Nuclear Compliance: Assessing the Utility of LLMs with RAG

Isak Hwang[a], Yoon Pyo Lee[ab]*
[a]*Hanyang University Nuclear Engineering, 222 Wangsimni-ro, Seongdong-gu, Seoul, 04763, South Korea*
[b]*The Grainger College of Engineering, Nuclear, Plasma & Radiological Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA*
**Corresponding author: yoonpyo2@illinois.edu*

***Keywords : Nuclear safety, LLM, RAG, LLM Evaluation**

1. Introduction

Advances in artificial intelligence (AI) and robotics are reshaping the industrial landscape, but their sustainability hinges on a stable and large-scale power supply [1]. Nuclear energy offers a viable solution to meet growing demand while achieving carbon neutrality [2]. In particular, the introduction of next-generation nuclear power systems is dramatically increasing the complexity of the operational and regulatory environment. In this safety-critical environment, integrating advanced AI-based systems can help improve staff efficiency and support compliance with nuclear regulatory documents.

Large language models (LLMs), such as GPT 5.2 and Gemini 3.0 Pro [4,5], offer the advantage of fast document processing speed, but the potential for generating incorrect information poses a significant risk in security-sensitive areas [6]. To mitigate this, researchers have explored Retrieval-Augmented Generation (RAG) to ensure accuracy by grounding AI responses in technical documentation [7]. For instance, Anwar et al. [8] demonstrated the necessity of RAG for nuclear domain-specific data, and Lee et al. [9] explored LLM agents for reactor operation assistance. While previous research has demonstrated the potential of RAG, the most appropriate LLM model to meet the strict requirements of nuclear safety regulations remains to be determined.

This study uses the International Atomic Energy Agency (IAEA) safety standards document [3]. To process this document for system design and evaluation, two different chunking methods are used. The RAG provided by LLM utilizes a common "token-based chunking" approach [10]. The evaluation dataset was generated using a "clause-based semantic chunking" approach [11]. This method chunks documents into clauses.

To quantify reliability, we utilize two key metrics: citation accuracy and Bidirectional Encoder Representations from Transformers (BERT) score [12]. These metrics prevent LLM-as-a-judge [13], enhancing the objectivity of nuclear safety assessments. We introduce RAG Quality Evaluation Score (RAG-QES), a novel quantitative evaluation metric that integrates citation accuracy and BERTScore. RAG-QES comprehensively evaluates system performance in terms of nuclear regulatory compliance.

Therefore, in this study, we benchmarked several state-of-the-art (SOTA) models, including Gemini 3.0 Pro/Flash, GPT-5.2, and Claude Sonnet 4.5 [14], to determine which LLM is most suitable for nuclear regulatory documents.

2. Methodology

This study aimed to demonstrate the feasibility of attaching RAG to commercial LLMs. The methodology consists of (1) data generation, (2) experimental setup, and (3) evaluation.

2.1. Data Generation

The evaluation requires Question-Answer (QA) sets. The source of information for these QA sets is the IAEA safety standards. To create these sets, we adopted a semantic chunking strategy that explores the structural components of regulatory documents, rather than the typical token-based chunking approach. We first analyzed the documents by clause, then chunked them into sub-clauses within each clause. By separating individual sub-clauses, we maintained the specific context of each clause and avoided information loss or semantic ambiguity that commonly occurs when multiple sub-clauses are merged into a single chunk. From the extracted chunks, we generated a total of 220 QA sets using Gemini 3.0 Pro.

2.2. Experimental Setup

This experiment evaluates all LLMs using the evaluation dataset. The testing procedure is shown in Figure 1.

1. RAG: Build a RAG pipeline for each regulatory document using the LLM embedding model recommended by the developer. Table 1 below shows which embedding model was used for each LLM.

2. QA: From the extracted sub-clauses, we generated a total of 220 corresponding QA sets using Gemini 3.0 Pro to construct the evaluation dataset.

Table 1. Embedding models used

Model Name	Developer	Embedding model
Gemini 3.0 Pro	Google	gemini-embedding-001
Gemini 3.0 Flash	Google	gemini-embedding-001
GPT 5.2	OpenAI	text-embedding-3-small
Claude Sonnet 4.5	Anthropic	voyage-4-lite

3. Evaluation: The target LLMs generate responses to the prepared queries from the QA dataset. The accuracy and semantic similarity of citations between the responses generated by the LLM and those in the dataset are evaluated. These scores are used in RAG-QES.

