International Journal of Engineering and Technology Innovation, vol. 15, no. 4, 2025, pp. 476-487

Knowledge Representation Strategies for Reducing Hallucinations in Retrieval-Augmented Domain-Specific Question Answering

Cheng-Hsiu Li*

Department of Information Management, National Taitung Junior College, Taitung, Taiwan, ROC

Received 18 September 2025; received in revised form 17 October 2025; accepted 20 October 2025

DOI: https://doi.org/10.46604/ijeti.2025.15708

Abstract

To address the limitations of hallucinated responses in large language models (LLMs), an artificial intelligence (AI) chatbot featuring a retrieval-augmented generation system is designed to assist with subject-based certification instruction. Focusing on the Level B certification curriculum for computer hardware repair as an example, this study develops six distinct knowledge base structures (Type0–Type5) and integrates them into two open-source 7B-parameter LLMs (LLaMA2 and Qwen2) with a custom-built question and answer system. Response accuracy to 10 standardized questions is evaluated by domain experts. Knowledge structure significantly affects performance, with the enriched Type5 base yielding the highest accuracy (Qwen2: 98 points; LLaMA2: 73 points). Statistical tests confirm significant improvements with knowledge base enhancement across knowledge types and between models. These findings highlight the critical role of knowledge representation and LLM selection in domain-specific AI applications, proffering practical design guidelines for intelligent teaching assistants in technical education.

Keywords: retrieval-augmented generation, knowledge representation, large language models, hallucination reduction, technical education

1. Introduction

The importance of Skill certification has been accentuated by advances in industry and increasing sophistication. In this regard, companies are increasingly adopting comprehensive certification systems [1]. Vocational certificates function as both quality assurance mechanisms and as instruments for credential verification in the job market. The effective integration of certification into technical and vocational education facilitates alignment between education and industry, strengthens communication between supply-side and demand-side operators, and promotes seamless transitions from school to the workplace. These improvements create a win–win–win scenario for industries, educational institutions, and students by increasing students' employability and competitiveness [2].

According to Article 10 of Taiwan's Regulations for Technician Skills Certification and Issuance, to obtain certification, a technician candidate must pass both subject-based and practical tests at the same occupational level (Workforce Development Agency, Ministry of Labor, 2021). The subject-based test contains items drawn from a predefined bank, and the practical test is tailored to the technical characteristics of each occupation. Study guides and reference materials for these tests are publicly available to support candidates' preparation efforts [3]. Technical and vocational schools incorporate training for relevant certification into their curricula. Additionally, regional vocational training institutions offer both pre-employment and in-service training programs [4]. Statistics from the Skills Certification Center indicate that the average pass rates are 12%, 38%, and 67% for Levels A–C, respectively [5]. These rates demonstrate a clear correlation between the certification level and pass rate. To improve overall success in skills certification, both subject-based and practical performance must be strengthened.

* Corresponding author. E-mail address: chsiu.lee@gmail.com

Large language models (LLMs) are foundational to conversational agents capable of responding to a diverse range of user queries by referencing sources. However, the inherent characteristics of LLMs introduce unpredictability in their responses. Additionally, given that LLMs are trained on static datasets, their knowledge is limited to the information available up to a specific cutoff date, which limits access to the most recent information. Retrieval-augmented generation (RAG) is a technique designed to optimize LLM output by allowing them to retrieve relevant content from external, authoritative knowledge bases before generating a response. LLMs are trained on vast corpora and apply billions of parameters to perform tasks, such as question answering, language translation, and sentence completion. The RAG framework strengthens these capabilities by grounding responses in domain-specific or organizational knowledge without the need for model retraining. Based on these advantages, RAG is a cost-effective approach to improving LLM outputs while maintaining relevance, accuracy, and utility across use cases [6-7].

