Development of an Automated Pipeline for International Nuclear Information System Literature Registration using a LLM Multi-Agent Framework

Yohan Lee[a*], Anna Yoo[b] and Yonggyun Yu[a]
[a]*Applied Artificial Intelligence Section, Korea Atomic Energy Research Institute, Republic of Korea*
[b]*Technical Information Team, Korea Atomic Energy Research Institute, Republic of Korea*
[*]*Author E-mail: yhlee7021@kaeri.re.kr*

***Keywords : INIS, Large Language Model, Multi-Agent System, t-SNE, Literature Registration Automation**

1. Introduction

The International Nuclear Information System (INIS), established by the IAEA in 1970, serves as the world's most comprehensive repository for nuclear science literature [1]. INIS operates as a collaborative database where member countries contribute by registering nuclear-related literature. Although the Republic of Korea is one of the countries with high global INIS utilization, its literature contribution rate is significantly lower, raising concerns about the imbalance between its responsibility and its contribution to the international nuclear information ecosystem.

The Korea Atomic Energy Research Institute (KAERI) processes an average of 500 documents annually through manual registration, a figure that falls significantly short of what is needed to align Korea's contribution rate with its utilization rate in the INIS system. Achieving this balance would require a substantial increase in processing capacity that cannot be realistically accomplished through traditional manual methods. The manual registration process, which relies on external domain experts to assign INIS Subject Categories and Descriptors, faces inherent constraints due to staff limitations and the cognitive complexity of the task. Unlike routine data processing operations, this work requires deep domain expertise that cannot be easily automated through simple algorithmic approaches, creating a fundamental bottleneck that limits Korea's ability to fulfill its collaborative obligations within the international nuclear information ecosystem.

This study presents a novel automated pipeline leveraging Large Language Model (LLM) multi-agent architecture to overcome these operational constraints, achieving a 12x increase in processing speed while reducing costs by over 95%.

2. System Architecture and Methodology

The developed automated pipeline integrates two primary modules that work in tandem to transform unstructured documents into INIS-compliant records: the Metadata Extractor & INIS Tag Generator and the LLM Document Analyzer, as illustrated in Figure 1.

2.1. Dual-Module Processing Architecture

The system employs a bifurcated processing approach where document metadata extraction operates in parallel with semantic content analysis. Upon document ingestion, the pipeline simultaneously initiates metadata extraction through DOI-based API queries and comprehensive text analysis through the LLM Document Analyzer module. This parallel processing architecture ensures optimal throughput while maintaining data integrity throughout the transformation process.

2.2. Metadata Extractor and INIS Tag Generator

The left module handles structured information extraction and final record compilation. When a PDF document enters the system, the metadata extractor parses bibliographic elements including title, authors, institutional affiliations, and abstract. For documents containing Digital Object Identifiers, the system queries CrossRef API to retrieve authoritative metadata, ensuring accuracy and completeness of bibliographic records. This verified metadata forms the foundation of the INIS record structure, with the module ultimately generating the complete INIS tag that encapsulates both bibliographic information and the semantic descriptors produced by the LLM Document Analyzer.

2.3. Multi-Agent LLM Document Analyzer

The core analytical capability resides in the LLM Document Analyzer module, which orchestrates multiple specialized agents operating in parallel. Each agent—represented by distinct LLM instances with role-specific system prompts—analyzes the full text content from a unique professional perspective. The Technical Agent focuses on reactor designs and experimental methodologies. The Safety Agent evaluates radiation protection and risk assessment aspects, while the Policy Agent identifies non-proliferation and regulatory implications. This multi-perspective approach generates a comprehensive pool of candidate keywords that captures the document's multifaceted nature more effectively than single-pass analysis. In this study, we utilized the Gemini 2.5 Pro API provided by Google.

