Pre-Execution Dual-Gate Verification and Dynamic Procedure Reconfiguration for LLM Agent System in Nuclear Reactor Operation

Gayeon Kim[a,b], Seungjin Baek[a,b], Joowon Cha[a,c], Yonggyun Yu[a,c], Seung Geun Kim[a,*]
[a]*Applied Artificial Intelligence Section, Korea Atomic Energy Research Institute, Daejeon 34057, Republic of Korea*
[b]*Korea University of Technology and Education (KOREATECH), Cheonan 31253, Republic of Korea*
[c]*University of Science and Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Republic of Korea*
[*]*Corresponding author: sgkim92@kaeri.re.kr*

***Keywords :** Large Language Models, Retrieval-Augmented Generation, Adaptive Planning, Dynamic Reconfiguration

1. Introduction

Large Language Models (LLMs) have gained increasing attention as a core technology for autonomous agent systems capable of performing complex, multi-step tasks, owing to their strong plan generation and procedural reasoning capabilities. Efforts to leverage LLMs in domains requiring high reliability, such as the nuclear industry, have been ongoing, and prior studies have proposed LLM agent frameworks for reactor operation assistance in simulator environments. These works explored the potential of LLMs as autonomous decision-support tools, including real-time detection and proactive mitigation capabilities [1, 2].

Regulatory bodies have also begun responding to these technological developments. The U.S. Nuclear Regulatory Commission (NRC) has acknowledged the potential of AI to enhance nuclear safety and operational efficiency, and is revising and strengthening its regulatory framework to complement existing safety standards [3]. In particular, the NRC is integrating AI technologies within the existing digital instrumentation and control (I&C) regulatory structure, while establishing concrete guidelines through guidance documents and policy statements to build an institutional foundation for technology adoption [3].

Despite these promising developments, significant structural limitations and safety concerns remain in applying LLM-based systems to real plant environments. According to Lee et al. (2025), LLMs inherently exhibit hallucination phenomena and possess a probabilistic nature in their decision-making processes [4]. As a result, model-generated outputs may appear logically plausible while being factually inaccurate or ambiguous, making it difficult to guarantee complete predictability and reproducibility. In particular, current LLM agent systems lack rigorous formal verification methodologies, safety assessment frameworks, and the regulatory structures necessary to support them [4].

Therefore, deploying LLMs in safety-critical environments such as nuclear power plants requires a structure that maintains flexible reasoning capabilities while proactively intercepting potential errors before execution. In a system where errors in generated procedures—whether infeasible or in violation of safety policies—are only detected after the execution stage has been entered, additional risks may arise during post-execution recovery. Furthermore, discarding the entire plan and restarting introduces unnecessary computational cost and structural instability. From a practical system design perspective, what matters is not the plan generation capability itself, but rather the ability to systematically verify generated procedures prior to execution and, when issues are identified, to flexibly revise the affected steps rather than discarding the entire plan.

To address these challenges, this study proposes an architecture that introduces a Dual-gate Verification structure between plan generation and execution, along with an Adaptive Planner that performs Dynamic Procedure Reconfiguration upon verification failure. The proposed system independently conducts feasibility verification (Feasibility Gate) and safety verification (Safety Gate), and when either gate detects a failure, the Adaptive Planner generates alternative procedures reflecting the cause of failure and reconfigures the plan accordingly. This enables the system to form a closed-loop architecture of plan, verification, and reconfiguration, supporting reliable multi-step task execution in safety-critical environments.

2. Prior Work and Motivation

Prior studies proposed an LLM-based multi-agent system for nuclear reactor operation assistance in a simulator environment [2]. In this system, an anomaly monitoring module (Detection Server) continuously collects and evaluates key parameters such as reactor power, control-rod position, and RCS pressure and temperature. Upon detecting an abnormal state, the server immediately notifies the agent, which then summarizes the current issue and its evidence for the operator and proposes a remedial procedure. In addition to responding to anomaly alerts, the operator can interact with the agent at any time through natural language to issue commands directly, and the agent acts accordingly.

