

Article

A Retrieval-Augmented Generation with Dual-Similarity Monitoring for Nuclear Energy Knowledge Q&A

Cheng-Hsing Chiang and Kun-Chou Lee * 

Department of System and Naval Mechatronic Engineering, National Cheng Kung University, Tainan 701, Taiwan; p16131139@gs.ncku.edu.tw

* Correspondence: kcle@ncku.edu.tw

Abstract

We present a Retrieval-Augmented Generation (RAG)-based question-answering system for nuclear energy science communication, characterizing retrieval quality in generated responses. The system introduces a dual-similarity analysis that jointly measures (i) question-to-context ($Q \rightarrow C$) and (ii) answer-to-context ($A \rightarrow C$) semantic consistency, serving as “retrieval-side semantic alignment signal” and “post-generation semantic alignment indicator” respectively. Built with LangChain, FAISS retrieval, and a large language model, our pipeline separates offline indexing from online inference and is grounded on authoritative Taiwanese Nuclear Safety Commission documents. We evaluate two settings: (a) in-domain prompts derived from the corpus and (b) out-of-domain, randomly generated nuclear energy questions. Results show that generated answers are, on average, more semantically similar to retrieved contexts than the original questions under the present setup, while the overall association between retrieval-side and answer-side signals remains stronger in the in-domain setting. Out-of-domain questions show weaker but still observable answer-to-context alignment patterns, contingent on corpus overlap. These findings suggest that combining RAG with dual-similarity analysis offers a practical and audit-oriented approach for educational Q&A, and we discuss potential improvements in versioned regulations, re-ranking, and abstention strategies. In this study, the RAG technique and dual-similarity analysis are combined together to promote nuclear energy knowledge. The research flow chart of this study can be applied to many other fields of scientific knowledge.

Keywords: retrieval-augmented generation; FAISS; LangChain; semantic similarity; semantic alignment; nuclear energy education; Q&A



Academic Editor: Luis Javier Garcia Villalba

Received: 22 January 2026

Revised: 16 March 2026

Accepted: 19 March 2026

Published: 26 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. Motivation

As generative AI rapidly permeates knowledge dissemination and educational practice, ensuring that answers are “evidence-grounded, source-traceable, and auditable” has become a central challenge. While large language models (LLMs) excel at fluent generation, their outputs can be constrained by the recency of training data and susceptible to hallucinations, yielding inconsistencies or unverifiable statements in high-stakes domains such as policy, regulation, technical terminology, and quantitative facts. Recent surveys synthesize mechanisms and taxonomies of LLM hallucinations, highlighting reliability risks for high-stakes uses [1,2]. This gap is especially acute in nuclear energy and radiation protection—fields that are highly specialized, tightly regulated, and time-sensitive.

Documents are frequently revised; versioning and effective dates are complex; and public trust in risk communication hinges on authoritative sources and transparency. Without rigorous evidence constraints and post hoc monitoring, generative systems struggle to earn credibility and may even lead to poor decisions or communicative breakdowns. These reliability concerns are widely documented in recent studies on LLM hallucinations [3], reinforcing the need for auditable grounding in high-stakes domains.

Retrieval-Augmented Generation (RAG) offers a promising path to mitigate these risks by separating concerns: retrieval delimits the visible knowledge scope, while generation organizes language. Constraining responses to citable passages should, in principle, support both timeliness and traceability. RAG explicitly couples a non-parametric retriever with a parametric generator to ground outputs in external knowledge [4]. RAG formalizes the coupling of non-parametric retrieval with parametric generation, and recent surveys systematize its components and evaluation practices across tasks. However, many existing RAG implementations face two persistent gaps:

- (1) The lack of quantitative criteria for “did retrieval fetch the right evidence?” and “is the generated answer semantically aligned with the retrieved evidence?”;
- (2) Insufficient, practice-oriented evaluation of RAG robustness under in-domain versus out-of-domain user queries (i.e., whether questions align semantically with the corpus).

In short, even with RAG, absent explicit ingress and egress measurements, it is difficult to satisfy educational and public-sector requirements for auditability and risk control.

Nuclear energy science communication further presents three structural challenges:

- (i) Strong interdisciplinarity—spanning engineering, medicine, regulation, and public communication;
- (ii) Local contextualization—Taiwan-specific norms and institutional terminology that general-purpose corpora may not cover;
- (iii) Trust sensitivity—audiences expect explicit sources, versions, and statutory bases.

These traits make general LLMs alone insufficient for real-world needs and underscore the necessity of a system that is grounded in local, authoritative documents, quantitatively monitorable, and traceable. Similar efforts in other socially sensitive domains also highlight the value of retrieval grounding: Jung and Joe [5] demonstrated a domain-optimized docent system integrating sLLMs with FAISS/OpenSearch-based RAG, while Castellanos-Nieves and García-Forte [6] developed a human-centered RAG framework for migrant integration, emphasizing contextualization and societal trust.

Motivated by these considerations, this study builds a RAG pipeline grounded in Taiwanese official documents on nuclear energy and radiation protection, and proposes a dual-similarity analysis as a system-level metric. Recent evaluations formalize component-level metrics—contextual recall/precision for retrieval and answer faithfulness/attribution for generation—to operationalize grounding [7,8]. In contrast, our A→C score is used only as a semantic alignment indicator in embedding space. The first measure quantifies Question-to-Document semantic consistency for ingress gating and passage ranking during retrieval; the second assesses Answer-to-Document consistency for egress auditing and risk signaling after generation. By combining “front-end assurance of topicality” with “back-end monitoring of semantic alignment with retrieved evidence”, we aim to meet three requirements in educational settings:

- (A) Accuracy and traceability of content;
- (B) Observability and explainability of system behavior;
- (C) Robustness to both in-domain and out-of-domain queries.

In education-specific contexts, emerging evidence suggests both opportunities and risks: empirical studies report measurable impacts of generative AI on learning [9], while domain prototypes show that RAG can scaffold scientific text comprehension and course assistance [10,11].

