Hierarchical Multi-Agent LLM Framework for Causal Factor Classification in Nuclear Power Plants

Sang Beom Kang[a], Jeong Jin Park[b], Dae Young Lee[a*]
[a]*FNC Technology Co., Ltd., 13 Heungdeok 1-ro, 32F, Giheung-gu, Yongin-si, Gyeonggi-do, 16954, Korea*
[b]*Central Research Institute, KHNP*
[*]*Corresponding author: ldy242@fnctech.com*

Keywords : Large Language Models (LLMs), Multi-Agent Systems, Prompt Engineering, Operational Experience (OE)

1. Introduction

In nuclear power plants, reliable classification of causal factors in Operational Experience (OE) reports is essential for recurrence prevention, safety enhancement, and regulatory compliance. International guidance such as WANO MN-01 requires a clear distinction between Direct Cause (DC) and Root Cause (RC), yet manual interpretations often suffer from ambiguous label definitions and inconsistent judgments across analysts.

Recent advances in Natural Language Processing (NLP), particularly Large Language Models (LLMs), provide new opportunities for automating OE analysis [1]. Initial evaluations demonstrated that LLM-based classification outperforms traditional keyword- or embedding-based approaches, especially when a Top-3 candidate strategy is applied to mitigate single-label errors [2].

2. Background

OE reports contain detailed descriptions of events, causes, and corrective actions, forming a critical basis for causal analysis and long-term safety trend identification. However, manual classification is labor-intensive and lacks reproducibility. Conventional text-matching or statistical methods cannot capture the semantic complexity of event narratives, underscoring the need for advanced automated solutions. This necessity forms the basis of the LLM-based multi-agent framework proposed in this study [3].

3. Classification Approaches

Using power-plant event reports, we systematically compared six representative classification methods, ranging from keyword-centric retrieval to LLM-based inference. Each experiment highlighted structural characteristics and limitations, providing the rationale for transitioning toward LLM-based multi-agent strategies.

BM25-Based Text Similarity Search:
A lexical retrieval method effective when keyword overlap exists, but limited in capturing semantic similarity and prone to superficial matches.

Embedding-Based Cosine Similarity:
Captures conceptual similarity through vector representations, yet loses nuance at detailed label levels, causing cumulative errors.

Hybrid Retrieval (BM25 + Embeddings):
Combines lexical filtering with semantic ranking, improving recall and coherence, though at the cost of higher complexity.

Cross-Encoder Direct Matching:
Jointly encodes report–label pairs for strong semantic modeling, but inference cost is too high for large-scale use.

Reranker-Based Reordering:
Refines candidate rankings with LLMs, reducing false positives, but cannot fully resolve structural mismatches in narratives.

After applying and evaluating the preceding classification methods, we found that LLM-based direct inference offered the most promising balance between accuracy and flexibility. We directly tested the ability of LLMs to classify event narratives. While single Top-1 predictions often resulted in frequent misclassifications, adopting a Top-3 candidate strategy substantially improved both recall and precision. This adjustment enabled the model to surface multiple plausible interpretations for ambiguous reports, ultimately providing the key motivation for extending the approach into a multi-agent framework.

4. Proposed Method: Hierarchical Multi-Agent LLM Framework

Single-model Top-1 predictions could not adequately capture event complexity, often resulting in misclassification. We therefore propose a hierarchical multi-agent framework in which specialized LLM agents collaboratively analyze reports in stages. Multi-agent architectures decompose complex tasks into modular roles, offering greater scalability, fault tolerance, and interpretability—advantages particularly beneficial in nuclear safety analysis.

4.1 Single-LLM Inference

A single LLM summarized each report and generated candidate codes using a prompt enriched with label definitions. While this surpassed keyword-based methods by enabling semantic classification, it suffered from compounding errors, reduced consistency, and limited interpretability.

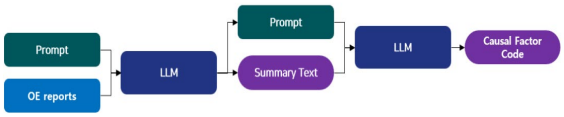

Fig. 1. Single LLM-Based Inference Architecture.

4.2 Proposer–Critic Feedback Loop

A Proposer–Critic structure was introduced to address these limitations. The Proposer generated summaries and candidate codes, while the Critic cross-checked them against the label schema, identifying inconsistencies for iterative refinement. This feedback enabled self-correction and stabilized Top-3 predictions, simulating the peer-review process of human experts and enhancing both reliability and explainability [4].

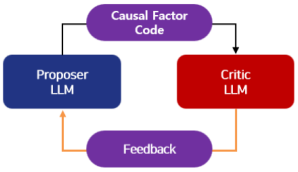

Fig. 2. Mutual Verification Architecture.

4.3 Direct Subclassification Pipeline

To improve operational efficiency, we developed a structured pipeline: summary → keyword extraction → DC classification → RC classification → consistency checking, with each step managed by a dedicated agent. This design streamlined the process compared to the feedback loop, providing greater efficiency and reproducibility in real-world scenarios. However, events with deeply layered causal chains continued to challenge interpretive depth.