The system proposed in the prior study is implemented as a multi-agent architecture consisting of a Supervisor, Planner, Approval Router, and Worker, as illustrated in Figure 1. The Supervisor determines the

next role at each turn, while the Planner presents a plan to the operator comprising a problem summary and step-by-step actions without directly invoking any tools. Once the operator reviews and approves the plan, the Approval Router interprets the response and determines whether to proceed with execution, upon which the Worker calls the registered tools to issue control commands to the simulator. The study demonstrated the practical potential of LLM-based systems as nuclear reactor operation assistance tools in a simulator environment.

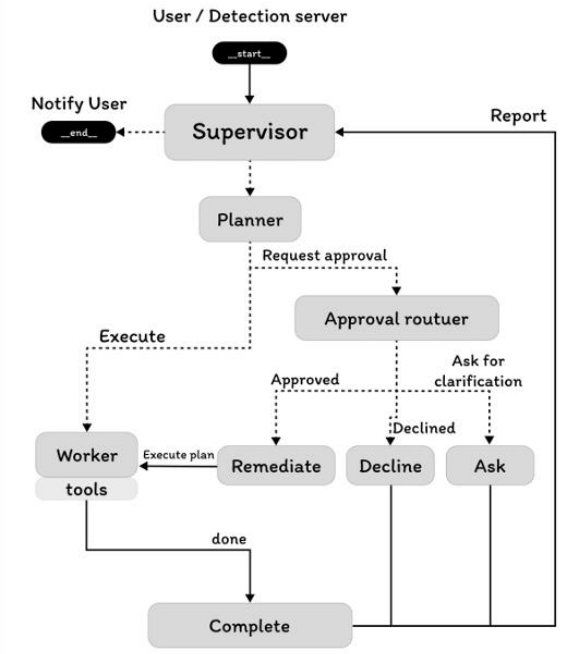

Fig. 1. Workflow and roles of the prior agent

However, the architecture of these prior studies does not include an explicit verification layer between plan generation and execution. In actual plant environments, equipment availability and safety policies shift in real time, and a plan that was valid when generated can become inapplicable by the time execution begins. In a structure where a fixed plan is executed sequentially, violations of feasibility or safety conditions may only be detected after the execution stage has been entered, which can introduce additional risk during post-execution recovery. To address this structural gap, the present study introduces a dual-gate verification structure prior to execution and an Adaptive Planner that performs Dynamic Procedure Reconfiguration, thereby constructing a more reliable system with a closed-loop architecture of plan, verification, and reconfiguration.

3. Methodology

The proposed system operates on a centralized orchestration structure, and is designed such that all generated procedures must pass through a dual-gate verification stage before execution. Each step derived during the initial planning phase is not executed immediately but is stored and managed in the system State, and subsequently passes through two sequential verification layers: the Feasibility Gate and the Safety Gate. Only steps that pass both gates are forwarded to the Worker/Tool Execution stage for actual task performance. The overall workflow and interactions between modules are illustrated in Figure 2. As shown in the figure, the system achieves high reliability suitable for industrial environments through the dual-gate verification structure that ensures normal execution flow, as well as the Adaptive Planner that forms a feedback loop to the Approval Router upon verification failure.

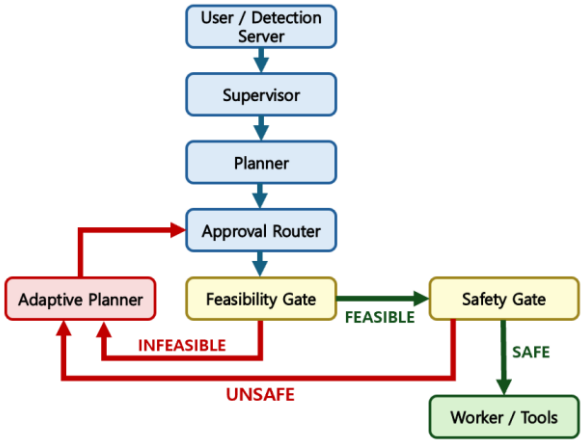

Fig. 2. Workflow of the proposed agent system featuring dual-gate verification and dynamic procedure reconfiguration

3.*1 Feasibility Gate*

The Feasibility Gate is the first verification stage, responsible for determining whether a generated procedure can be executed within the current technical and physical configuration of the system. This stage goes beyond simply checking for equipment faults; it comprehensively assesses structural consistency from the perspective of internal module flow and resource availability.