In sum, our motivation is that in high-expertise, high-risk knowledge communication, fluent and complete language alone cannot sustain trust. Only by embedding authoritative-source retrieval and quantitative semantic-alignment monitoring can we deliver a Q&A service that is usable, credible, and auditable. This is why we adopt RAG as the system backbone and dual-similarity as the quantitative basis for governance.

1.2. Objectives

This study aims to propose a refreshable and audit-oriented workflow for nuclear energy knowledge question answering: beyond merely using Retrieval-Augmented Generation (RAG) to compensate for large language models' (LLMs') intrinsic lag in timeliness, we separate offline indexing from online inference and bound knowledge to authoritative corpora (e.g., documents from the Nuclear Safety Commission of Taiwan), thereby promoting locality-aware, source-traceable responses. This design makes the questions “Did we retrieve the right evidence?” and “Is the generated answer semantically aligned with the retrieved text?” observable. Whereas prior educational systems emphasize tooling—e.g., knowledge graph-enhanced RAG tutors or document Q&A pipelines [10,12]—our focus is to operationalize governance through dual-similarity signals for retrieval-side alignment ($Q \rightarrow C$) and post-generation answer-to-context alignment monitoring ($A \rightarrow C$). We introduce a dual-similarity monitoring scheme spanning two measurement stages of the RAG pipeline: question-to-context similarity ($Q \rightarrow C$) is used for Top-k passage ranking and selection before generation, while answer-to-context similarity ($A \rightarrow C$) is computed after generation and reported as a post-generation semantic alignment signal relative to the retrieved context chunks. The terms “entry” and “exit” are used here as conceptual labels for these two measurement stages; the present experiments do not implement formal thresholds, automatic re-retrieval, or uncertainty-triggering logic. Instead, the two scores are recorded and analyzed as observable signals associated with potential off-topic drift or weak answer–context alignment.

We further design two contrasted conditions to assess robustness: an in-corpus setting where queries are supported by the knowledge base, and an out-of-corpus (weak-support) setting where overlap with the corpus is limited. In the former, the average $A \rightarrow C$ exceeds $Q \rightarrow C$ (e.g., 0.633 vs. 0.593), indicating that the generated answers are, on average, more semantically similar to the retrieved contexts than the original questions under this experimental setup. In the latter, both metrics degrade overall and the association between the two signals is weaker, which is consistent with greater dependence on corpus overlap and retrieval support. However, these signal-level observations should not be interpreted as a substitute for independent human or expert assessment of correctness and attribution.

In sum, the objective of this section is to (i) address LLM staleness and locality gaps via RAG, (ii) operationalize “findable and semantically aligned” answers through dual-similarity monitoring as measurable monitoring signals, and (iii) empirically characterize quality degradation when database support is absent, thereby establishing a reusable, governance-oriented technical foundation for education-oriented Q&A in nuclear energy and radiation protection.

2. Materials and Methods

2.1. System Architecture

The proposed nuclear-energy popular-science question-answering system adopts a Retrieval-Augmented Generation (RAG) architecture, as shown in Figure 1. Using LangChain [13] as the workflow orchestration framework, it integrates FAISS [14]-based vector retrieval (Retriever) and a large language model (ChatGPT 5.2 is used in this study) [15] as the two core modules to achieve generation grounded in trustworthy documents. The overall data flow is divided into two stages: offline construction and online inference.

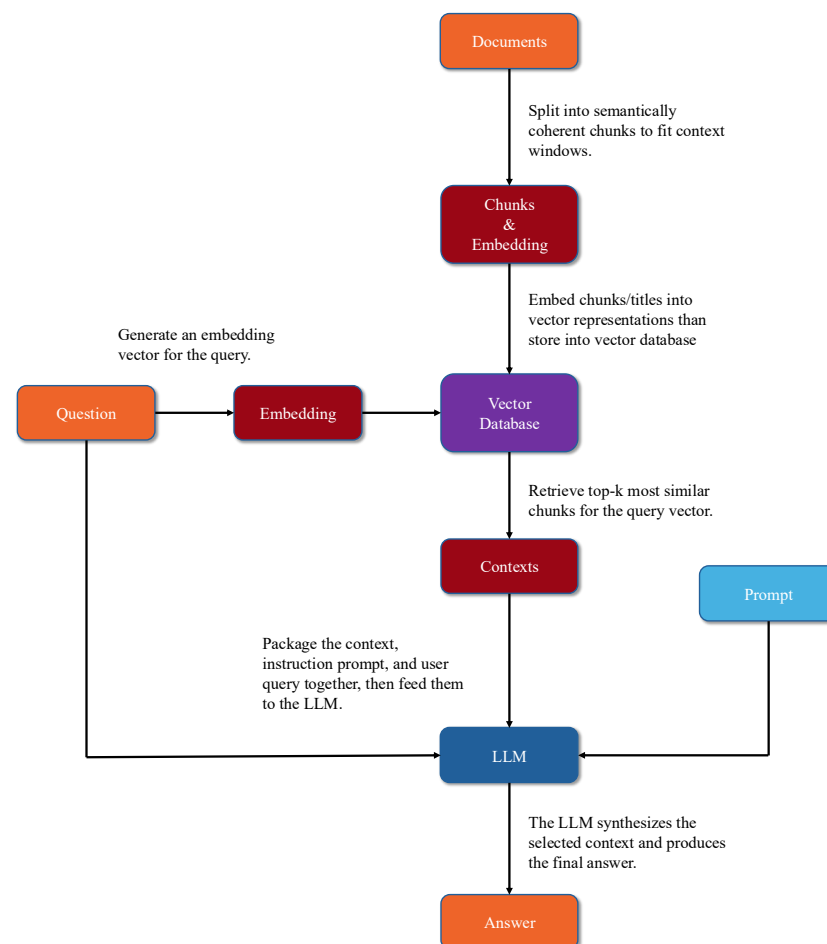


Figure 1. The proposed Retrieval-Augmented Generation (RAG) system architecture in this research.

