Transactions of the Korean Nuclear Society Autumn Meeting
Gyeongju, Korea, October 29-30, 2026

Stage-wise Evaluation of an On-Premises LLM Pipeline for INES Event Rating

MinJe Shin, YooJoon Seoung Seung Jun Lee*
Department of Nuclear Engineering
Ulsan National Institute of Science and Technology (UNIST)
50 UNIST-gil, Ulju-gun, Ulsan, 44949, Republic of Korea
**Corresponding author: sjlee420@unist.ac.kr*

Keywords : INES, event rating, large language model, regulatory decision support, rating path verification

1. Introduction

Large language models (LLMs) are increasingly being applied to domains such as law, healthcare, and regulatory compliance, where large volumes of unstructured documents must be reviewed and contextually relevant judgments must be derived. In the nuclear regulatory domain, applications including regulatory document question answering, ontology-based requirement extraction, and automation of review procedures have also been reported [1, 2]. In particular, small language models offer practical advantages in nuclear facilities where operation within a closed network is required, as they can process documents without transmitting them outside the facility.

For nuclear power plants, the INES (International Nuclear Event Scale) level of an accident or failure is deliberated by the Nuclear Power Plant Event Rating Committee on the basis of an investigation report [3, 4]. The determination follows a multi-stage procedure in which three axes are evaluated separately, the maximum value among them is selected, and additional factors are subsequently applied. At each stage, the decision is supported by specific provisions in the relevant manual and appendices to regulatory notices. Within the scope examined in this study, both of the first two axes were determined to be zero, and the final level was therefore determined by the defence-in-depth axis. Determination along this axis involves multiple decision branches concerning whether an initiating event actually occurred, its frequency of occurrence, the operability of safety functions, and other factors, with extensive cross-references among provisions. Accordingly, the substantive task of the deliberation is not simply to determine a numerical rating, but rather to identify which statements in the investigation report correspond to which regulatory provisions and to document those connections. This process is document-intensive, although the effort it requires depends on how investigation and review roles are divided within each institution (Section 4.4).

Automation is an obvious way to reduce this burden, but evaluating such a system by accuracy alone is difficult for two reasons. First, most domestic nuclear events are assigned Level 0. Consequently, a system that simply assigns Level 0 to every event could still achieve high accuracy. High accuracy alone therefore does not demonstrate that the tool has contributed to the actual determination process. Second, if the source of a given value cannot be traced back to the specific regulatory provision and statement from which it was derived, reviewers cannot verify the value and must re-examine the case from the beginning, potentially making the tool an additional burden rather than a support mechanism. What is required in regulatory determination is not merely the correct numerical rating, but a record showing which regulatory provision and which statement in the report support each individual value. The manual itself reflects this characteristic by distinguishing between the absence of evidence and evidence of absence; unless there is an affirmative statement that a safety measure operated as designed, the corresponding decision gate cannot be opened. Existing studies have primarily evaluated constraint compliance or classification accuracy as performance metrics, but have not addressed a mechanism for withholding an output when sufficient evidence is unavailable.

In this paper, we propose a pipeline designed to assist, rather than replace, the deliberation process and evaluate its operational characteristics. The determination process is decomposed into individual slots, and only the regulatory provisions relevant to each slot are provided to the model. Once a slot value is determined, the system records both the corresponding provision identifier and the supporting evidence sentence; if the value cannot be determined with sufficient confidence, no value is produced and the case is instead passed to a human reviewer. Table lookups, maximum-value selection, and upper-bound handling are performed programmatically, while the language model is responsible for determining only a single slot value, thereby ensuring reproducibility for identical inputs. The proposed system employs an on-premises small language model, Qwen3-4B. Evaluation was conducted by comparing the decision path step by step, rather than by evaluating only the final INES level. Because different decision paths can lead to the same final level, the final rating alone is insufficient to verify the correctness of the underlying decision process.

2. Methodology

This section describes the pipeline architecture, reference-answer construction, and comparison procedure. The overall architecture is shown in Fig. 1.