In terms of implementation, the feasibility gate module analyzes real-time plant state data (Plant State) to determine the executability of each step. For example, if a valve essential to a specific operation is in a non-operational or communication-failure state, the corresponding step is immediately determined to be infeasible. The module also pre-validates pipeline-level structural consistency—such as whether the retriever is properly loaded when a step requires RAG-based information retrieval, and whether the output data format of a preceding step matches the input requirements of the following step—thereby preventing runtime errors that could occur during execution. When an infeasibility determination is made, the failure reason is recorded in the State and control flow is immediately transferred to the Adaptive Planner.

3.2 Safety Gate

Procedures confirmed as feasible are then evaluated at the Safety Gate, the second verification stage, for compliance with plant safety policies. This stage is implemented as a hybrid structure combining deterministic rule-based inspection with probabilistic LLM-based evaluation, centered on the safety gate module.

The rule engine first rapidly detects violations of explicitly definable conditions based on predefined safety rules, such as exceedance of core temperature limits or violation of rod withdrawal constraints. Rule-based inspection is well-suited for efficiently filtering clear numerical violations due to its fast processing speed and deterministic outputs. The LLM evaluator then reasons over complex contextual factors that are difficult to capture through numerical rules alone, performing a final safety determination regarding the potential risk of the current action sequence considered as a whole. This enables complementary evaluation of ambiguous boundary conditions that are difficult for rule-based systems to handle.

When an Unsafe determination is made, the system generates a specific violation reason (unsafe reason) and records it in the State, rather than issuing a simple rejection message. This metadata is subsequently used as an essential constraint for the Adaptive Planner to derive an appropriate alternative procedure. Only steps that pass both gates are forwarded to the Worker for actual task execution.

3.3 Adaptive Planner: Dynamic Procedure Reconfiguration

When a verification failure occurs at either the Feasibility Gate or the Safety Gate, the system invokes the Adaptive Planner to perform Dynamic Procedure Reconfiguration. This process is not a retry mechanism that simply repeats the failed procedure, but an adaptive strategy that modifies the structure of the plan itself by reflecting the root cause of the failure.

The Adaptive Planner performs reconfiguration by comprehensively utilizing the failure reason metadata recorded by the verification gates, the current real-time plant state data, and standard procedures retrieved from vectorized operational manual PDFs via the RAG module. If no relevant manual data is available, an LLM Fallback mechanism is used to generate a safe and logically sound alternative step.

Reconfiguration operates by identifying the specific step where the failure occurred and either replacing that step with an alternative or inserting additional steps to satisfy prerequisite conditions, rather than discarding the entire plan. When a safety violation is the cause, the procedure is redirected to a lower-risk alternative while preserving the original operational objective. In cases where certain detailed operations cannot be automated through MCP tools, a manual execution procedure is additionally generated for the operator's reference, ensuring practical applicability in semi-automated operational environments. Once reconfiguration is complete, the revised procedure is returned to the Approval Router to re-enter the dual-gate verification process, and the system updates the UI and XML so that the operator can clearly recognize the changes made and the rationale for reconfiguration.

4. Case Study

Three scenarios were constructed to validate the behavior of the proposed system. Scenario A represents a case in which all verification steps are passed and the procedure is executed sequentially. Scenario B represents a case in which verification fails at the Feasibility Gate, and Scenario C represents a case in which verification fails at the Safety Gate. In each scenario, (I) denotes equipment operability, (II) denotes stability of plant operating conditions, (III) denotes the Feasibility Gate decision, (IV) denotes the Safety Gate decision, (V) denotes the system response, and (VI) denotes the reconfiguration method. The results are summarized in Table 1.

Table I: Scenario Comparison

	A	B	C
(I)	O	X	O
(II)	O	-	X
(III)	PASS	FAIL	PASS
(IV)	PASS	-	FAIL
(V)	Work/Tool Execution	Adaptive Planner invoked	Adaptive Planner invoked
(VI)	-	Step replaced with alternative equipment	Stabilization steps inserted

In Scenario A, the target equipment was assumed to be operable and the plant operating conditions were within a stable range. Both the Feasibility Gate and Safety Gate were passed, and the system executed each step sequentially in accordance with the defined procedure, confirming that the procedure matched the expected response.