2.1.1. Indexing Pipeline

The pipeline proceeds as follows: (1) document loading and chunking → (2) embedding (vectorization) → (3) indexing and document storage. In implementation, we first load the curated PDF corpus, then split the text into 500-character chunks with a 100-character overlap; next, we vectorize using OpenAIEmbeddings (text-embedding-3-large) and add the L2-normalized vectors to FAISS IndexFlatIP (inner-product index) to enable cosine-similarity retrieval. We simultaneously build an InMemoryDocstore and the index_to_docstore_id mapping, and finally persist them to the specified paths (index.faiss and docstore.pkl).

2.1.2. RAG Inference Pipeline

(1) Query embedding → (2) Top-k retrieval → (3) prompt assembly → (4) generation → (5) post hoc similarity assessment. Upon receiving a question, the system embeds the query

and retrieves the $k = 5$ most relevant chunks from FAISS. It then uses PromptTemplate to merge the chunks into the context and generates an answer via ChatOpenAI(model="gpt-4o", temperature=0). Finally, it performs dual similarity analyses to quantify the consistency between the question/answer and the evidence chunks, returning these scores together with the answer to provide a traceable response.

Design principles: prioritize “document-grounded evidence”; retrieval delimits the visible knowledge scope, while generation handles linguistic organization and synthesis; post hoc alignment monitoring may improve interpretability and support downstream auditing.

2.2. Data Sources and Pre-Processing

2.2.1. Data Sources

Data collection is the foundational groundwork for building a RAG system and directly determines the quality and reliability of subsequent generation. The corpus used in this study is drawn entirely from Taiwan’s Nuclear Safety Commission (NSC) [16], an agency under the Executive Yuan. The NSC was established to strengthen Taiwan’s regulatory framework for nuclear safety and radiation protection, in line with international trends and the recommendations of the International Atomic Energy Agency (IAEA) [17]. The database centers on Taiwan’s policy developments, scientific applications, and public communication in nuclear-related fields, forming a local knowledge base that is professional, traceable, and timely. Its coverage includes nuclear safety, radiation protection, medical radiation technology, food irradiation, radioactive-waste management, nuclear legislation, and energy transition. After curation, textual materials are consolidated into PDF files for downstream ingestion and use.

2.2.2. Loading and Chunking

After the data was collected, tables, covers, appendices, headers and footers were removed beforehand, and titles and bar numbers were retained. Text is extracted page by page with PyPDFLoader, and paragraphs are then split using RecursiveCharacterTextSplitter. Chunk length and overlap are measured in characters. We set 500/100 to balance information density and retrieval discriminability: overly long chunks reduce retrieval precision, whereas overly short chunks disrupt semantic coherence; a 100-character overlap mitigates cross-chunk fragmentation.

2.2.3. Semantic Normalization

After chunking, we compute embeddings with text-embedding-3-large. To make the inner product directly interpretable as cosine similarity, we apply L2 normalization [18] to each embedding vector $\mathbf{x} \in \mathbb{R}^d$:

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \quad (1)$$

This reduces similarity to $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} \in [-1, 1]$, enabling fast matching with IndexFlatIP.

2.2.4. FAISS Index

FAISS (Facebook AI similarity search) is an open-source vector search library developed by Meta AI for similarity search and nearest-neighbor matching in large-scale, high-dimensional data. It supports multiple index families (e.g., Flat, IVF, PQ, HNSW) and can leverage GPU acceleration to balance efficiency and accuracy. FAISS is widely used across search engines, recommender systems, natural language processing (RAG and semantic search), and computer vision (image retrieval and face recognition), making it a core tool for handling embeddings and similarity analysis.

The FAISS index is persistently stored as `index.faiss`, while the document store is saved as `docstore.pkl` [19]. These artifacts can be loaded directly in subsequent runs to avoid recomputing embeddings and rebuilding the index, thereby saving computation.

2.3. Retrieval Techniques and Generation Strategy (Retriever and Generator)

2.3.1. Vector Retriever (FAISS Retriever)

We use `faiss.IndexFlatIP(d)` to perform exact inner-product search over d -dimensional, L2-normalized vectors. Because $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = \cos \theta$, the returned similarity scores can be interpreted as cosine similarities. In the present study, this exact-search configuration is sufficient for the corpus size used in the experiments and is executed in a standard single-machine environment. The FAISS index is stored as `index.faiss`, while the associated document store is stored separately as `docstore.pkl`.

2.3.2. Generator (LLM via LangChain)

1. Model and temperature

We use `ChatOpenAI(model="gpt-4o", temperature=0)`. A temperature of 0 promotes reproducibility and rigor (reduced stochasticity). To constrain the response to retrieved material and reduce the risk of off-context generation, the `PromptTemplate` explicitly instructs: "Please answer as much as possible based on the documents provided below and keep your response within the scope of those documents; do not mention the documents in your answer; use Taiwan-specific wording and conventions." In this study, the RAG system is configured to generate user-facing responses in Traditional Chinese. This combination of "style constraint" and "source boundedness" supports both localization and traceability.

2. Prompt assembly

The retrieved Top- k chunks are concatenated in order as the context, then injected together with the question into the template and passed to the LLM. To control context length, we apply chunk-length limits and keyword weighting; for multi-topic, long-form queries, we design "query decomposition" to retrieve sub-queries separately.

2.4. Implementation Details and Reproducibility

All experiments were executed in Google Colaboratory (Colab) on a Linux runtime (Linux-6.6.105+, x86_64; glibc 2.35) using Python 3.12.12. To support reproducibility, we report the key software dependencies used in this study: `langchain` v1.2.8, `langchain-community` v0.4.1, `langchain-openai` v1.1.9, the OpenAI Python SDK (`openai`) v2.17.0, FAISS (`faiss-cpu`) v1.13.2, and `pypdf` v6.7.0; tokenization utilities were provided by `tiktoken` v0.12.0. For retrieval, we built a FAISS `IndexFlatIP` index and applied L2-normalization to embedding vectors so that inner product similarity is equivalent to cosine similarity.

2.5. Dual-Similarity Analysis and Post-Generation Alignment Monitoring

To enhance transparency and monitor semantic alignment with retrieved documents, we apply a dual-similarity analysis after generation [20].

