Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 2156–2165
https://doi.org/10.54941/ahfe1007027

From Checklists to Chatbots: Reimagining HRA With Generative AI

Michael Hildebrandt and Awwal M. Arigi

Institute for Energy Technology, 1777 Halden, Norway

ABSTRACT

This paper evaluates the capability of Large Language Models (LLMs) to support Human Reliability Assessment (HRA) through a systematic test using the Integrated Human Event Analysis System for Event and Condition Assessment (IDHEAS-ECA) methodology. Using Claude Opus 4.1, we generated Steam Generator Tube Rupture scenarios and subsequently tasked the model with producing a comprehensive HRA analysis, which was then independently reviewed by two IDHEAS-ECA method experts. The LLM demonstrated substantial domain knowledge, generating technically coherent scenarios with appropriate procedural details and system responses, and produced a structured analysis covering cognitive functions and performance influencing factors. However, expert review identified critical methodological gaps including conflation of concepts from different HRA methods, omission of formal task analysis steps required by NUREG-2256, and inadequate human failure events identification. While current LLMs show promise as auxiliary tools for scenario generation and preliminary analysis, they require significant enhancement before supporting safety-critical HRA applications. Future work should focus on method-specific training, integration with structured knowledge representations (e.g. knowledge graphs), and development of validation protocols to ensure appropriate application boundaries.

Keywords: Large language models, Human reliability assessment, Nuclear power operations, Knowledge graphs, IDHEAS-ECA

INTRODUCTION

The integration of generative artificial intelligence such as Large Language Models (LLMs) with safety analysis methodologies represents an emerging frontier in risk assessment, though direct applications to Human Reliability Assessment (HRA) remain limited in current literature.

Current research shows some progress in integrating LLMs with established safety analysis frameworks. The co-hazard analysis (CoHA) approach demonstrates that LLMs can effectively support hazard cause elicitation through context-aware interactions with human analysts (Diemert & Weber, 2023). Similarly, ChatGPT's integration with Failure Mode and Effects Analysis (FMEA) has shown significant potential for automating conventional risk assessment processes, generating failure modes that human analysts might overlook while accelerating analysis timelines (Xu, 2025). Multi-agent LLM systems employing Retrieval Augmented Generation

(RAG) for risk analysis have been developed to interface directly with FMEA knowledge databases, generating targeted recommendations for completing FMEA spreadsheets (Xia et al., 2024). These developments extend to Systems-Theoretic Process Analysis (STPA), where LLM-integrated frameworks have been shown to support the synthesis of unsafe control actions (UCAs) and loss scenarios (Raeisdanaei et al., 2025).

However, a critical gap emerges when examining direct applications to traditional HRA methods. While research demonstrates that machine learning systems can be trained to automatically classify accident reports involving human error, thereby expanding existing databases (Morais et al., 2022), the specific application of generative LLMs to produce structured HRA analyses remains largely unexplored.

Several critical challenges emerge from this literature for LLM-generated HRA analyses. First, the inherent uncertainty in human reliability analyses may be compounded by LLM hallucination tendencies, potentially generating plausible but incorrect Human Error Probabilities (HEPs) or Performance Shaping Factor (PSF) assessments. Research examining LLM safety through verification and validation techniques has categorized vulnerabilities into inherent issues, attacks, and unintended bugs (Huang et al., 2024), highlighting the complexity of ensuring reliable outputs in safety-critical applications. Second, many traditional HRA methods rely on structured decision trees and quantitative multipliers that require precise application—areas where LLM reliability remains questionable. Third, the regulatory and certification requirements for HRA in safety-critical industries demand traceable, repeatable analyses that may conflict with the probabilistic nature of LLM outputs.