In Scenario B, the target equipment was set to an inoperable state, and the corresponding step was blocked at the Feasibility Gate. The Adaptive Planner reconfigured the procedure to utilize alternative equipment capable of achieving the same operational objective while preserving the intent of the original procedure. The reconfigured procedure re-entered the dual-gate verification process before execution, confirming that goal-preserving reconfiguration was performed rather than simple retry.

In Scenario C, the equipment itself was operable, but the Safety Gate determined that executing the operation under the current unstable plant conditions could induce unintended transient responses. The Adaptive Planner performed state-aware reconfiguration by inserting preparatory stabilization steps prior to the main operation and adjusting control parameters

incrementally. The reconfigured procedure was required to re-pass the same dual-gate verification, ensuring that execution was permitted only when safety conditions were satisfied.

The comparison across the three scenarios confirms that the proposed system performs Dynamic Procedure Reconfiguration in a manner that structurally reflects the specific cause of verification failure, rather than applying a uniform retry strategy. By independently verifying feasibility and safety conditions prior to execution and performing state-aware reconfiguration accordingly, the proposed architecture demonstrates its suitability as a safety-focused design for industrial environments.

The verification structure proposed in this study carries significant engineering value in that it combines the deterministic reliability of rule-based systems with the flexible reasoning capability of LLM-based systems. This serves as an essential design mechanism for ensuring the practical applicability of LLMs in industrial environments where physical hazards are inherent. Furthermore, the proposed architecture is considered extensible to broader scenarios beyond the valve control case examined in this study, including emergency operating procedures involving multiple interconnected systems and multi-equipment recovery processes following abnormal events. Ongoing efforts are being made to expand these case studies to include more complex operational scenarios involving multi-component failures, thereby further demonstrating the system's robustness under diverse plant conditions.

5. Conclusion

This study proposed an architecture for LLM agent systems in safety-critical industrial environments such as nuclear power plants, built around a pre-execution dual-gate verification structure combined with Dynamic Procedure Reconfiguration. In the proposed system, the Feasibility Gate and Safety Gate independently assess each generated procedure before execution, and any verification failure activates the Adaptive Planner, which reconfigures the plan centered on the failed step by reflecting the cause of failure, forming a closed-loop architecture of plan, verification, and reconfiguration. The pre-execution dual-gate verification structure proactively ensures the safety and feasibility of procedures, while the integration of LLM-based reasoning and RAG-based manual retrieval enables flexible reconfiguration in situations where rigid predefined procedures alone would be insufficient, offering a structural foundation for the reliable deployment of agent systems in safety-critical industrial environments. Future research will prioritize the experimental validation of these expanded scenarios to ensure the scalable reliability of the proposed framework.

ACKNOWLEDGMENT

This research was supported by a grant from Korea Atomic Energy Research Institute (KAERI) (No. KAERI-526140-26)

REFERENCES

[1] Y. P. Lee and S. G. Kim and Y. Yu, "LLM-Based Integrated Control Agent System for Nuclear Reactor," Proceedings of the Korean Nuclear Society Spring Meeting, Jeju, Korea, 2025.

[2] G. Han and Y. Yu and S. G. Kim, " Development of an Autonomous Reactor Operation LLM Agent for Real-Time Detection and Proactive Mitigation," Proceedings of the Korean Nuclear Society Autumn Meeting, Changwon, Korea, 2025.

[3] J. Y. Keum and J. G. Choi, " Study on Regulatory Status of Artificial Intelligence Technology in Nuclear Instrumentation and Control Fields by the U.S. NRC," Proceedings of the Korean Nuclear Society Autumn Meeting, Changwon, Korea, 2025.

[4] Yoon Pyo Lee, Joowon Cha, Yonggyun Yu, Seung Geun Kim, "Large language model agent for nuclear reactor operation assistance," Nuclear Engineering and Technology, 2025.