Cosine similarity:

$$\cos_sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \Rightarrow \hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = \cos_sim(\mathbf{x}, \mathbf{y}) \quad (2)$$

From the above, once text embeddings are L2-normalized, they become unit vectors, so the similarity score can be obtained directly from their inner product [18].

2.5.1. Question-to-Context Similarity

During retrieval, we compute the similarity between the query vector $\hat{\mathbf{q}}$ and each document chunk $\hat{\mathbf{x}}_i$:

$$s_{QC}^{(i)} = \hat{\mathbf{q}} \cdot \hat{\mathbf{x}}_i \quad (3)$$

This metric evaluates whether the retrieved evidence is on target—i.e., whether the documents match the question—and is used for Top- k filtering and ranking (we set Top- $k = 5$ in this study).

2.5.2. Answer-to-Context Similarity

After generating the answer a , we embed it with the same model to obtain $\hat{\mathbf{a}}$ and compute

$$s_{AC}^{(i)} = \hat{\mathbf{a}} \cdot \hat{\mathbf{x}}_i \quad (4)$$

In the RAG workflow, before the final response is produced, a large language model performs consolidation and refinement; during this step, the answer may still drift from the intent of the question or exhibit hallucinations. By evaluating the answer-to-context similarity, we can flag potential issues: if the $s_{AC}^{(i)}$ scores for the leading context chunks are markedly low, this may indicate an increased risk that the answer has drifted semantically from the retrieved context [21].

2.5.3. Question Generation

The evaluation used 10 questions in total: five corpus-referenced questions and five general nuclear energy questions. These questions were generated using ChatGPT under fixed category-specific prompts in Traditional Chinese and were manually reviewed by the authors before use. The use of Traditional Chinese was intended to match the target public-facing question-answering setting in Taiwan. The full prompt set and the resulting question list are provided in the Supplementary Materials. Because the current study is exploratory and uses a small evaluation set, the results should be interpreted as a pilot analysis rather than a benchmark-scale evaluation.

1. Questions randomly generated with reference to the corpus:
 - (a) According to the “Ionizing Radiation Protection Act”, which radioactive materials or devices require a license application to the Nuclear Safety Commission (NSC)?
 - (b) After a nuclear incident occurs, within what time frame must the facility operator complete both the telephone and written notifications?
 - (c) What are the seven fundamental laws that currently govern nuclear and radiation safety regulation in Taiwan?
 - (d) On what date was the No. 2 unit of Taiwan’s Maanshan (Nuclear Power Plant 3) shut down, marking the formal step toward a “nuclear-free homeland”?
 - (e) What are the four main stages in the full decommissioning process of a nuclear power plant?
2. Randomly generated general nuclear energy questions:
 - (a) How do the energy sources of nuclear fission and nuclear fusion differ, and why is fusion regarded as a cleaner energy option?
 - (b) In a nuclear power plant, what is the role of control rods, and what materials are they typically made of?
 - (c) What is the difference between the radiation dose units sievert (Sv) and gray (Gy)?

- (d) What were the primary causes of, and key differences among, the Three Mile Island, Chernobyl, and Fukushima nuclear accidents?
- (e) In medical imaging (e.g., PET and CT), why are radioactive isotopes used, and how is radiation safety ensured?

2.6. Algorithmic Summary

The algorithm utilized in this study is summarized in Algorithm 1.

Algorithm 1: RAG with Dual-Similarity Monitoring

1. Input: query string q ; Output: answer a and $\{(s_{QC}^{(i)}, s_{AC}^{(i)})\}_{i=1}^k$.
 2. $\hat{q} \leftarrow \text{Normalize}(\text{Embed}(q))$.
 3. $\{x_i\}_{i=1}^k \leftarrow \text{FAISS.TopK}(\hat{q})$ to obtain the Top- k chunks.
 4. $\text{context} \leftarrow \text{Concat}(\{x_i\})$; inject (context, q) into the PromptTemplate.
 5. $a \leftarrow \text{LLM.Generate}(\text{context}, q; \text{temperature} = 0)$.
 6. $\hat{a} \leftarrow \text{Normalize}(\text{Embed}(a))$.
 7. For each x_i : keep the retrieval score $s_{QC}^{(i)}$ and compute $s_{AC}^{(i)} \leftarrow \hat{a} \cdot \hat{x}_i$.
-

3. Results

3.1. Question Similarity and Answer Similarity

Although both “question similarity” and “answer similarity” measure semantic proximity, they serve different stages and mitigate different risks. The former (Q–C) compares the user’s question with candidate evidence chunks during retrieval; it is used for ranking and sampling with the goal of “finding the right material.” A high Q–C indicates that the chunk is topically close to the question, but it does not guarantee that it actually contains facts sufficient to answer it. The latter (A–C) compares the model’s generated answer with the retrieved chunks after generation; it is used as a post-generation semantic alignment indicator. A high A–C suggests that the generated answer is semantically close to the retrieved contexts in embedding space.

The timing also differs: Q–C is computed before generation to decide which Top- k chunks are fed to the model; A–C is computed after generation to characterize semantic alignment between the generated answer and the retrieved contexts. In the present study, this score is reported for post hoc analysis only. At present, this mechanism is not linked to an automated re-retrieval trigger or uncertainty flag; incorporating these functions will be part of future work. In practice, both often use cosine similarity over text embeddings. If Q–C is high but A–C is low, retrieval likely found on-topic material, but the generation step drifted. If Q–C is low but A–C is high, the generated answer may still be semantically close to some retrieved chunks, although the initial retrieval may have missed the most relevant passages. High–high is ideal (easy to find; easy to answer). Low–low is a dual warning (both data and answer are off). Both signals then indicate a semantic misalignment.