RAG has been widely adopted across domains to improve response accuracy, factual consistency, and information traceability in LLM outputs. Yang et al. [8] demonstrated that the incorporation of external knowledge sources into LLMs, such as clinical guidelines, improved the accuracy, contextual relevance, and reliability of medical responses. Similarly, Magesh et al. [9] reported that RAG reduced hallucinated case citations and strengthened source traceability when integrated with legal research tools. In medical applications, Bora and Cuayáhuitl [10] combined RAG with fine-tuned models for medical chatbot development, reporting significant improvements in response generation and multiple-choice question answering. These examples demonstrate the adaptability of RAG across knowledge-intensive and regulation-heavy domains where factual grounding is critical.

Several studies have highlighted the benefits of structured knowledge representation for increasing the retrieval precision and generation fidelity of RAG. Wang et al. [11] proposed the K-RagRec framework, which integrates knowledge graphs into LLM-based recommendations through semantic indexing, popularity-selective retrieval, and subgraph re-ranking strategies to enhance both accuracy and efficiency. Wang et al. [12] developed a mineral resource knowledge graph by integrating geological entity recognition and semantic relations, employing template matching and Cypher-based queries for retrieval, and delivering lightweight question answering through subgraph and abstract outputs. Knollmeyer et al. [13] introduced the GraphRAG framework, which leverages knowledge graphs derived from document structures with keyword-based semantic linking to refine retrieval strategies, thereby improving context relevance and multi-hop reasoning performance. These works indicate that structured data, especially in the form of knowledge graphs, can serve as a robust foundation for semantically meaningful retrieval and effective response generation in RAG systems.

Numerous scholars have explored how prompt design and system modularization affect RAG performance. For instance, Bora and Cuayáhuitl [10] demonstrated that few-shot prompting enhanced answer selection accuracy, while the adoption of Modular RAG architectures improved flexibility and contextual relevance in medical chatbot systems. Nie et al. [14] enhanced generation quality in their medical QA system through customized structured prompts and demonstrated a clear modular system design with a five-layer architecture. Similarly, Woesle et al. [15] reported that the combination of prompt engineering and modular strategies with knowledge-rich retrieval can increase the consistency and factuality of RAG outputs. These studies underscore the central roles of architecture and prompt design as controllable variables for system improvement.

Hallucination, or the production of factually incorrect or unsupported content by generative artificial intelligence (AI) systems, remains a critical challenge for LLMs. Multiple studies have identified RAG as a key solution to this concern. Zhang and Zhang [16] grounded LLM generation in retrieved evidence, which significantly reduced unsupported outputs, particularly when retrieval design and reranking mechanisms were optimized. Similarly, Woesle et al. [15] categorize hallucination mitigation strategies into four architectural types and highlight that fully integrated LLM-DMS architectures are the most effective yet significantly underrepresented. In a similar vein, Abdelghafour et al. [17] categorize hallucination mitigation

approaches into four main strategies, such as RAG, knowledge augmentation, human feedback training, and controlled text generation, and each furnishes distinct advantages for improving factual consistency in LLMs. Collectively, these findings establish RAG as a foundational tool for hallucination suppression, although its success is contingent upon careful system design.

Despite its promise, RAG faces several unresolved challenges. Wang et al. [11] note that current approaches continue to face issues such as retrieval noise, limited exploitation of structural information, and constraints in knowledge representation, while further exploration is needed in scaling to larger models, diversifying knowledge sources, and designing more advanced retrieval policies. Similarly, Woesle et al. [15] identify several unresolved challenges and methodological gaps in hallucination mitigation for LLMs, including the low prevalence of fully integrated architectures, the underrepresentation of domains such as education and autonomous systems, and the dominance of heuristic or single-stage approaches—highlighting the need for more sophisticated and cross-domain system designs. Zhang and Zhang [16] indicated that the absence of standardized retrieval strategies or knowledge representation formats produces inconsistent results across implementations. Although RAG is widely recognized as a method to improve LLM accuracy, no unified framework exists for selecting the optimal combination of retriever, data type, and prompt format.