2.1 Slot Decomposition and Provision Injection

The determination procedure was decomposed into individual slots, including out-of-scale screening, axis-specific gates, initiating-event occurrence, degraded safety functions, initiator frequency, safety function operability, potential consequence range, remaining safety layers, adjustment factors, and the impact and barrier axes.

The INES Manual and Appendix 2 of the relevant regulatory notice were divided into eight rule documents, and only the provisions required for each slot were injected. This fixed the applicable provisions at inference time without searching the full manual. Each provision was assigned a source-and-section identifier, such as IL§2.1.

The model outputs a slot value and the identifier of the cited provision. Citations are limited to provisions in the rule documents, not sentences in the investigation report. Since valid identifiers form a finite set, the code verifies each citation and records invalid ones. If no valid citation is available, predefined categorical slots such as initiator frequency and safety function operability are left unknown. For path-determining slots such as initiating-event occurrence, the value is retained but marked as unsupported.

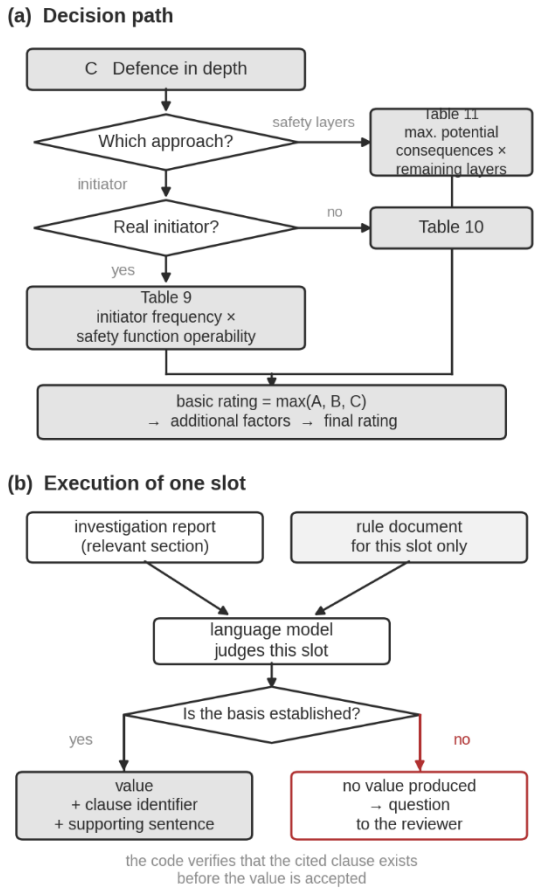


Fig. 1. (a) The decision path on the defence-in-depth axis. Shaded boxes are executed by code; diamonds are judged by the model, one slot at a time. (b) Execution of a single slot. When the basis is not established, no value is produced.

2.2 Division of Responsibilities

The language model determines one slot value at a time, while table lookup, maximum selection across axes, and upper-bound handling are performed by code. Deterministic values, such as whether an initiating event appears in the normative list in Appendix 2, are also resolved programmatically.

An upward adjustment factor is applied only when both its provision citation and the corresponding condition in the investigation report are verified; otherwise, the adjustment is suppressed and flagged.

Each slot is processed independently without shared conversation history. Only required values from previous stages are passed to subsequent stages. For each slot, the system records the value, provision identifier, supporting sentence, and source. Unresolved slots are retained for human review rather than assigned a value.

The model is an on-premises Qwen3-4B with a fixed temperature and a fixed random seed.

2.3 Reference Answers

Reference answers were taken from 435 deliberation cases resolved by the Nuclear Power Plant Event Rating Committee. Because these finalized documents contain both the determination and its rationale, they were used as stage-level reference answers without additional labeling.

The cases vary in format, with some containing rating summary tables and others consisting of narrative text. Tables may also span page boundaries. Table rows and narrative paragraphs were therefore extracted independently, and conflicting values were left unfilled.