$$\text{Question similarity} = \cos(f(Q), f(C_i)) \quad (5)$$

$$\text{Answer similarity} = \cos(f(A), f(C_i)) \quad (6)$$

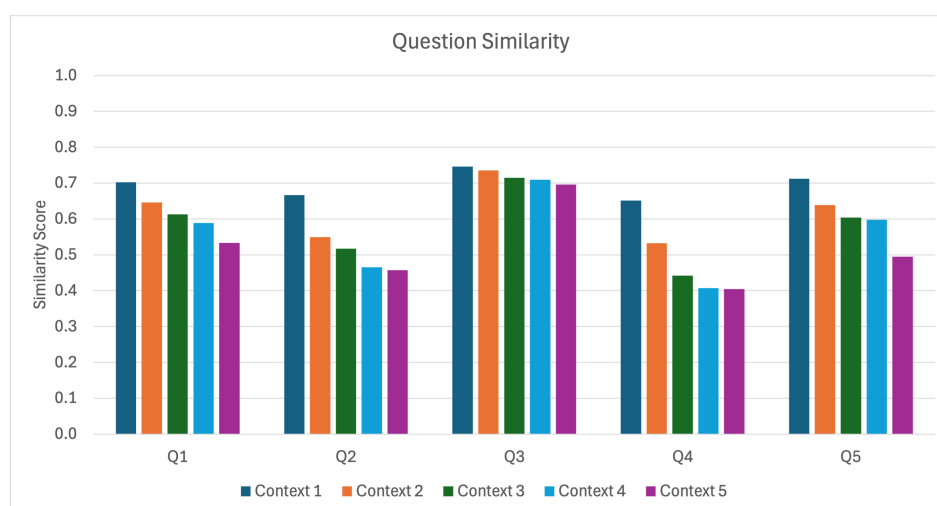
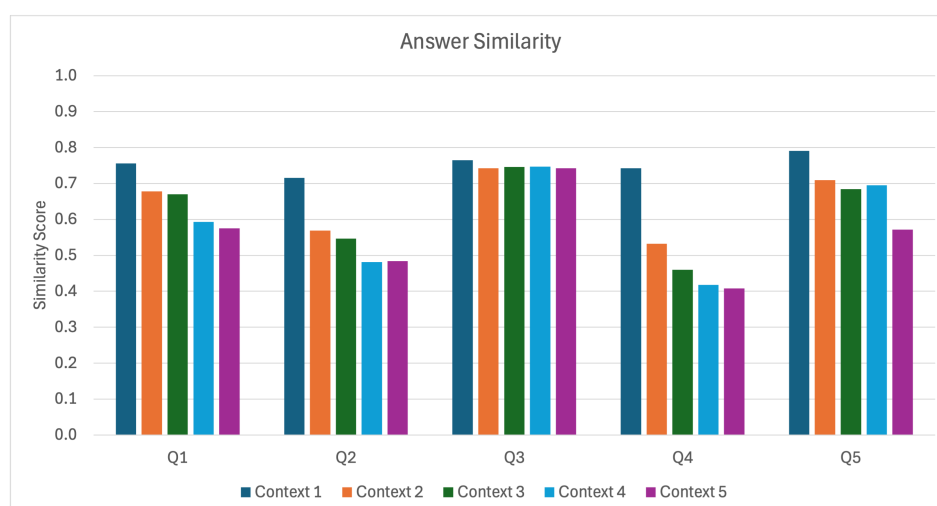
Here, $f(\cdot)$ denotes the text embedding model, and C_i is the i -th context chunk.

3.2. Questions Generated with Corpus References

The similarity analysis for questions generated with reference to the corpus is shown in Table 1. Next, the question similarity analysis (corpus-referenced) is given in Figure 2. Finally, the answer similarity analysis (corpus-referenced) is given in Figure 3.

Table 1. Similarity analysis for questions generated with reference to the corpus.

Item	Similarity	Questions Generated with Corpus References				
		Chunk 1	Chunk 2	Chunk 3	Chunk 4	Chunk 5
Q1	Question	0.702	0.646	0.613	0.589	0.533
	Answer	0.756	0.678	0.670	0.593	0.575
Q2	Question	0.666	0.549	0.517	0.465	0.457
	Answer	0.716	0.569	0.547	0.481	0.484
Q3	Question	0.746	0.735	0.715	0.709	0.696
	Answer	0.765	0.742	0.746	0.747	0.742
Q4	Question	0.651	0.532	0.442	0.407	0.404
	Answer	0.742	0.532	0.460	0.418	0.408
Q5	Question	0.712	0.639	0.604	0.598	0.495
	Answer	0.791	0.709	0.684	0.695	0.572

**Figure 2.** Question similarity analysis (corpus-referenced).**Figure 3.** Answer similarity analysis (corpus-referenced).

3.2.1. Overall Trends by Corpus Questions

- The overall averages are question similarity = 0.593 and answer similarity = 0.633, with the answer similarity being higher by 0.040 (a relative gain of about +6.7%, since $0.040/0.593 \approx 0.067$).

- In terms of overall association, the Pearson correlation across the 25 (chunk, score) pairs is approximately $r \approx 0.973$ (extremely high). This indicates a strong observed association between $Q \rightarrow C$ and $A \rightarrow C$ in this corpus-referenced setting: chunks that are more semantically similar to the question also tend to yield higher answer-to-context similarity. However, this association should not be interpreted as an independent demonstration that retrieval quality determines answer quality, because correctness and attribution were not separately evaluated by human or expert judgment.

3.2.2. Per-Question Observations by Corpus Questions

The comparison of question versus answer similarity for per-question observations by corpus questions is shown in Table 2.

Table 2. Comparison of question vs. answer similarity for per-question observations by corpus questions.

Question	Mean Question Similarity	Mean Answer Similarity	Difference	Relative Improvement
Q1	0.617	0.654	+0.037	+6.0%
Q2	0.531	0.559	+0.028	+5.3%
Q3	0.720	0.748	+0.028	+3.9%
Q4	0.487	0.512	+0.025	+5.1%
Q5	0.610	0.690	+0.080	+13.1%

- Question 3 is the most stable and highest overall (both metrics ≥ 0.72), indicating strong semantic overlap between the question and the corpus and consistently high alignment scores across the retrieved contexts.
- Question 4 is noticeably lower (question 0.487, answer 0.512), especially dropping to 0.404/0.408 by context 5; this suggests the key semantics of the question are hard to align within the primary contexts.
- Question 5 shows the largest answer-versus-question similarity gain (+0.081), indicating that the generated answer was more semantically aligned with the retrieved contexts than the original question in this case. We interpret this as an observed alignment pattern rather than as direct evidence of improved factual correctness.