In summary, numerous studies have demonstrated the potential of RAG systems to increase the accuracy, factual consistency, and reasoning capabilities of LLMs across diverse application domains. This improvement is particularly evident in healthcare [8], legal services [9], and medical applications [10, 14], where RAG can effectively reduce hallucinated outputs and facilitate interpretability by grounding responses in external knowledge. Furthermore, scholars have explored the integration of structured data, such as knowledge graphs [11-13], into RAG systems and the roles of prompt engineering and retrieval strategies in RAG effectiveness [15-16]. These studies highlight key design considerations for improving RAG performance. Despite these advances, prior research remains largely domain-specific and lacks systematic comparisons of knowledge representation formats, retriever configurations, and structured response generation techniques.

Due to the sheer volume of information processed by LLMs, they may struggle to accurately identify the most relevant content for a given query, often resulting in hallucinated outputs, which are plausible but incorrect or unverifiable responses. To address this problem, in this study, hallucination mitigation efforts were applied to the subject domain of the Level B certification for computer hardware fabrication. Specifically, RAG was integrated with LLMs to minimize hallucination and develop an AI-powered teaching assistant chatbot for subject-based technical certification preparation, alleviating the instructional workload of educators. Importantly, the accuracy of AI teaching assistant responses to learners' subject-related questions is crucial. Therefore, the RAG knowledge base structure yielding the most effective results when LLMs were used to answer standardized subject-based questions was investigated. Additionally, whether different LLMs exhibit performance variation when integrated with the same RAG structure was assessed. The goal of this study was to identify the optimal combination of an LLM and a RAG knowledge base for the development of AI teaching assistants tailored to technical certification subjects.

2. Materials and Methods

This section describes the design and implementation of the proposed AI teaching system integrating an RAG framework with locally deployed LLMs. The system was developed using the Ollama framework, the Python programming language, and the Visual Studio Code environment. It combines vector storage based on Facebook AI Similarity Search (FAISS) and user feedback in the format of JavaScript Object Notation Lines (JSONL) to support reinforcement learning from human feedback (RLHF). In the following subsections, Section 2.1 details the system environment, operational workflow, core functionalities design, knowledge base management, and application of reinforcement learning concepts, while Section 2.2 presents the system's implementation and interface for both standard and RAG-enhanced operational modes.

2.1. System design

(1) System environment

The system employed in this study was developed with the Ollama framework, Python programming language, and Visual Studio Code editor.

(2) Operational workflow

System operation was designed according to the following four phases. During the user interaction phase, users first select a preferred local LLM on the Ollama platform and may upload one or more PDF documents to serve as external knowledge sources. During the knowledge base construction phase, PyPDFLoader is applied to parse and segment the uploaded documents into semantically coherent chunks with RecursiveCharacterTextSplitter. Then, OllamaEmbeddings is used to embed these chunks, which are subsequently indexed into a FAISS-based vector database, supporting persistence and caching. If available, prior user feedback in JSONL format is also vectorized and incorporated into a separate feedback retrieval database. During the question-answering and retrieval phase, users submit timestamped queries. If a knowledge base has been uploaded, the system harnesses ConversationalRetrievalChain and EnsembleRetriever to integrate document and feedback sources into an RAG workflow; otherwise, it defaults to a general ConversationChain.

During the feedback and learning optimization phase, the system displays the generated response and labeled source segments. Users may submit feedback or corrections, which are logged and vectorized for integration into future sessions, creating an RLHF loop that progressively improves retrieval quality and response accuracy. The operational workflow of the system is illustrated in Fig. 1.