The decision path—initiating-event or safety-layer approach—was identified only from explicit statements in the case body. Frequency and operability fields in the summary table were not used for this purpose because they are conventionally populated even in cases ultimately resolved through the safety-layer approach.

2.4 Comparison

Comparison required both an investigation report and a corresponding deliberation case for the same event. Investigation reports were linked to 433 of the 440 deliberation cases.

Of these, 162 follow a 1990s summary format whose optical character recognition quality is poor and which lacks the structure required for slot-based extraction. Excluding them leaves 271 cases for comparison.

The comparison focused on the decision path rather than only the final rating, since different paths can produce the same final level. Stage-level agreement was calculated only when both the pipeline and reference

answer produced a value. Unresolved pipeline outputs and unavailable reference values were excluded from the denominator. The modal reference value at each stage was also reported as a baseline.

3. Results

Both Axis A and Axis B were assigned a value of 0 by both the reference answers and the pipeline in all but one of the comparison cases. The final rating was therefore determined by the defence-in-depth axis, which is examined below.

3.1 Stage-by-Stage Comparison

The stage-level agreement rates for the 271 cases are shown in Table I. Agreement with the committee determination was 77.5% for approach selection, 97.1% for the initiating-event slot, 98.6% for initiator frequency, 88.5% for safety function operability, 91.8% for the basic rating, 78.1% for additional upward adjustment factors, and 72.8% for the final rating.

Table I also reports the agreement obtained by always predicting the modal reference value. At all eight stages, this baseline was equal to or higher than the pipeline. This was most pronounced for safety function operability, where 98.4% of the reference answers were A (fully satisfied). Thus, high agreement rates largely reflect concentrated reference distributions rather than decision difficulty.

The reference-value distribution is shown in Table II. Of the 12 cells in Table 9, three contained events, with one cell accounting for 98.0%. The modal-value proportion remained similar for frequency × operability (98.0%) and the basic rating (98.0%), but decreased to 86.7% for the final rating. The main increase in value dispersion therefore occurred at the additional upward adjustment stage.

Table I: Stage-wise agreement with committee decisions (n = 271)

Stage (n)	**Agr. (%)**	**Modal (%)**
Approach: initiator / safety layers (271)	77.5	77.9
Real initiator, slot (170)	97.1	97.1
Table 9 vs 10 selection (174)	84.5	96.6
Initiator frequency (213)	98.6	99.1
Safety function operability (217)	88.5	98.6
Basic rating (147)	91.8	98.0
Additional factors (270)	78.1	86.3
Final rating (250)	72.8	89.2

Table II: Distribution of reference values

Axis	**Modal value**	**Share**	**n**
Initiator frequency	Expected	99.2 %	250
Safety function operability	A (Full)	98.4 %	250
Table 9 cell	(Expected, A)	98.0 %	249
Basic rating	0	98.0 %	147
Final rating	0	86.7 %	271

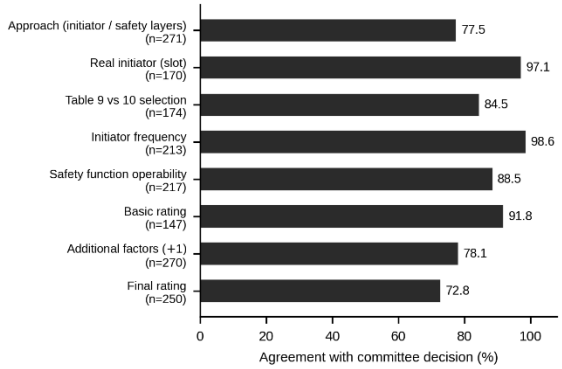


Fig. 2. Agreement by stage (n = 271). Read against Table II: the modal value accounts for most of the ground truth at every stage up to the basic rating.

3.2 Divergence in Path Selection

Separating the initiating-event determination from the subsequent Table 9/Table 10 selection revealed a discrepancy (Fig. 3). While the initiating-event slot agreed in 97.1% of cases, table selection agreed in only 84.5%. Of the 27 cases with incorrect table selection, 22 had a correct initiating-event determination.