3.2.3. Per-Context Observations by Corpus Questions

The comparison of question versus answer similarity for per-context observations by corpus questions is shown in Table 3.

Table 3. Comparison of question vs. answer similarity for per-context observations by corpus questions.

Context	Mean Question Similarity	Mean Answer Similarity	Difference
Context 1	0.695	0.754	+0.059
Context 2	0.620	0.646	+0.026
Context 3	0.578	0.621	+0.043
Context 4	0.554	0.587	+0.033
Context 5	0.517	0.556	+0.039

- Context 1 is the best across all questions (both metrics rank first), effectively serving as the primary or most central source.
- From Context 1 $\rightarrow$ Context 5 the scores decline monotonically (question similarity 0.695 $\rightarrow$ 0.517; answer similarity 0.754 $\rightarrow$ 0.556), indicating increasing semantic distance for later contexts. If retrieval is ranked by similarity, the current ordering

aligns with the score trend and is consistent with the retrieval ranking produced in this setting.

3.2.4. Answer vs. Question: Difference Analysis

- The largest differences occur in “Question 5 × Context 4” (+0.097) and “Question 4 × Context 1” (+0.091), indicating that even when the question–context semantic match is not very high, the generated answer is, in these cases, more semantically similar to the retrieved contexts than the original question, indicating an observed increase in alignment scores after generation.
- A zero difference is observed only in “Question 4 × Context 2” ($0.532 \rightarrow 0.532$), suggesting that, for this combination, generation did not materially increase the alignment score relative to the retrieved context.

3.3. Randomly Generated Nuclear Energy Questions

The similarity analysis for randomly generated nuclear energy questions is shown in Table 4. Next, the question similarity analysis (randomly generated) is given in Figure 4. Finally, the answer similarity analysis (randomly generated) is given in Figure 5.

Table 4. Similarity analysis for randomly generated nuclear energy questions.

Item	Similarity	Randomly Generated Nuclear Energy Questions				
		Context 1	Context 2	Context 3	Context 4	Context 5
Q1	Question	0.362	0.345	0.318	0.309	0.308
	Answer	0.460	0.435	0.403	0.392	0.411
Q2	Question	0.407	0.373	0.369	0.368	0.366
	Answer	0.413	0.381	0.365	0.360	0.387
Q3	Question	0.353	0.277	0.271	0.258	0.255
	Answer	0.422	0.348	0.344	0.317	0.314
Q4	Question	0.401	0.351	0.342	0.332	0.332
	Answer	0.508	0.388	0.433	0.405	0.411
Q5	Question	0.446	0.432	0.416	0.394	0.393
	Answer	0.463	0.453	0.454	0.419	0.397

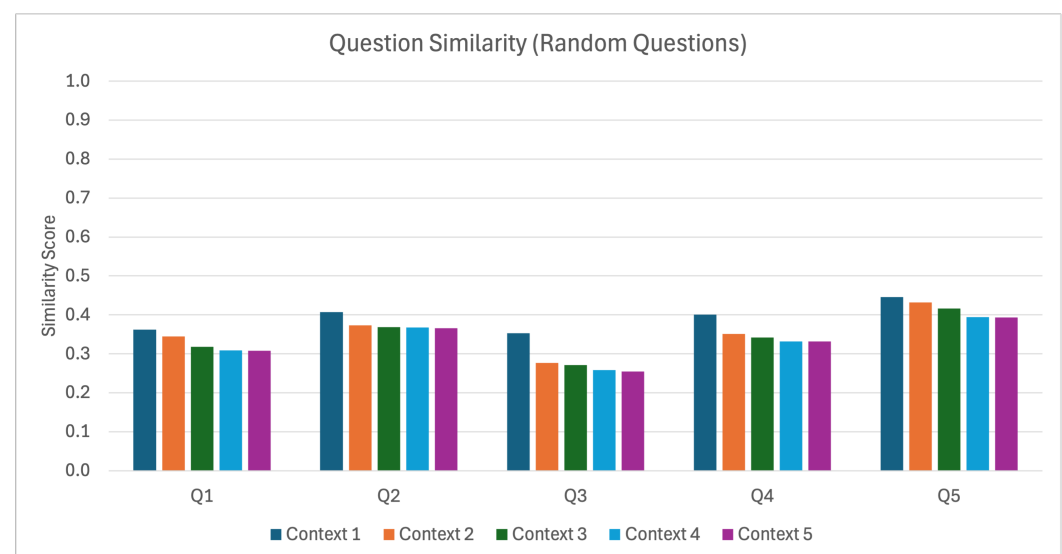


Figure 4. Question similarity analysis (randomly generated).

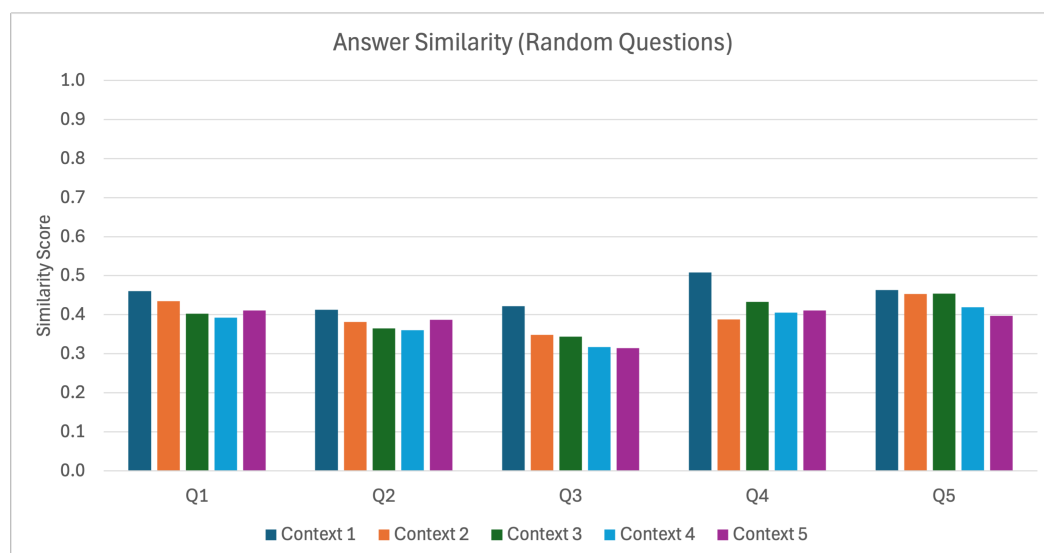


Figure 5. Answer similarity analysis (randomly generated).