The literature suggests that while LLMs show promise for supporting certain aspects of safety analysis, their direct application to generating complete HRA analyses represents a significant leap requiring careful validation. Research on AI's role in detecting and mitigating human errors across safety-critical industries (Gursel et al., 2025) provides a comprehensive overview but focuses primarily on detection rather than reliability assessment. The maturity of LLM integration with structured safety methods like FMEA and STPA provides a foundation, but the analysis precision and regulatory requirements of HRA methods present unique challenges not yet addressed in current research.

The IDHEAS-ECA methodology was developed by the U.S. Nuclear Regulatory Commission and formally introduced around 2020 to improve consistency and cognitive grounding in HRA. Led by researchers from the Office of Regulatory Research, it addresses limitations in earlier methods by integrating cognitive science and operational data. IDHEAS-ECA is especially suited for modeling operator actions under severe conditions, including scenarios outside control rooms and during low-power or shutdown states (Xing et al., 2022). Its selection here is based on its empirical foundation, structured approach, and well-developed qualitative analysis framework.

IDHEAS-ECA stands out by modeling human performance through five macrocognitive functions—Detection, Understanding, Decision-Making, Action Execution, and Interteam Coordination—each linked to specific

Cognitive Failure Modes (CFMs). Analysts using this method typically assess scenario context, define Human Failure Events (HFEs), and apply 20 Performance Influencing Factors (PIFs) to evaluate error likelihood. The method's stepwise process, supported by standardized worksheets and data sources like Scenario Authoring, Characterization and Debriefing Application (SACADA), ensures traceable and consistent HRA outcomes.

This paper aims to practically examine the idea of supporting the HRA analyst become more efficient by moving from the conventional structured *checklist*-type analysis to leveraging the potentials of generative AIs like LLMs. The following sections of the paper outline the methodology employed, presents the results, discusses the findings, implications, limitation of the study, and potential solutions. We also outline potential future directions for this research and provide a conclusion.

METHOD

This study employed a multi-stage approach to evaluate LLM capabilities for Human Reliability Assessment applications, using Claude Opus 4.1 for all generation tasks. The method consisted of scenario generation, HRA analysis generation, expert review, and synthesis of findings.

The first stage employed an LLM for scenario generation to assess multiple capabilities simultaneously. Scenario generation requires understanding of nuclear plant systems, operational procedures, accident phenomenology, and human-system interactions, allowing evaluation of the model's domain knowledge across these interconnected areas. This approach tested whether LLMs can generate technically coherent accident progressions suitable for HRA analysis while providing insights into their ability to structure complex technical information.

Critically, all scenario information came entirely from within the model's training data—no plant-specific information, procedures, or human performance data were provided as input. This constraint tested the model's intrinsic nuclear operations knowledge, revealing it must have ingested detailed emergency operating procedures, system descriptions, and operational practices during training. The model likely synthesized knowledge from multiple plants, regulatory documents, and technical publications, though this aggregation could create inconsistencies where elements from different plants or vendors are combined.

The prompt instructed the model to assume a senior reactor operator role at a Westinghouse-type Pressurized Water Reactor (PWR) plant and generate two Steam Generator Tube Rupture (SGTR) scenarios—baseline and complex—in JSON format (JavaScript Object Notation). JSON was selected for its unambiguous field delineation, explicit relationships between data elements, and consistent syntax that reduces interpretation ambiguity in LLM applications. The baseline scenario represented a straightforward single tube rupture, while the complex scenario introduced multiple complications including equipment failures, degraded conditions, and reduced crew staffing.

The generated scenarios were subsequently provided to the same LLM with instructions to perform an HRA using the IDHEAS-ECA method. The model

produced a comprehensive qualitative HRA analysis including human failure event identification, cognitive function analysis for each event, performance influencing factor assessment, and comparative analysis between the baseline and complex scenarios.

The generated HRA analysis was forwarded to two method experts with instructions to review the analysis quality, methodological accuracy, and technical coherence. Both experts provided detailed comments directly annotated in the document, identifying strengths and limitations in the LLM's application of the method.