2.4. Addressing the Challenge of INIS Thesaurus Compliance

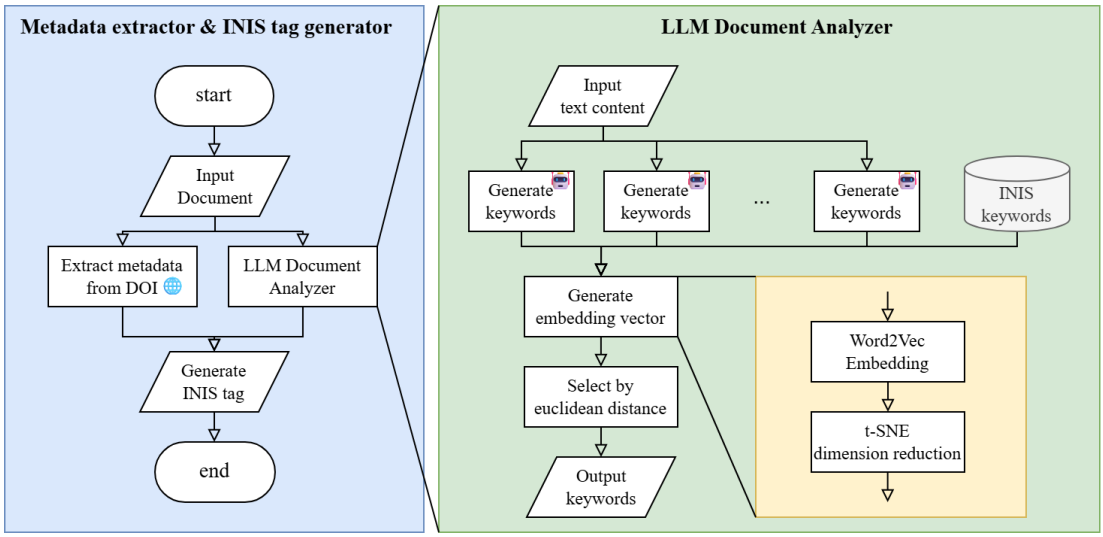

Fig 1. System Architecture Diagram showing the dual-module pipeline with metadata extraction, multi-agent LLM analysis, and low-dimensional keyword mapping using t-SNE

A critical constraint in INIS literature registration is the requirement to use only officially defined categories and descriptors from the INIS Thesaurus, which contains over 30,000 strictly controlled terms [2]. This presents a significant technical challenge: the system must map freely generated LLM keywords to this fixed vocabulary with high precision. Traditional word embedding models generate vectors in hundreds of dimensions, typically 300 to 768 dimensions depending on the model architecture. When attempting to calculate cosine similarity directly between these high-dimensional vectors, the results become extremely unstable due to the curse of dimensionality, where distance metrics lose their discriminative power and tend to converge toward constant values.

Furthermore, natural language exhibits inherently nonlinear semantic relationships that cannot be adequately captured through linear distance metrics in high-dimensional spaces. Words with similar meanings may be distributed across complex manifolds in the embedding space rather than forming simple linear clusters. This nonlinearity necessitates a more sophisticated approach to preserve semantic relationships while enabling reliable similarity calculations.

2.5. Low-Dimensional Keyword Mapping via t-SNE

To overcome these challenges, we implement a nonlinear dimensionality reduction strategy using t-Distributed Stochastic Neighbor Embedding (t-SNE) [3]. This algorithm is particularly well-suited for preserving local neighborhood structures in nonlinear data, making it ideal for capturing the complex semantic relationships inherent in natural language. The process operates in two phases:

During the preprocessing phase, all INIS Thesaurus descriptors undergo Word2Vec embedding to generate high-dimensional semantic vectors. These vectors are then transformed through t-SNE into a three-dimensional representation. This dramatic dimensionality reduction—from hundreds of dimensions to just three—eliminates the computational instability while preserving the essential semantic relationships between terms. The t-SNE algorithm's strength lies in its ability to maintain local structures, ensuring that semantically related INIS descriptors remain clustered together in the reduced space.

During runtime processing, candidate keywords generated by the multi-agent system are embedded using the same Word2Vec model and projected into the pre-computed three-dimensional t-SNE space. The system then calculates simple Euclidean distances between the projected candidates and all official INIS descriptors. This approach provides stable, reproducible keyword selection that would be impossible with direct high-dimensional comparisons. The three-dimensional space allows for reliable distance calculations while the nonlinear transformation preserves the complex semantic relationships necessary for accurate keyword matching.