3.3.1. Overall Trends by Random Questions

- Overall averages: question similarity = 0.351, answer similarity = 0.403; the answer is higher by +0.052 (relative improvement +14.8%, since $0.052/0.351 \approx 0.148$).
- Overall association: Pearson $r \approx 0.729$ over the 25 (context, score) pairs (moderate-to-high). The observed association between $Q \rightarrow C$ and $A \rightarrow C$ remains positive in the randomly generated question setting ($r \approx 0.729$), but it is notably weaker than in Table 1. This suggests a weaker consistency between retrieval-side and answer-side similarity signals under lower corpus overlap, rather than providing independent evidence of answer correctness.
- Distribution: question similarity ≥ 0.40 in 5/25 cases; answer similarity ≥ 0.40 in 14/25 (with ≥ 0.45 in 5/25). This suggests that, even under lower corpus overlap, the generated answers may still exhibit moderate semantic alignment with some retrieved contexts.

3.3.2. Per-Question Observations by Random Questions

The comparison of question versus answer similarity for per-question observations by random questions is shown in Table 5.

Table 5. Comparison of question vs. answer similarity for per-question observations by random questions.

Question	Mean Question Similarity	Mean Answer Similarity	Difference	Relative Improvement
Q1	0.328	0.420	+0.092	+28.0%
Q2	0.377	0.381	+0.005	+1.2%
Q3	0.283	0.349	+0.066	+23.4%
Q4	0.352	0.429	+0.077	+22.0%
Q5	0.416	0.437	+0.021	+5.0%

- Question 3 is the lowest on both metrics and is most sensitive across contexts (from Context 1 $\rightarrow$ 5: question -0.098 , answer -0.108).
- Question 4 has the highest mean answer similarity (0.429), and in “Q4 $\times$ Context 1”, it shows the largest gain in the entire table, +0.107 (0.401 $\rightarrow$ 0.508).

3.3.3. Per-Context Observations by Random Questions

The comparison of question versus answer similarity for per-context observations by random questions is given in Table 6.

Table 6. Comparison of question vs. answer similarity for per-context observations by random questions.

Context	Mean Question Similarity	Mean Answer Similarity	Difference
Context 1	0.394	0.453	+0.059
Context 2	0.356	0.401	+0.045
Context 3	0.343	0.400	+0.057
Context 4	0.332	0.379	+0.046
Context 5	0.331	0.384	+0.053

- Context 1 is the best landing spot across questions (both metrics are the row's highest for each question) and yields the largest average gain (+0.059).
- Question similarity decreases monotonically from Context 1 $\rightarrow$ 5; answer similarity generally declines as well, with a slight rebound at Context 5 ($0.384 > 0.379$).

4. Conclusions

This study evaluates consistency in the RAG pipeline—across retrieval and generation—using two question sets: (i) questions generated with reference to the corpus (Table 1), and (ii) randomly generated nuclear energy questions (Table 4). The evaluation metrics comprise “question similarity” (semantic similarity between the question and each context) and “answer similarity” (similarity between the model's answer and each context). The main conclusions are as follows:

1. $A \rightarrow C$ is on average higher than $Q \rightarrow C$ under the present setup.
In both tables, answer similarity exceeds question similarity: Table 1 averages 0.633 vs. 0.593 (+0.040, +6.7%), and Table 4 averages 0.403 vs. 0.351 (+0.052, +14.8%). Across both settings, $A \rightarrow C$ is on average higher than $Q \rightarrow C$, indicating that the generated answers were more semantically similar to the retrieved contexts than the original questions under this experimental setup. We interpret this as an observed post-generation alignment pattern, not as independent evidence of factual improvement.
2. $Q \rightarrow C$ and $A \rightarrow C$ show stronger association in the corpus-referenced setting.
A strong association between $Q \rightarrow C$ and $A \rightarrow C$ was observed in the corpus-referenced setting ($r \approx 0.973$), whereas a weaker but still positive association was found in the randomly generated setting ($r \approx 0.729$). These results suggest that retrieval-side and answer-side similarity signals tend to vary together, especially when the questions are closely aligned with the corpus. However, these correlations should not be interpreted as independent evidence that retrieval quality determines answer quality, because correctness and attribution were not separately evaluated.
Importantly, the reported $Q \rightarrow C$ – $A \rightarrow C$ correlations describe associations between two semantic-similarity signals within the same pipeline and should not be interpreted as a substitute for direct evaluation of answer correctness, completeness, or source attribution. These aspects would require independent human or expert assessment.
3. A clear “core-document” structure emerges.
In both tables, Context 1 consistently ranks near the top across questions (hosting many of the highest question/answer similarities), and scores decline monotonically from Context 1 to Context 5. This suggests semantic density is concentrated, and core contexts disproportionately influence final quality.

4. Overall finding: corpus-guided vs. random questions.
Compared with Table 1, both metrics drop notably in Table 4 (0.593 $\rightarrow$ 0.351 for question similarity; 0.633 $\rightarrow$ 0.403 for answer similarity). Compared with the corpus-referenced setting, both similarity signals are lower for the randomly generated questions. This pattern is consistent with reduced corpus overlap and weaker alignment between the retrieved contexts and the generated answers.

In summary, this study has successfully combined the RAG technique and dual-similarity analysis to promote nuclear energy knowledge. The application is not limited to nuclear energy knowledge. In fact, the research flow chat of this study can also be applied to many other fields of scientific knowledge. This is the contribution of our study.

