operators used the LLM to seek explanations for observed plant behavior, test emerging hypotheses, and retrieve relevant documentation. The system was not treated as an authoritative source, but as a supplementary tool for exploring possible fault causes and identifying additional lines of inquiry, which were evaluated against operators' own understanding of the plant. Participants emphasized that the LLM's value lay in surfacing relevant information and offering suggestions, while final diagnostic decisions remained the responsibility of the crew, particularly the shift manager.

Operators expressed caution regarding use of such a system in a nuclear control room, citing concerns about cybersecurity, control over training data and outputs, time criticality, and over-reliance on AI-generated suggestions. Trust was viewed as developing through repeated use and validation, with outputs cross-checked against training, plant parameters, and referenced documentation, often through crew discussion. Conflicting outputs prompted further deliberation. Initial assumptions by some participants that the system functioned like a simulator underscored the need to clearly communicate its capabilities and limitations during training.

Both crews agreed that interaction with the LLM should not distract operators responsible for monitoring or procedural execution. Instead, use of the system was seen as most appropriate for an additional resource, such as a shift technical advisor or extra reactor or senior reactor operator. During normal operations, broader exploratory access was considered acceptable, while decision authority remained with the shift manager, reinforcing the view of the LLM as a collaborative support tool rather than integrated control-room automation.

5.2. *Identified use cases beyond the study scenario*

Beyond real-time fault diagnosis, operators identified several potential applications for LLM-based tools in nuclear operations. These included contingency planning prior to actions that affect plant conditions, preparation for maintenance activities, and assessment of risks associated with simultaneous operations. Participants also viewed the LLM as a promising advanced search interface for accessing system information, historical events, industry operating experience, and procedure content. Finally, operators suggested that incorporating historical trend data could enable the system to support expectations of plant behavior following specific actions, provided the model was sufficiently plant-specific.

6. Discussion: implications for LLM-based operator support

Operator interaction with the LLM suggests that its primary value lies in supporting cognitive processes rather than providing definitive answers. Operators used the system to explore alternative explanations, test emerging hypotheses, and retrieve relevant documentation—behaviors consistent with sensemaking and hypothesis-driven diagnostic reasoning. This aligns with established models of nuclear operator cognition, in which diagnosis is iterative and shaped by evolving information and uncertainty. The LLM functioned as an additional cognitive resource that expanded the space of considered explanations while leaving diagnostic authority with the operators.

Operators consistently framed the LLM as an advisory information source rather than a decision-making authority, reflecting established safety principles emphasizing conservative decision-making, accountability, and human control. Positioning the system as suggestive rather than directive appears critical for acceptance and may help mitigate automation bias. These findings indicate that acceptance of AI-based tools in nuclear operations depends not only on technical capability but also on alignment with existing roles, responsibilities, and decision hierarchies.

6.1. *Trust, transparency, and mental model alignment*

Trust in the LLM was not immediate but was described as something that would develop over time through repeated use and validation. Operators evaluated system outputs by comparing them with their understanding, training, and experience, cross-checking real-time plant parameters, and consulting referenced documentation, often through crew discussion. Outputs that conflicted with expectations were treated as inputs for further deliberation rather than authoritative guidance.

Operators emphasized that trust calibration depends strongly on presentation and transparency. Clear response structure, use of domain-specific terminology, and explicit references to source documentation were viewed as prerequisites for meaningful engagement. Misconceptions about system capabilities—such as assumptions that the LLM performed physical simulations or generated trend plots—highlighted the need for training that clearly communicates system functionality and limitations. These observations reinforce that trust is shaped not only by accuracy, but by transparency and alignment between operator mental models and actual system behavior.

6.2. *Workflow integration and broader operational value*

Operators emphasized that LLM interaction should not interfere with primary monitoring or procedural execution tasks. Instead, interaction was seen as most appropriate for an additional resource, such as a shift technical advisor or extra operator, particularly during abnormal events. This highlights the importance of integrating AI tools in ways that respect existing team structures and attentional demands, supporting collaboration rather than competing for cognitive resources.

Beyond real-time fault diagnosis, operators identified additional promising applications for LLM-based tools, including contingency planning, training, document retrieval, and access to historical operating experience. These use cases involve lower time pressure and risk and align well with LLM strengths in information synthesis and retrieval. Prioritizing such applications may offer a more feasible pathway for early deployment, enabling the development of experience, trust, and governance structures before considering more time-critical operational roles.

7. Limitations and future directions

The study's generalizability is limited by its exploratory nature, small participant sample, reliance on a single simulated fault scenario, and primarily qualitative evaluation. As such, the results should be interpreted as hypothesis-generating rather than evidence of benefit.

Despite these limitations, this work provides a foundation for further investigation of LLM-based support tools in nuclear operations. Future research should consider broader scenario coverage, longer-term evaluations, and complementary quantitative measures of performance and workload, alongside technical advances in document structuring, contextual awareness via simulator data integration, and selective agentic capabilities such as guided information gathering or hypothesis tracking. Furthermore, future research should investigate team collaboration practices and team roles related to LLM-based support. Together, these directions aim to clarify the role of LLM-based systems in supporting safe, effective, and human-centered nuclear plant operation.

8. Conclusions

This exploratory study examined the use of a large language model as a diagnostic assistance tool for nuclear control room operators in a simulated environment. The findings indicate that experienced operators were able to meaningfully engage with the system and perceived it as a useful supplementary source of information during a challenging fault diagnosis task. Operators primarily used the LLM to support hypothesis generation, information retrieval, and exploration of potential fault causes, rather than as an authoritative decision-making system.