2.6. Integration and Output Generation

The final stage synthesizes outputs from both processing modules. The metadata extractor combines structured bibliographic information with the semantically validated keywords from the LLM Document Analyzer to generate a complete INIS-compliant record. This integrated approach ensures that each document receives comprehensive treatment—accurate bibliographic representation coupled with expert-level semantic categorization—while maintaining the processing efficiency necessary for large-scale deployment. The strict adherence to INIS Thesaurus terminology through the t-SNE mapping ensures that all generated records meet the system's-controlled vocabulary requirements, enabling seamless integration

into the international database.

3. Quantitative Efficiency Evaluation

To assess the operational improvements achieved by the automated pipeline, we conducted a comparative analysis based on processing 500 documents, which represents the typical annual workload under the current manual system. Table I presents the transformation in key performance metrics between manual and automated processing approaches.

Table I. Efficiency Comparison: Manual vs. Automated Processing

Metric	Manual process	Automated process
Processing time	~ 1 year	~ 1 month
Cost	Tens of millions KRW	Tens of thousands KRW
Required personnel	More than 6 experts	1-2 reviewers
Human error rate	10-20%	< 1%

The automated pipeline demonstrates transformative improvements across all operational metrics. While manual processing requires approximately one year for domain experts to analyze and register 500 documents, the automated system completes the computational analysis within a single day through API-based document processing, with the total timeline extending to approximately one month when including comprehensive human review and validation. This twelve-fold reduction in processing time is accompanied by a cost reduction exceeding 95%, shifting from tens of millions of won in labor costs to mere tens of thousands in API usage fees. The required workforce transitions from more than six domain experts to just one or two reviewers who focus on quality assurance rather than content generation.

Most significantly, the system reduces human error rates from 10-20% to less than 1%. Manual registration errors typically occur in metadata transcription, such as incorrectly counting figure numbers, misspelling author names, or inaccurately recording institutional affiliations. The automated system eliminates most of these transcription errors through direct extraction from source documents and API-based verification. These quantitative improvements establish that the automated pipeline not only enables Korea to meet its 2,000-document annual target but also provides a sustainable, cost-effective framework for maintaining contribution-utilization parity in the INIS system.

4. Discussion

The developed pipeline enables KAERI to exceed its 2,000-document annual target, effectively resolving the contribution-utilization imbalance. The t-SNE dimensionality reduction proved crucial for stable keyword selection, overcoming the instability observed in high-dimensional cosine similarity calculations. The multi-agent architecture successfully captured diverse analytical perspectives, producing more comprehensive keyword sets than single-model approaches.

However, two limitations require consideration. First, the inherent subjectivity in subject categorization necessitates maintaining human oversight for final validation. Second, security concerns regarding sensitive nuclear information currently restrict the system to publicly available documents when using commercial LLM APIs.

5. Conclusion and Future Work

This research successfully demonstrates an LLM multi-agent pipeline that revolutionizes INIS literature registration, achieving dramatic improvements in efficiency, cost, and scalability. The system represents a practical application of advanced AI to scientific information management, providing a pathway for enhanced international collaboration in nuclear knowledge sharing.

Future development will focus on: (1) Implementing on-premises LLMs (e.g., Llama 3, Mistral, or AtomicGPT [4]) to eliminate security constraints while maintaining performance [5], and (2) Developing autonomous document collection agents using web APIs and classification models to create a fully automated end-to-end registration system [6].

REFERENCES

[1] International Atomic Energy Agency, International Nuclear Information System (INIS), IAEA, Vienna, 2024.

[2] International Atomic Energy Agency, INIS Thesaurus English, IAEA-INIS Reference Series IAEA-INIS-01, ISSN 1684-095X, IAEA, Vienna, September 2018.

[3] van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data, Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.

[4] Yeom Seung Don, ChangSu Choi, Lim KyungTae, & Yu Yong Gyun (2024-11-20). Development and Performance Evaluation of a Domain-Specific Language Model for Nuclear: A Comparative Study Using a Custom-Built Dataset. Proceedings of Symposium of the Korean Institute of communications and Information Sciences, Gyeongsangbuk-do.

[5] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*(2), 100211. doi:10.1016/

[6] Wang, L., Ma, C., Feng, X. *et al.* A survey on large language model based autonomous agents. *Front. Comput. Sci.* 18, 186345 (2024). https://doi.org/10.1007/s11704-024-40231-1