The discrepancy arose from how an unresolved occurrence status was represented. Cases in which occurrence could not be established were handled in the same way as non-occurrence and routed to Table 10. Since all A and B rows in Table 10 yield 0, initiator frequency did not affect the result in these cases. In 20 cases, the initiating event was identified by name while its occurrence remained unresolved, leaving the two fields in contradiction.

These cases were not detectable from final-rating agreement because their final ratings still matched the committee's results. The discrepancy became visible only when slot determination and path routing were evaluated separately.

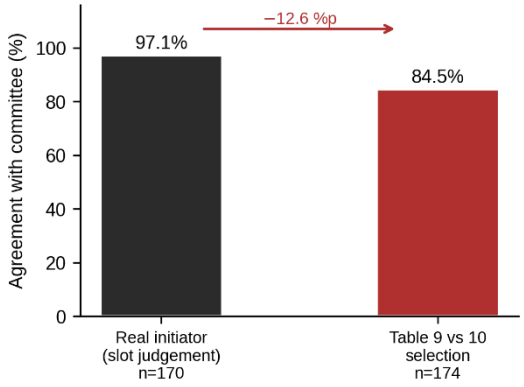


Fig. 3. The real-initiator slot and the table selection that follows from it. Most table-selection mismatches are cases where the slot itself agreed with the ground truth.

3.3 Safety-Layer Slot

Determination was withheld in 21 of the 271 cases. Unknown slots included safety layers (16), the barrier axis (13), the impact axis (7), and initiator frequency (5); a case could contain multiple unknown slots.

The safety-layer slot was particularly difficult because the number of effective layers was generally not stated explicitly in the investigation report. Determining it could require combining descriptions across multiple sections and considering common-cause failures. For example, two physically redundant radiation monitors were considered one effective layer because both failed due to the same design-change error.

Among the 44 cases routed to Table 11, the safety-layer slot was set to 0 without supporting evidence in 22 cases, left unresolved in 16, and assigned a row value in 7. Separately, in 16 cases the maximum potential consequence column was derived from the reactor state rather than from the event itself.

Termination at 0 without evidence is inconsistent with Section 6.2.3.1, which permits termination at 0 only under specified conditions and directs all other cases to Table 11. Withholding, by contrast, reflects the intended evidence-based policy. Deriving the consequence column from the reactor state also bypasses the exception in the manual that allows Level 0 when sufficient response time remains despite a single residual layer.

The committee reached Table 11 far less often than the pipeline. Among 14 cases handled by the committee using the safety-layer approach, only 4 included entries for both Table 11 axes; the others were terminated through Section 6.2.3.1(1), specific-type guidance in Section 6.3, or sufficient response time. In contrast, all 44 cases classified by the pipeline as using the safety-layer approach entered Table 11.

Finally, 30 cases involved an identified upward adjustment factor that was not applied because the corresponding condition was not supported by the investigation report. The suppression and its reason were recorded as flags.

4. Discussion

4.1 What Accuracy Measures

At all eight stages, always predicting the modal value achieved an agreement rate equal to or higher than that of the pipeline. This reflects the concentration of reference values rather than necessarily poor model performance. For example, 98.4% of the reference answers for safety function operability were A, leaving little room for improvement over the modal prediction.

Therefore, agreement rate alone does not adequately represent decision-making capability. High agreement was observed mainly at stages with highly concentrated values, whereas the stages that actually affected the final rating showed greater variation and lower agreement. For a regulatory decision-support tool, which decisions were reached correctly is more informative than the overall agreement rate.

4.2 Regulatory Compliance vs. Actual Practice

The pipeline and the committee differed in their handling of the safety-layer approach. Although the manual directs the use of Table 11 based on the maximum potential consequence and remaining safety layers, the committee often terminates cases through other gates or specific-type guidelines. The pipeline, which directly implements the written rules, therefore enters Table 11 more frequently.

