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

1. Introduction

Motivation: Advanced AI systems are required to support staff efficiency and strict regulatory compliance as nuclear operational environments become increasingly complex.

Problem: Standard Large Language Models (LLMs) suffer from hallucinations—generating incorrect information—which poses significant security risks in safety-critical nuclear domains.

Objective: This study introduces RAG-QES, a novel quantitative evaluation metric, to benchmark state-of-the-art LLMs (GPT, Gemini, Claude) using IAEA safety standards and RAG technology.

2. Methodology

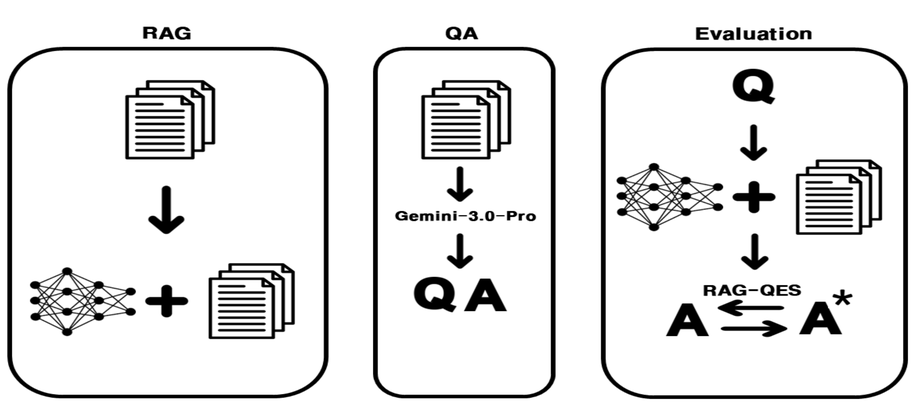
1) Data Generation

Data Source: IAEA Safety Criteria Document.

Data Chunking Method: To prevent information loss and contextual ambiguity, a semantic chunking strategy centered on document structure (clauses and sub-clauses) was applied instead of general token units.

Generation Result: Based on the extracted chunks, a total of 220 QA sets were generated using the Gemini 3.0 Pro model.

2) Experimental Setup



RAG Pipeline Construction: A RAG based on safety regulatory documents is constructed using the embedding model recommended by each LLM developer (e.g., gemini-embedding-001 for Google).

Evaluation Dataset (QA) Generation: Based on sub-clause partitioned by semantic units, a set of 220 Q&A (QA) items is generated using Gemini 3.0 Pro.

Final Evaluation (RAG-QES): The final RAG-QES score is calculated by comparing the Citation Accuracy and Semantic Similarity (using BERTScore) between the answers generated by the evaluated LLMs and the correct answer dataset.

3) Evaluation Metrics

To avoid subjectivity (LLM-as-a-judge), we introduced the RAG Quality Evaluation Score (RAG-QES), integrating logical reliability and semantic accuracy.