The expert reviews were analyzed to identify common themes, including areas where the LLM demonstrated strong understanding versus those requiring improvement.

RESULTS

LLM-Generated Scenario

The LLM produced seven main sections: SGTR overview, plant-specific information, baseline scenario, complex scenario, procedural appendices, human factor analysis information, and timing windows. The baseline scenario depicted a single tube rupture with straightforward diagnosis, while the complex scenario introduced multiple complications including stuck-open atmospheric dump valve (ADV), loss of instrument air, failed pressurizer pilot-operated relieve valve (PORV), and loss of offsite power.

Generated procedures followed standard Westinghouse Owners Group formats:


```
"critical_steps": [{
    "step_number": "E-0 Step 17",
    "full_text": "Check If SG Tubes Are Intact: Check SG levels - NOT INCREASING
IN AN UNCONTROLLED MANNER, Check secondary radiation - NORMAL, Air ejector
radiation - NORMAL..."
    }]
```


This demonstrates understanding of SGTR diagnostic methodology, including multiple radiation monitoring points and transition logic. The scenario progression incorporated realistic complications:


```
"time": "T+6 min",
"plant_response": "SG-A ADV cycles open but fails to close when pressure
decreases",
"complications": "ADV-A stuck 75% open, continuous steam release to atmosphere",
"indications": "ADV-A position indication 75%, acoustic monitor confirms flow"
```


This accurately portrays a stuck-open ADV with appropriate position indication and acoustic confirmation—realistic control room indications for diagnosing valve position.

The model demonstrated sophisticated understanding of integrated plant response. Time windows distinguished between predicted performance and requirements:

```
"prevent_sg_overfill": {
    "time_available": "60-90 minutes depending on break size (460 gpm = ~75 min
to 100% NR)",
    "consequence_of_failure": "Water relief through safety valves, potential
containment bypass"
}
```


This appropriately scales time with leak rate and identifies critical safety concerns of water relief through safety valves.

However, the model's aggregated training data introduces potential inconsistencies. The SI actuation setpoint (1,865 psig) and main steam isolation setpoint (585 psig) may not correspond to the same plant design. Emergency core cooling systems (ECCS) configuration and auxiliary feedwater capacities might reflect composite designs rather than coherent configurations. While not invalidating scenarios for training purposes, this highlights the importance of plant-specific validation for actual safety analyses.

LLM-Generated IDHEAS-ECA Analysis

The LLM generated an HRA analysis structured in six sections:

1) Scope and objectives
2) Human failure event identification
3) Scenario 1 – baseline SGTR Analysis
4) Scenario 2 – complex SGTR Analysis
5) Comparative analysis of baseline versus complex scenarios
6) Conclusions and HRA recommendations

Six HFEs were identified (presented by the LLM as shown on Table 1) for both scenarios and analyzed across five cognitive functions: Detection, Understanding, Decision-Making, Action Execution, and Teamwork. The analysis identified applicable Cognitive Failure Modes (CFMs) and PIFs for each scenario.

Table 1: HFEs identified by the LLM.

HFE ID	Description	Success Criteria	Time Window	Procedure
HFE-1	Diagnose SGTR and identify ruptured SG	Correctly identify SGTR and affected SG(s)	To to To + 15 min	E-0 Steps 11-18
HFE-2	Isolate ruptured SG	Complete isolation of steam and feed flow	To + 10 to To + 25 min	E-3 Steps 1-4
HFE-3	Perform RCS cooldown	Cool RCS to target temperature	To + 15 to To + 45 min	E-3 Steps 11-14
HFE-4	Depressurize RCS to terminate break flow	Reduce RCS pressure below ruptured SG pressure	To + 25 to To + 50 min	E-3 Steps 15-18
HFE-5	Prevent SG overfill	Maintain ruptured SG level <88% NR	Continuous	E-3 Steps 26-27
HFE-6	Terminate SI when criteria met	Stop ECCS injection at appropriate conditions	To + 30 to To + 60 min	E-3 Steps 19-23

Methos Expert's Review of the LLM-Generated Analysis

This section presents the HRA experts' key findings with specific examples from the LLM-generated analysis.