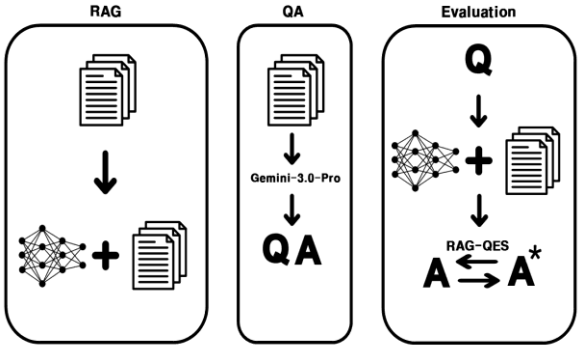

Fig 1. A workflow that shows the experiment progress.

2.3. Evaluation Metrics

These metrics implemented an evaluation protocol to verify logical reliability, and accuracy of citation sources of the generated answers. The RAG system was validated using the following two key metrics:

1. Citation Accuracy: To assess the reliability of the references provided by the model, we evaluated citation accuracy. We extracted the cited regulatory numbers from the generated responses and compared them with the actual source clause numbers.

2. BERTScore: Instead of relying on simple word matching, we used BERTScore to evaluate how well the generated responses retain the meaning of the actual answers. This metric calculates the semantic similarity between the model-generated answers and the answers in Section 2.1, capturing subtle contextual differences.

To provide a holistic assessment of the model's trustworthiness and accuracy, we introduced the RAG-QES (Eq 1). This metric is a weighted sum of semantic similarity and reliability.

$$\text{(Eq 1) RAG-QES} = \omega_1\{\text{CA}\} + \omega_2\{\text{BERT}\}$$

{CA} indicates citation accuracy. A match between the subclause regulation numbers earns 1 point, a match between only the requirement numbers earns 0.5 points, and no match earns 0 points. {BERT} indicates the normalized BERTScore. Since the score distribution ranges from approximately 0.8 to 1.0, normalization was performed to improve evaluation accuracy. Normalization was performed using Eq 2.

$$\text{Eq 2. } \{\text{BERT}\} = \frac{BERTScore - 0.8}{1 - 0.8}$$

Since both citation accuracy and semantic accuracy are important, we set the weights to ω_1=0.5 and ω_2=0.5 for calculation.

3. Results

This section presents the normalized BERTScore, citation accuracy, and RAG-QES evaluation results for the four LLMs (Gemini 3.0 Pro, Gemini 3.0 Flash, GPT-5.2, and Claude Sonnet 4.5). All experiments were conducted under identical conditions using the QA sets.

Figure 2 illustrates the distribution of normalized BERTScore, which quantifies the semantic similarity between the generated responses and the ground truth. Gemini 3.0 Pro recorded the highest median score, indicating a strong baseline performance in capturing technical semantics. Based on median score, Gemini 3.0 Flash ranks next, followed by Claude Sonnet 4.5. Meanwhile, GPT-5.2 has the most outliers and the lowest median.

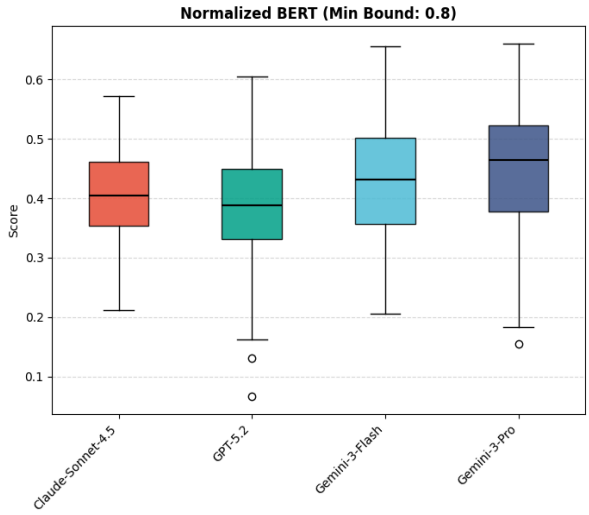

Fig 2. Normalized BERTScore distribution of LLMs

Figure 3 presents the citation accuracy, a metric assessing whether the model references the correct regulatory provision. Claude Sonnet 4.5 achieved the highest reliability with a score of 0.62, followed by Gemini 3.0 Flash and Gemini 3.0 Pro. Conversely, GPT-5.2 recorded the lowest score of 0.46, demonstrating a high frequency of incorrect citations compared to the other models.