This difference is not necessarily an implementation error. The manual itself contains branches whose interpretation depends on reactor operating mode, and the same status description may lead to different paths depending on the reactor type. Moreover, some committee decisions cannot be directly explained by the literal wording of the manual. Thus, a system designed for strict regulatory compliance and one designed to reproduce actual review practice may produce different results. The evaluation criterion should therefore be defined explicitly.

4.3 The Boundary of Automation

Two limitations of automation were observed. Some stages have nearly fixed values and provide little benefit from automated judgment, while others require expert judgment that is not explicitly stated in the investigation report. The number of remaining safety layers is one example.

When unsupported values are inferred, the direction of the error is unpredictable. In this study, the absence of evidence for safety-layer degradation was treated as 0 in some cases and as the worst case in others. Therefore, the role of automation should be to identify where evidence is insufficient and defer those cases to human reviewers rather than to fill unsupported values. In this sense, withholding a value can serve as a useful signal rather than a failure.

4.4 Applicability to Institutional Workflows

The value of a tool of this kind depends on the workflow into which it is placed. In current domestic practice the draft rating is prepared by the department that conducted the investigation and is then submitted to the committee; where that department already holds the investigation context, the benefit of automated tracing is correspondingly smaller. The benefit is larger where the reviewer is separated from the investigation, where consistency across many cases and reviewers is required, or where a draft must be prepared while the investigation is still in progress.

The last of these points to the setting for which the pipeline is ultimately intended. The cases examined here were rated from completed investigation reports in which the rationale had already been documented, so

the measurement asks whether a determination can be reconstructed from a finished document. A draft prepared shortly after an event works from far less. Under those conditions the useful output is not a rating but the set of points at which a rating cannot yet be established: which slot is unresolved, which provision governs it, and what statement would settle it. The withheld cases reported in Section 3.3 already take this form, each carrying the unresolved slot, the levels that remain possible, and the question that separates them. Evaluating the pipeline in that setting calls for measuring the questions it raises rather than its agreement with a completed determination, and this is left for future work.

4.5 Limitations

Path reference answers. The reference path was taken only from cases that state the selected approach explicitly, and such cases are a subset of the dataset. On those cases, a path inferred without the explicit statement agreed 87.5% of the time; comparable error may therefore remain in cases where no explicit statement exists.

Evidence sentences. Provision identifiers are verified against a finite set of valid identifiers, and extracted evidence sentences are guaranteed to exist in the source reports. However, evidence statements generated by the model are not automatically verified against the original text and therefore still require human review.

Reasoning mode and sample. The model was run with its reasoning mode disabled, a setting inherited from an earlier configuration in which slots performed deterministic extraction only; its suitability for the present rule-based setting was not measured. The analysis covers the 271 cases for which an investigation report was available and suitable for slot-based processing, and reflects a single model and a single execution.

ACKNOWLEDGMENTS

This work was supported by Korea Institute of Energy Technology Evaluation and Planning(KETEP) grant funded by the Korea government(MOTIE)(RS-2024-00403194, Next-Generation Nuclear Technology Creation IP-R&D Talent (Human Resources) Development Project)

REFERENCES

[1] I. S. Hwang, Y. P. Lee, Nuclear Compliance: Assessing the Utility of LLMs with RAG, Trans. Korean Nucl. Soc. Spring Meeting, 2026.
[2] B. K. Kim, Y. G. Yu., Ontology-Driven Knowledge Graph Construction from Text Using Large Language Models: A Nuclear Domain Case Study, Trans. Korean Nucl. Soc. Spring Meeting, 2026.
[3] IAEA and OECD/NEA, INES: The International Nuclear and Radiological Event Scale User's Manual, 2008 Edition, IAEA, Vienna, 2013.
[4] Nuclear Safety and Security Commission, Notice on Reporting and Disclosure of Accidents and Incidents at Nuclear Facilities, Appendix 2.