Missing Process Steps. The analysis omitted critical IDHEAS-ECA steps including scenario analysis with operational narrative and context analysis, timeline establishment, and task analysis before CFM identification. For example, the LLM began directly with HFE definitions without providing the required operational narrative:

The experts noted: "Missing IDHEAS-ECA 3.1 Step 1 Scenario Analysis... Missing 3.1.1 Develop the Operational Narrative... The HFEs are not incorrect, but they are not at the level of HFEs defined in IDHEAS."

HFE Definition Issues. The identified HFEs were too granular for IDHEAS-ECA methodology. The LLM separated actions that should be modeled together:

```
HFE-3: Failure to perform adequate RCS cooldown
HFE-4: Failure to depressurize RCS to terminate break flow
HFE-5: Failure to prevent SG overfill
HFE-6: Failure to terminate SI
```

Expert comment: "HFE-3 and 4 are commonly modeled as an HFE because they are iterative actions and one affects the other. HFE-5 and 6 are embedded in HFE-3 & 4."

Cue Identification Problems. While the model identified technically accurate cues, it included pre-trip indications that wouldn't be credited in actual PRA analysis:

```
PIF Category Status                                    Impact
Procedures   Adequate - Clear diagnostic steps in E-0  Positive
Experience   Mixed - Crew familiar with procedures     Neutral
Complexity   Low - Single tube rupture, clear symptoms Positive
```

Expert comment: "Even though the detection occurs before reactor trip, PRA analysis typically only credits the cues explicitly identified in the procedures (E0 here)."

Documentation Deficiencies. The analysis did not follow IDHEAS-ECA's structured worksheet format. When identifying PIFs, the LLM used non-standard categories:

```
PIF Category Status                                    Impact
Procedures   Adequate - Clear diagnostic steps in E-0  Positive
Experience   Mixed - Crew familiar with procedures     Neutral
Complexity   Low - Single tube rupture, clear symptoms Positive
```

Expert comment: "IDHEAS-ECA PIF structure begins with 'no impact' state then becomes 'negative' if certain PIF attribute is assessed as 'present.' PIFs have no 'positive' state."

CFM Identification. The analysis failed to explicitly document when CFMs were not applicable. For HFE-6 in the baseline scenario, the LLM evaluated only three cognitive functions without explaining why *Action Execution* and *Interteam Coordination* were omitted:

```
3.6 HFE-6: Terminate SI (Baseline)
DETECTION
        - Cues: Subcooling, pressurizer level, RCS pressure trend
UNDERSTANDING
        - Cognitive Demand: Verify all termination criteria met
DECISION-MAKING
        - Decision: Confirm SI termination appropriate
```

Expert comment: "Action CFM is missing here. The error of understanding and decision-making is negligible."

Method Contamination. The analysis conflated terminology from different HRA methods. Most notably, it consistently used "Teamwork" instead of the correct IDHEAS-ECA term:

```
TEAMWORK
Requirements: SRO/RO/BOP coordination for diagnosis
CFMs Applicable:
CFM T1: Miscommunication (low - standard terminology)
```

Expert comment: "Teamwork should be replaced with interteam coordination... Responding to a SGTR event typically does not need interteam coordination."

CFM Terminology: The LLM used incorrect CFM nomenclature throughout:

```
CFMs Applicable:
CFM D1: Failure to detect visual cue (low probability - multiple redundant
indications)
CFM E1: Omission of step (low - critical step with verification)
```

Expert comment: "IDHEAS-ECA has five CFMs... There is no 'CFM D1'. In IDHEAS, the symbols D1-D5 (and U1, U2...) are for processors that together achieve the macrocognitive function."