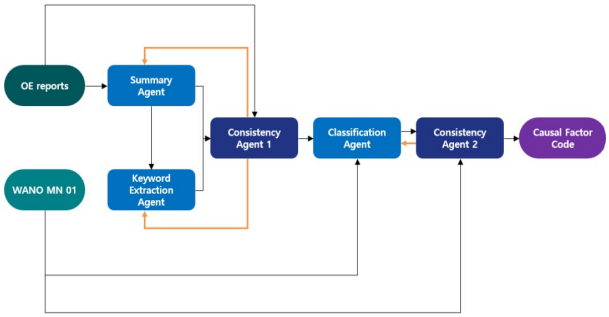

Fig. 3. Direct Subclassification-Based Architecture.

4.4 Three-Tier Hierarchical Scheme (Advanced Design)

The final framework structured DC and RC classification into three tiers (Level-1, Level-2, Level-3). Multiple agents collaborated to classify events, and a dedicated D/R Consistency Agent verified hierarchical coherence before finalizing Top-1 and Top-3 results. Beyond code prediction, the framework yields an explainable representation of multi-dimensional causal structures, providing actionable input for procedure improvement, training design, and policy development [5].

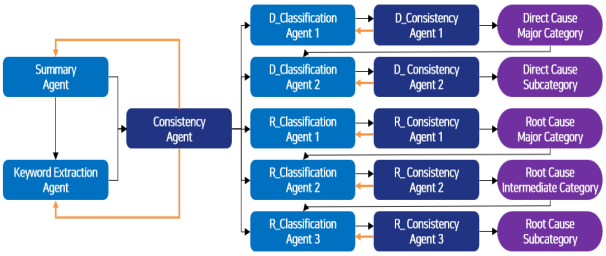

Fig. 4. Three-Tier Hierarchical Architecture.

4.5 Results

CASE 1 used the original dataset; CASE 2 employed consistency-verified data. DC classification achieved 95% and 90% accuracy, respectively, while RC classification reached 85% and 77.5%. The lower RC performance reflects their higher abstraction level, the propagation of small misjudgments across hierarchical steps, and the cumulative effects of ambiguous label definitions and inconsistent annotations. These findings underscore the importance of refining label schemas and incorporating domain-specific training to further enhance RC performance.

Table I: Experimental results of the multi-agent classification system.

Classification Level	CASE 1		CASE 2	
	Direct Cause	Root Cause	Direct Cause	Root Cause
High-Level Category	97.5	92.5%	92.5%	85%
Mid-Level Category	-	87.5%	-	82.5%
Low-Level Category	95%	85%	90%	77.5%

5. Conclusion

This study proposed a hierarchical multi-agent framework leveraging LLMs to overcome the limitations of manual, subjective DC/RC classification in power-plant event analyses. The framework evolved progressively—from a single-LLM baseline, to a Proposer–Critic feedback loop, to a direct

subclassification pipeline, and ultimately to a three-tier hierarchical scheme. Experimental results demonstrated DC accuracy above 90% and RC accuracy between 77.5–85%, confirming that even complex event narratives can be systematically structured into reliable causal codes.

Importantly, performance gains were driven not only by architectural sophistication but also by the explicit integration of precise label definitions within prompts. Iterative expert review and refinement of label schemas enhanced both consistency and explainability, establishing a solid foundation for further accuracy improvements [6].

Future work will focus on refining label schemas, expanding curated datasets, and applying domain-specific fine-tuning (e.g., instruction tuning, LoRA) to enhance classifier robustness. Further directions include validation in real operational environments, multilingual OE report analysis, and visualization-based trend exploration. Collectively, these efforts aim to advance the automation and reliability of nuclear event analysis, strengthening its applicability to long-term safety management and regulatory decision-making.

REFERENCES

[1] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, Xiangliang Zhang, Large Language Model-based Multi-Agents: A Survey of Progress and Challenges, arXiv preprint, 2024.
[2] Peipei Wei, Dimitris Dimitriadis, Yan Xu, Mingwei Shen, Don’t Just Demo, Teach Me the Principles: A Principle-Based Multi-Agent Prompting Strategy for Text Classification, arXiv preprint, 2025.
[3] Alberto Caballero, Roberto Centeno, Álvaro Rodrigo, LLM-Based Multi-Agent Models for Multiclass Classification of Strategic Narratives, CEUR Workshop Proceedings, DIPROMATS 2024.
[4] Thorsten Händler, A Taxonomy for Autonomous LLM-Powered Multi-Agent Architectures, KMIS 2023 Proceedings; extended arXiv preprint, 2023.
[5] X. Li, A Survey on LLM-Based Multi-Agent Systems: Workflow, Springer Journal, 2024.
[6] Hediyeh Baban, Sai A. Pidapar, Aashutosh Nema, Sichen Lu, Enhancing Text Classification with a Novel Multi-Agent Collaboration Framework Leveraging BERT, arXiv preprint, 2025.