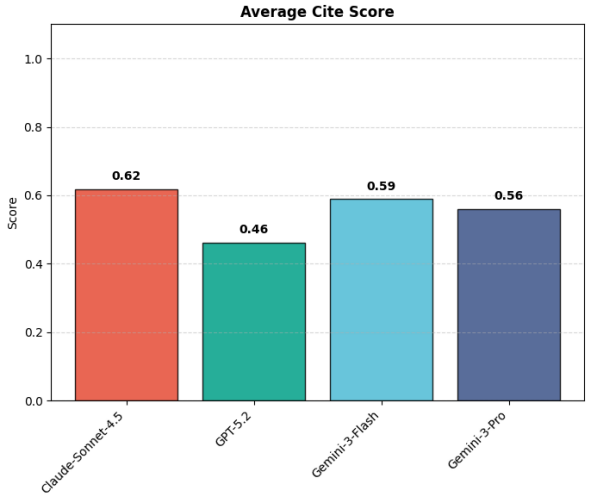

Fig 3. Average citation accuracy of the LLMs.

Figure 4 displays the comprehensive RAG-QES results, integrating both semantic and citation metrics. The highest median RAG-QES score was observed in Claude Sonnet 4.5, reflecting consistent overall performance. GPT-5.2, however, recorded the lowest overall and minimum scores, indicating the widest performance gap among the tested models.

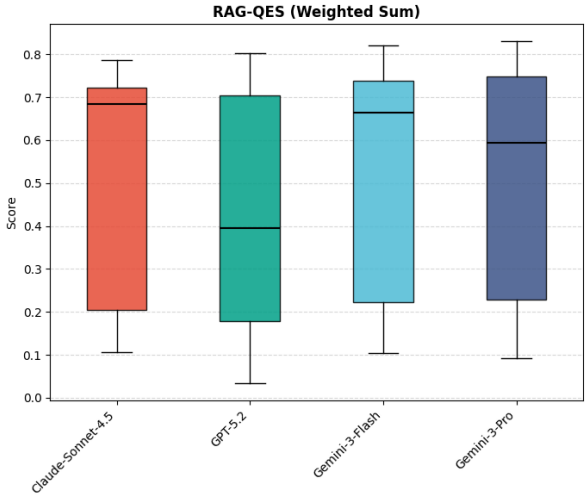

Fig 4. Distribution of RAG-QES for LLMs

4. Discussion & Conclusion

As shown in Table 2, recent benchmarks such as Massive Multitask Language Understanding (MMLU) -Pro assess the ability of an inference engine to solve complex inference tasks. Therefore, models with high MMLU-Pro scores are expected to demonstrate semantic understanding (high normalized BERTScore) in a complex regulatory environment. The results in Figure 2 support this hypothesis, revealing alignment between the models' overall MMLU-Pro rankings and their median normalized BERTScore rankings. Gemini 3.0 Pro achieved the highest median normalized BERTScore, demonstrating a capability to interpret technical nuances. As shown in the median score rankings in Table 2, Gemini 3.0 Flash and Claude Sonnet 4.5 followed in that order. GPT-5.2 exhibited the lowest median scores and two outliers, indicating instability in semantic inference. This correlation suggests that a model's general MMLU-Pro performance can serve as a reliable predictor of its semantic accuracy in specialized technical applications.

Table 2. MMLU-Pro Score

Model Name	Developer	Score
Gemini 3.0 Pro	Google	89.8
Gemini 3.0 Flash	Google	89.0
GPT 5.2	OpenAI	85.9
Claude Sonnet 4.5	Anthropic	87.5

Figure 3 shows the results of evaluating citation accuracy. Key factors determining citation accuracy in a RAG system include the data chunking method, the retrieval performance of the embedding model, and the LLM's ability to adhere to prompt constraints. In this study, the same chunking method was applied as a control variable to all four target LLMs. Therefore, the difference in citation accuracy between models was expected to be primarily proportional to the objective performance of the embedding model and the language model's ability to adhere to guidelines. The specific metrics for these capabilities are provided in Table 3 for the retrieval embedding benchmark (RTEB) scores, which evaluate how accurately the model finds relevant documents, and Table 4 for the instruction following benchmark (IFBench) scores, which measure the model's precision in executing specific formatting and logical constraints, alongside Figure 5 which displays the common prompts. However, the actual evaluation results differed from these expectations. Claude Sonnet 4.5 achieved a citation accuracy of 0.62, while GPT-5.2 achieved 0.46, a difference difficult to explain based on prior metrics. Explicit evaluation metrics and controlled environments alone cannot fully explain this discrepancy, suggesting the existence of underlying factors that influence final citation performance. These findings indicate that existing benchmark scores cannot fully