Unsupported Conclusions: The analysis provided insights without completing necessary quantitative steps:

```
CFMs Applicable:
CFM D1: Failure to detect visual cue (low probability - multiple redundant
indications)
CFM E1: Omission of step (low - critical step with verification)
```

Expert comment: "What is the definition of 'dominant failure modes?' If they mean the CFMs that dominate the HEPs... the PIF attribute assessment and HEP calculation haven't been performed yet."

DISCUSSIONS

The LLM demonstrated good domain knowledge, generating technically coherent SGTR scenarios with realistic plant responses and procedural details. However, the expert review revealed methodological issues that limit its current utility for formal HRA applications.

The model's ability to synthesize complex technical information suggests strong potential as an auxiliary tool for scenario generation and preliminary analysis. It correctly identified critical human actions and understood the relationships between plant systems, procedures, and operator responses. This capability could support HRA analysts in the initial stages of analysis, particularly for training scenarios or preliminary hazard identification.

However, the systematic methodological errors indicate that current LLMs lack the structured reasoning necessary for formal safety analysis. The conflation of different HRA methods suggests the model's training data included multiple methodologies without clear delineation, leading to contaminated outputs. The omission of required analytical steps and improper use of terminology would make such analyses unsuitable for regulatory submissions or safety-critical applications.

The experts' emphasis on following structured processes highlights a fundamental challenge: HRA methods require strict adherence to sequential analytical steps where each output becomes input for subsequent analyses. Current LLMs, trained on diverse textual data, struggle to maintain this methodological rigor without explicit guidance.

PROPOSED SOLUTIONS

Integration With Knowledge Graphs

Knowledge graphs offer a promising solution to address the methodological rigor required for HRA applications. A knowledge graph for HRA would represent methods, procedures, plant systems, and their relationships as structured, interconnected data rather than unstructured text. This approach would provide several advantages:

For an HRA application, the scope of a knowledge graph would encompass:

- **Method ontologies**: Formal representations of HRA methods (IDHEAS-ECA, SPAR-H, THERP) with their specific steps, terminology, and requirements.
- **Plant system models**: Structured representations of system configurations, dependencies, and failure modes.
- **Procedural networks**: Emergency operating procedures linked to plant states and operator actions.
- **Human performance data**: Structured repositories of performance shaping factors and their quantitative relationships.

The primary challenge lies in the initial knowledge engineering effort required to construct comprehensive graphs. Once established, these graphs would be highly scalable, allowing addition of new plants, procedures, or methods through structured templates. Knowledge graphs enhance LLM capabilities by providing:

- **Contextual grounding**: LLMs can query structured data to verify terminology and methodological requirements.
- **Consistency enforcement**: Graph constraints ensure outputs comply with method-specific requirements.
- **Traceability**: All analytical decisions can be traced to specific nodes and relationships in the graph.
- **Validation support**: Outputs can be automatically checked against graph-encoded rules and constraints.

However, this approach also presents challenges, including development cost (creating comprehensive knowledge graphs requires significant expert input), maintenance effort (graphs must be updated as methods and procedures evolve), and integration complexity (effective LLM-knowledge graph interfaces require sophisticated query and reasoning mechanisms).

Method-Specific Fine-Tuning

Beyond knowledge graphs, LLMs could be fine-tuned on curated datasets of properly executed HRA analyses. This would involve creating training sets that explicitly demonstrate correct methodological application, proper terminology usage, and appropriate documentation standards. Combined with structured prompting techniques that enforce step-by-step analysis, this could significantly improve methodological adherence.

Hybrid Human-AI Systems

Rather than fully automated analysis prior to human review, hybrid systems where LLMs support specific subtasks under human supervision may be more appropriate for near-term applications. For example, LLMs could generate initial scenario descriptions that analysts refine, or provide real-time documentation assistance while analysts perform the core reliability assessment.