$$\textbf{RAG-QES} = \omega_1\{CA\} + \omega_2\{BERT\}$$

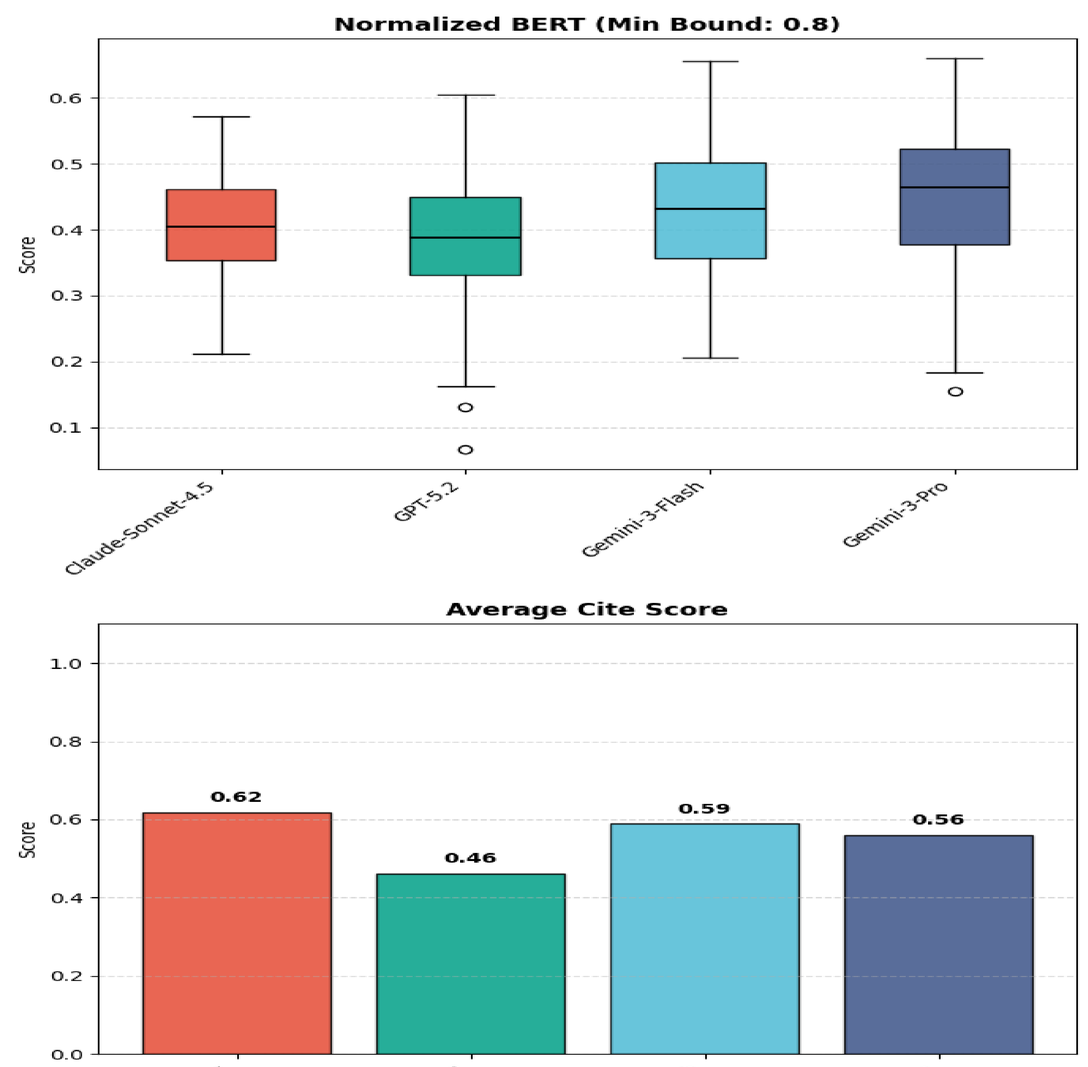
Citation Accuracy = {CA}
Measures accuracy of references. Match on sub-clause (1.0), Requirement only (0.5), or no match (0.0).