Overall, the results suggest that LLM-based diagnostic assistance can be feasible and acceptable when explicitly positioned as a non-autonomous, transparent support tool. Operators emphasized the importance of preserving operator-led reasoning, established procedures, and clear decision authority, and highlighted transparency, traceability to source documentation, and careful workflow integration as critical design requirements. These findings reinforce the need for human-centered, workflow-aware AI design in safety-critical nuclear operations.

References

Anh-Hoang, D., V. Tran, and L.-M. Nguyen (2025, September). Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence 8*, 1622292.

Berretta, S., A. Tausch, G. Ontrup, B. Gilles, C. Peifer, and A. Kluge (2023). Defining human-AI teaming the human-centered way: a scoping review and network analysis. *Frontiers in Artificial Intelligence 6*, 1–28.

Boring, R. L., K. D. Thomas, T. A. Ulrich, and R. T. Lew (2015, January). Computerized Operator Support Systems to Aid Decision Making in Nuclear Power Plants. *Procedia Manufacturing 3*, 5261–5268.

Braarud, P. Ø. and B. Kirwan (2011). Task complexity: What challenges the crew and how do they cope. In A. B. Skjerve and A. Bye (Eds.), *Simulator-based Human Factors Studies Across 25 Years*, pp. 233–251. Springer.

Carrasco, A., V. Rodriguez-Fernandez, and R. Linares (2025, September). Large language models as autonomous spacecraft operators in kerbal space program. *Advances in Space Research 76*(6), 3480–3497.

Choi, S., R. Jain, P. Emami, K. Wadsack, F. Ding, H. Sun, K. Gruchalla, et al. (2024, May). eGridGPT: Trustworthy AI in the Control Room. Technical Report NREL/TP–5D00-87740.

CNSC, U. ONR, and U. NRC (2024). Considerations for developing artificial intelligence systems for nuclear applications. Technical report, Canadian Nuclear Safety Commission, UK Office for Nuclear Regulation, US Nuclear Regulatory Commission.

Endsley, M. R. (2023). Supporting human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior 140*, 1–16.

Fei, H., T. Xue, Y. He, S. Lin, G. Du, Y. Guo, and Z. Wang (2025, July). Large language model-driven natural language interaction control framework for single-operator bimanual teleoperation. *Frontiers in Robotics and AI 12*.

Ferrigno, E., M. Rodriguez, and E. Davidsson (2024, September). *Revolutionizing Drilling Operations: Next-gen Llm-AI for Real-time Support in Well Construction Control Rooms*.

IAEA (2022). Conduct of operations at nuclear power plants. Technical Report SSG-76, International Atomic Energy Agency.

IFE (2026). Hammlab. `https://ife.no/en/laboratory/hammlab/`.

Kirwan, B. (2025, April). Human Factors Requirements for Human-AI Teaming in Aviation. *Future Transportation 5*(2).

Lee, Y. P., J. Cha, Y. Yu, and S. G. Kim (2025, December). Large language model agent for nuclear reactor operation assistance. *Nuclear Engineering and Technology 57*(12), 103842.

Massaiu, S. and L. Holmgren (2013). An empirical investigation of team decision-making with emergency procedures. In *Paper to the American Nuclear Society's 2013 Risk Management Embedded Topical Meeting: Risk Management for Complex Socio-technical Systems*.

Mendoza, M. and P. V. Tsvetkov (2024, June). An intelligent fault detection and diagnosis monitoring system for reactor operational resilience: Unknown fault detection. *Progress in Nuclear Energy 171*, 105167.

Mumaw, R. J., E. M. Roth, K. J. Vicente, and C. M. Burns (2000). There is more to monitoring a nuclear power plant than meets the eye. *Human Factors 42*(1), 36–55.

O'hara, J. M., J. C. Higgins, W. F. Stibler, and J. Kramer (2000). Computer-based procedure systems: Technical basis and human factors review guidance. Technical Report NUREG/CR-6634, US Nuclear Regulatory Commission.

Qi, B., J. Liang, and J. Tong (2023, February). Fault Diagnosis Techniques for Nuclear Power Plants: A Review from the Artificial Intelligence Perspective. *Energies 16*(4).

Roth, E. M., R. J. Mumaw, and P. M. Lewis (1994). An empirical investigation of operator performance in cognitively demanding simulated emergencies. Technical Report NUREG/CR-6208, US Nuclear Regulatory Commission.

Timothy J. Baek (2024). Open webui: User-friendly ai interface. `https://github.com/open-webui/open-webui`.

Xiao, X., J. Liang, J. Tong, and H. Wang (2024, May). Emergency Decision Support Techniques for Nuclear Power Plants: Current State, Challenges, and Future Trends. *Energies 17*(10).

Xie, Y., K. He, and A. Castellanos-Gomez (2025, August). Toward Full Autonomous Laboratory Instrumentation Control with Large Language Models. *Small Structures 6*(8), 2500173.

Yang, X., C. Lin, H. Liu, Q. Wang, and W. Wu (2025, July). Large Language Model as An Operator: An Experience-Driven Solution for Distribution Network Voltage Control.

Zhao, W., L. Li, H. Zhan, Y. Wang, and Y. Fu (2024, December). Applying Large Language Model to a Control System for Multi-Robot Task Assignment. *Drones 8*(12).