Future Research Directions

This study highlights several areas requiring further investigation:

1. Development of standardized evaluation frameworks for LLM-generated safety analyses.
2. Creation of benchmark datasets for HRA method training and validation.
3. Investigation of prompt engineering techniques specific to structured analysis methods.
4. Use of LLM reasoning capabilities for enhanced logical consistency.
5. Development of explanation mechanisms that make LLM reasoning transparent and auditable for safety applications.

The evolution of AI-assisted HRA will likely proceed incrementally, with initial applications in low-consequence training and preliminary analysis contexts, gradually expanding as validation methods and regulatory frameworks mature. Success will require close collaboration between AI researchers, HRA practitioners, and regulatory bodies to ensure that technological capabilities align with safety requirements and methodological standards.

CONCLUSION

This study provides empirical evidence of both the benefits and limitations of current LLMs for supporting Human Reliability Analysis. While Claude Opus 4.1 demonstrated substantial nuclear domain knowledge and generated plausible scenario descriptions, it failed to correctly implement the IDHEAS-ECA methodology, producing analyses that would be unsuitable for safety-critical applications.

The findings suggest that LLMs could serve as valuable auxiliary tools for HRA analysts, particularly in scenario generation, initial brainstorming, and documentation assistance. However, significant development is needed before these systems can reliably support formal HRA work. The path forward requires addressing both technical and methodological challenges through structured approaches.

ACKNOWLEDGMENT

The authors would like to thank J. Xing and J.Y. Chang for their expert review of the LLM-generated analysis. All conclusions in this paper reflect purely the views of the authors, not the reviewers or their agencies.

REFERENCES

Diemert, S., & Weber, J. H. (2023). Can large language models assist in hazard analysis? arXiv preprint arXiv:2303.15473. https://doi.org/10.48550/arXiv.2303.15473

Gursel, E., Madadi, M., Coble, J. B., Agarwal, V., Yadav, V., Boring, R. L., Khojandi, A. (2025). The role of AI in detecting and mitigating human errors in safety-critical industries: A review, Reliability Engineering & System Safety, Volume 256, 2025, 110682. https://doi.org/10.1016/j.ress.2024.110682

Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., Bensalem, S., Mu, R., Qi, Y., Zhao, X., Cai, K., Zhang, Y., Wu, S., Xu, P., Wu, D., Freitas, A., & Mustafa, M. A. (2024). A survey of safety and trustworthiness of large language models through the lens of verification and validation. Artificial Intelligence Review, 57(7), 1-69. https://doi.org/10.1007/s10462-024-10824-0

Morais, C., Yung, K. L., Johnson, K., Moura, R., Beer, M., Patelli, E. (2022). Identification of human errors and influencing factors: A machine learning approach, Safety Science, Volume 146, 2022, 105528. https://doi.org/10.1016/j.ssci.2021.105528

Raeisdanaei, A., Kim, J., Liao, M., Kochhar, S. (2025). An LLM-Integrated Framework for Completion, Management, and Tracing of STPA, *Preprint*. https://doi.org/10.48550/arXiv.2503.12043

Xia, Y., N. Jazdi, N., & Weyrich, M. (2024). Enhance FMEA with Large Language Models for Assisted Risk Management in Technical Processes and Products. 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA), Padova, Italy, 2024, pp. 1–4. https://doi.org/10.1109/ETFA61755.2024.10710996

Xing, J., Chang, Y. J., and DeJesus Segarra, J. (2022). Integrated Human Event Analysis System for Event and Condition Assessment (IDHEAS-ECA). U.S. Nuclear Regulatory Commission, Office of Nuclear Regulation, NUREG 2256.

Xu, M. (2025). Enhancing FMEA with ChatGPT: Structured outputs, qualitative evaluations, and AI-human hybrid FMEA. IEEE Conference: 2025 Annual Reliability and Maintainability Symposium (RAMS), 27–30 Jan 2025, Destin, FL, USA.