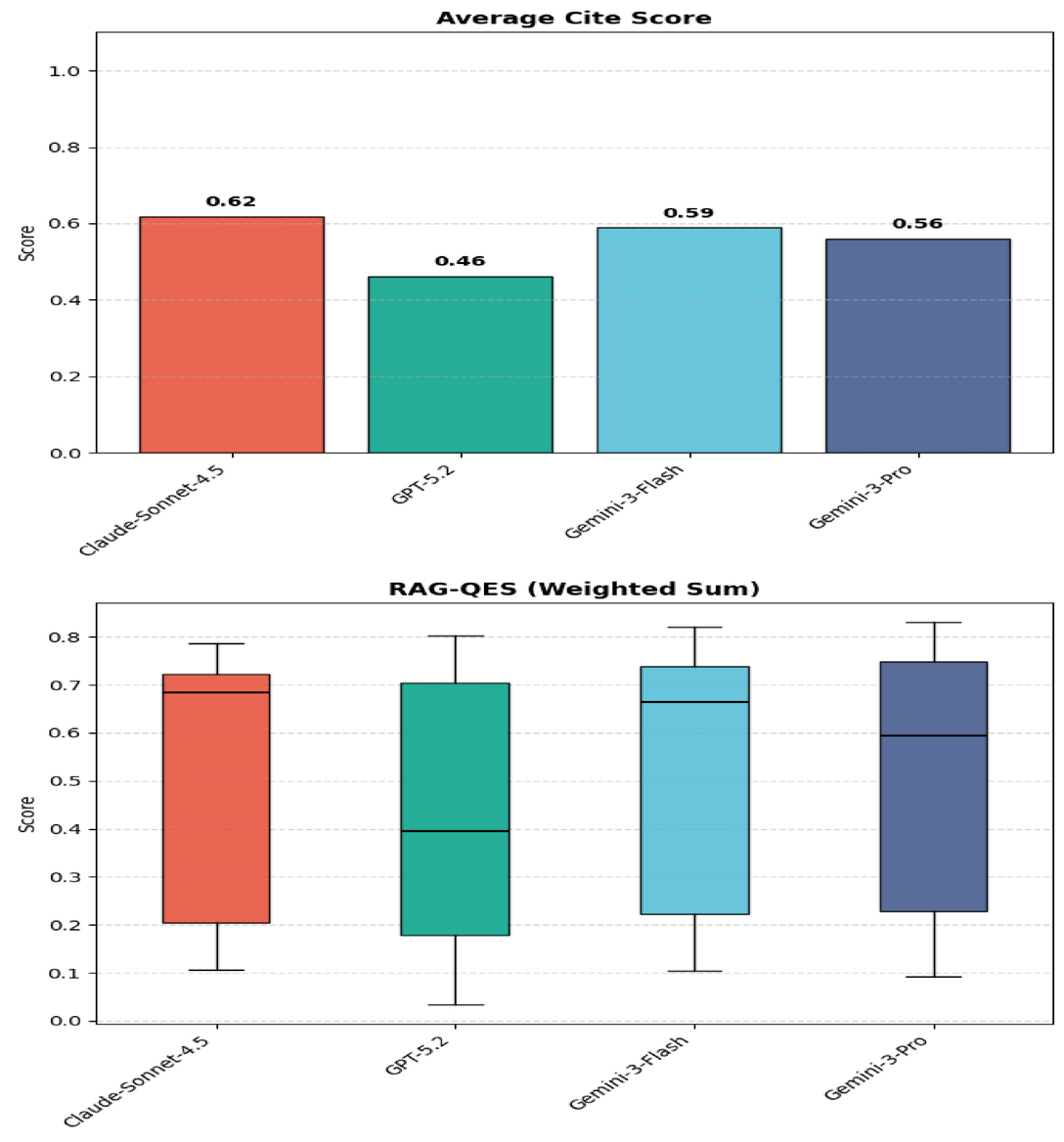
Normalized BERTScore = {BERT}

$$\{BERT\} = \frac{BERTScore - 0.8}{1 - 0.8}$$

3. Experiments & Key Finding

Experiments
Target LLMs generated responses based on the IAEA QA dataset. To calculate the comprehensive RAG-QES, equal weights (ω_1=0.5, ω_2=0.5) were applied to both citation accuracy and normalized BERTScore to ensure a perfectly balanced assessment of reliability and semantic understanding.





Key Finding

Semantic Accuracy: Gemini 3.0 Pro demonstrated the best baseline performance in capturing complex technical nuances.

Citation Reliability: Claude Sonnet 4.5 exhibited outstanding precision in adhering to strict formatting and citation guidelines.

Overall Capability (RAG-QES): Claude Sonnet 4.5 achieved the highest median score, proving the most consistent overall performance. Conversely, GPT-5.2 recorded the most outliers and lowest median, indicating severe instability in specialized AI-driven regulatory reasoning.

4. Conclusion

We introduced RAG-QES to evaluate SOTA LLMs on nuclear regulatory compliance. Because strict nuclear data security limits cloud-based applications, future research will develop secure, local LLMs and benchmark them against these commercial baselines using RAG-QES.