Table 3. RTEB Score

Model Name	Developer	Overall	English
gemini-embedding-001	Google	75.85	71.02
text-embedding-3-small	OpenAI	59.24	53.98
voyage-4-lite	Voyage AI	75.07	71.23

Table 4. IFBench Score

Model Name	Developer	Score
Gemini 3.0 Pro	Google	70.4
Gemini 3.0 Flash	Google	78.0
GPT 5.2	OpenAI	65.2
Claude Sonnet 4.5	Anthropic	57.3

account for the evaluation results. Specifically, since IFBench evaluates models outside of a RAG environment, it cannot serve as an absolute predictor of citation performance. Consequently, there is a clear need for a novel benchmark specialized for RAG-based compliance.

```
system_prompt = """You are a strict compliance auditor for IAEA Safety Standards.
Answer based **ONLY** on the provided [Context].

[Citation Rules - STRICT]
1. **Sub-clause Number:** You MUST cite the exact paragraph number (e.g., **3.14**, **4.2**).
2. **Requirement ID:** If applicable, cite the Requirement number (e.g., Requirement 12).
3. **Format:** Start your answer with the citation tag like [Source: Requirement 12, Para 3.14].
4. **Consistency:** Do not omit the number if it appears in the text.
```

Fig 5. Prompt Content

Figure 4 presents the evaluation results using RAG-QES, a metric designed to score system performance by assigning equal weight to semantic alignment and citation accuracy. While Gemini 3.0 Pro excels at interpreting complex regulatory logic, its median score is not the highest among the evaluated models. In contrast, Claude Sonnet 4.5 demonstrates outstanding consistency and response reliability, as evidenced by its highest median score and overall narrow distribution. These findings indicate that advanced reasoning capabilities alone are insufficient to guarantee strict regulatory compliance. Achieving optimal RAG performance requires the synergistic integration of multiple variables, including a high-performance inference engine and an accurate, domain-optimized embedding model.

This study aimed to analyze the information processing and compliance capabilities of SOTA LLMs in relation to nuclear regulatory documents. To overcome the subjectivity and bias of the 'LLM-as-a-judge' approach, this study introduced RAG-QES, a novel quantitative evaluation metric that combines citation accuracy and normalized BERTScore to comprehensively assess nuclear regulatory compliance. Using this metric, we empirically benchmarked SOTA commercial models, including Gemini 3.0 Pro/Flash, GPT-5.2, and Claude Sonnet 4.5, under the stringent conditions of nuclear safety regulations, providing a practical basis for identifying the most appropriate model for this application. While this experiment evaluated RAG performance using the recommended embedding models for each LLM, the stringent data security and closed-network requirements of nuclear facilities pose fundamental limitations on the practical application of these commercial cloud-based LLMs. Therefore, future research will focus on developing and optimizing local LLMs specialized for nuclear regulatory compliance. These local models will be evaluated using the proposed RAG-QES metric and compared with the baseline performance of the commercial models established in this study.

REFERENCES

[1] International Atomic Energy Agency. (2022). Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology. International Atomic Energy Agency.
[2] Yongsung Lee and Hyun Seok Kim. (2021). Comparison of Cost-Efficiency of Nuclear Power and Renewable Energy Generation in Reducing CO2. Environmental and Resource Economics Review, 30(4), 607-625.
[3] International Atomic Energy Agency. (2016). Safety Assessment for Facilities and Activities (IAEA Safety Standards Series No. GSR Part 4, Rev. 1). IAEA.
[4] OpenAI. (2025). GPT-5.2 technical report. OpenAI. https://openai.com/ko-KR/index/introducing-gpt-5-2/
[5] Google DeepMind. (2025). Gemini 3: A new era of intelligence. Google DeepMind. https://deepmind.google/models/gemini/
[6] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.
[7] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
[8] Anwar, M., de Costa, M., Hammad, I., & Lau, D. (2024). Evaluating ChatGPT on Nuclear Domain-Specific Data. arXiv preprint arXiv:2409.00090.
[9] Lee, Y. P., Cha, J., Yu, Y., & Kim, S. G. (2025). Large language model agent for nuclear reactor operation assistance. Nuclear Engineering and Technology, 103842.
[10] Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense Passage Retrieval for Open-Domain Question Answering. In EMNLP (1) (pp. 6769-6781).
[11] Latif, S., Ameer, H., Akram, M. H., & Fatima, M. (2025, August). The Chunking Paradigm: Recursive Semantic for RAG Optimization. In Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025) (pp. 137-145).
[12] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.
[13] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36, 46595-46623.
[14] Anthropic. (2025). Claude Sonnet 4.5 System Card. https://www.anthropic.com/claude-sonnet-4-5-system-card