5. Limitation and Future Work

A limitation of this study is that the evaluation questions were generated with the assistance of ChatGPT and the sample size remains small, which may not fully reflect real user information needs and may introduce distributional bias. In addition, Q $\rightarrow$ C and A $\rightarrow$ C are used here as semantic-alignment signals rather than direct measures of factual correctness or attribution, and no independent human/expert evaluation was conducted in the present experiments. The same embedding family was also used for both retrieval and post-generation semantic scoring, which may introduce methodological coupling between retrieval and evaluation. The current system is further limited to a single-pass RAG workflow, in which similarity scores are recorded for analysis but are not yet linked to threshold-based re-retrieval, abstention, or uncertainty control. Finally, the experiments were performed on a relatively limited Taiwan-specific corpus and a fixed retrieval configuration. Future work will therefore expand the evaluation set with human-authored questions, introduce human/expert judgment for correctness and attribution, test heterogeneous evaluation backbones to reduce methodological coupling, and extend the pipeline with uncertainty handling, re-retrieval, re-ranking, corpus expansion, and retrieval-configuration ablation studies.

Supplementary Materials: The following supporting information can be downloaded at https://drive.google.com/drive/folders/1QjuTdjK7zZjq5Z44HFS_U2jwyb5Z9G4X?usp=drive_link, accessed on 20 January 2026.

Author Contributions: Conceptualization, C.-H.C.; Methodology, C.-H.C.; Resources, C.-H.C.; System design, C.-H.C.; Software, C.-H.C.; Investigation, C.-H.C.; Formal analysis, C.-H.C.; Visualization, C.-H.C.; Writing—original draft, C.-H.C.; Supervision, K.-C.L.; Writing—review and editing, K.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science and Technology Council, Taiwan, under Grant NSTC 114-NU-E-006-001-NU.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Supplementary Materials provide question, retrieved contexts, generated answer, and similarity scores, enabling readers to inspect how each conclusion was derived. The research data of this study can be found at https://drive.google.com/drive/folders/1QjuTdjK7zZjq5Z44HFS_U2jwyb5Z9G4X?usp=drive_link, accessed on 20 January 2026.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.2) to generate a subset of the evaluation questions described in the Methods section and to assist with language-related editing (translation and proofreading) to improve clarity and fluency. The generated questions were manually reviewed and curated by the authors. The authors take full responsibility for the final content, validity, and interpretations reported in this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 42. [CrossRef]
2. Liu, Z.; Wang, P.; Song, X.; Zhang, X.; Jiang, B. Survey on Hallucinations in Large Language Models. *J. Softw.* **2025**, *36*, 1152–1185.
3. Khashabi, D.; Zhang, S.; Kummerfeld, J.K. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv* **2024**, arXiv:2401.11817. [CrossRef]
4. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020; pp. 9459–9474.
5. Jung, T.; Joe, I. An Intelligent Docent System with a Small Large Language Model (sLLM) Based on Retrieval-Augmented Generation (RAG). *Appl. Sci.* **2025**, *15*, 9398. [CrossRef]
6. Castellanos-Nieves, D.; García-Forte, L. Human-Centered AI for Migrant Integration Through LLM and RAG Optimization. *Appl. Sci.* **2025**, *15*, 325. [CrossRef]
7. Knollmeyer, S.; Caymazer, O.; Koval, L.; Akmal, M.U.; Asif, S.; Mathias, S.G.; Großmann, D. Benchmarking of Retrieval Augmented Generation: A Comprehensive Systematic Literature Review on Evaluation Dimensions, Evaluation Metrics and Datasets. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024)—Volume 3: KMIS*; SciTePress: Setúbal, Portugal, 2024; pp. 137–148.
8. Adlakha, V.; BehnamGhader, P.; Lu, X.H.; Meade, N.; Reddy, S. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 681–699. [CrossRef]
9. Fan, L.; Deng, K.; Liu, F. Educational impacts of generative artificial intelligence on learning and performance of engineering students in China. *arXiv* **2025**, arXiv:2505.09208v1. [CrossRef] [PubMed]
10. Dong, C.; Yuan, Y.; Chen, K.; Cheng, S.; Wen, C. How to Build an Adaptive AI Tutor for Any Course Using Knowledge Graph-Enhanced Retrieval-Augmented Generation (KG-RAG). *arXiv* **2023**, arXiv:2311.17696v7.
11. Thüs, D.; Malone, S.; Brünken, R. Exploring generative AI in higher education: A RAG system to support scientific text comprehension. *Front. Psychol.* **2024**, *15*, 1474892. [CrossRef] [PubMed]
12. Muludi, K.; Fitria, K.M.; Triloka, J.; Sutedi. Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 776–783. [CrossRef]
13. LangChain. Available online: <https://python.langchain.com/docs/introduction/> (accessed on 25 September 2025).
14. FAISS. Available online: <https://faiss.ai/> (accessed on 25 September 2025).
15. OpenAI. api-reference. Available online: <https://platform.openai.com/docs/api-reference/introduction> (accessed on 25 September 2025).
16. Nuclear Safety Commission (NSC), Taiwan. Available online: <https://www.nusc.gov.tw/> (accessed on 25 September 2025).
17. International Atomic Energy Agency (IAEA). Available online: <https://www.iaea.org/> (accessed on 25 September 2025).
18. Sohangir, S.; Wang, D. Improved sqrt-cosine similarity measurement. *J. Big Data* **2017**, *4*, 25. [CrossRef]
19. LangChain. FAISS Vector Store—API Reference. Available online: https://api.python.langchain.com/en/latest/vectorstores/langchain_community.vectorstores.faiss.FAISS.html (accessed on 25 September 2025).
20. Turney, P.D.; Pantel, P. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.* **2010**, *37*, 141–188. [CrossRef]
21. Wu, D.; Gu, J.-C.; Yin, F.; Peng, N.; Chang, K.-W. Synchronous Faithfulness Monitoring for Trustworthy Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 9390–9406. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.