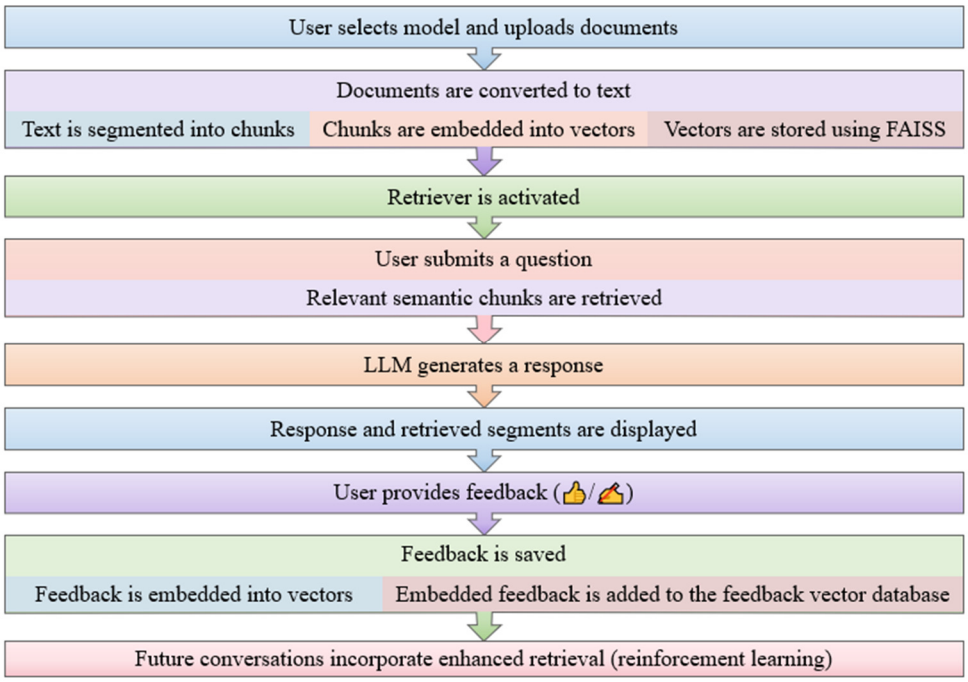


Fig. 1 System operational workflow

(3) Core functionalities design

The system includes five core components designed to support robust domain-specific question responses. First, users can select a locally deployed LLM on the Ollama interface, which generates responses in traditional Chinese text based on user input. Second, the RAG mode enables semantic retrieval from a knowledge base of uploaded PDF documents. FAISS segments and vectorizes these documents, enabling the system to retrieve contextually relevant content and generate grounded answers. Third, a user feedback mechanism establishes a reinforcement loop, allowing users to mark helpful responses as a "good answer" or provide corrections. These feedback entries are stored in JSONL format, embedded into vectors, and

integrated into the retrieval process to improve future outputs. Fourth, the system employs an ensemble retriever to combine results from the document database and feedback vectors, assigning greater weight (60%) to user-derived feedback to increase personalization and precision. Finally, conversation buffer memory is employed for dynamic memory management, preserving interaction history to support both retrieval-based and general conversational modes.

(4) Knowledge base management

The system incorporates a streamlined management process for knowledge bases to optimize efficiency and support scalable retrieval. After PDF documents are uploaded, the system automatically extracts, segments, and converts the textual content into vector embeddings. Each document is hashed to generate a unique identifier, enabling the system to detect duplicates and avoid redundant processing. If the uploaded file remains unchanged, the system loads the previously cached FAISS vector database, substantially reducing computation time. Additionally, historical user feedback stored in JSONL format is automatically loaded during system initialization to construct or update the feedback vector index. Thus, the system continuously leverages prior corrections and evaluations to improve retrieval quality. The pseudocode is as follows:

```
When PDF documents are uploaded:
    Extract and segment textual content
    Convert text into vector embeddings
    Generate unique hash for each document
    If document hash already exists:
        Load cached FAISS vector database
    Else:
        Build new FAISS vector database and save it
End if
On system initialization:
    Load historical feedback from JSONL file
    Construct or update feedback vector index
End initialization
```

(5) Application of reinforcement learning concepts

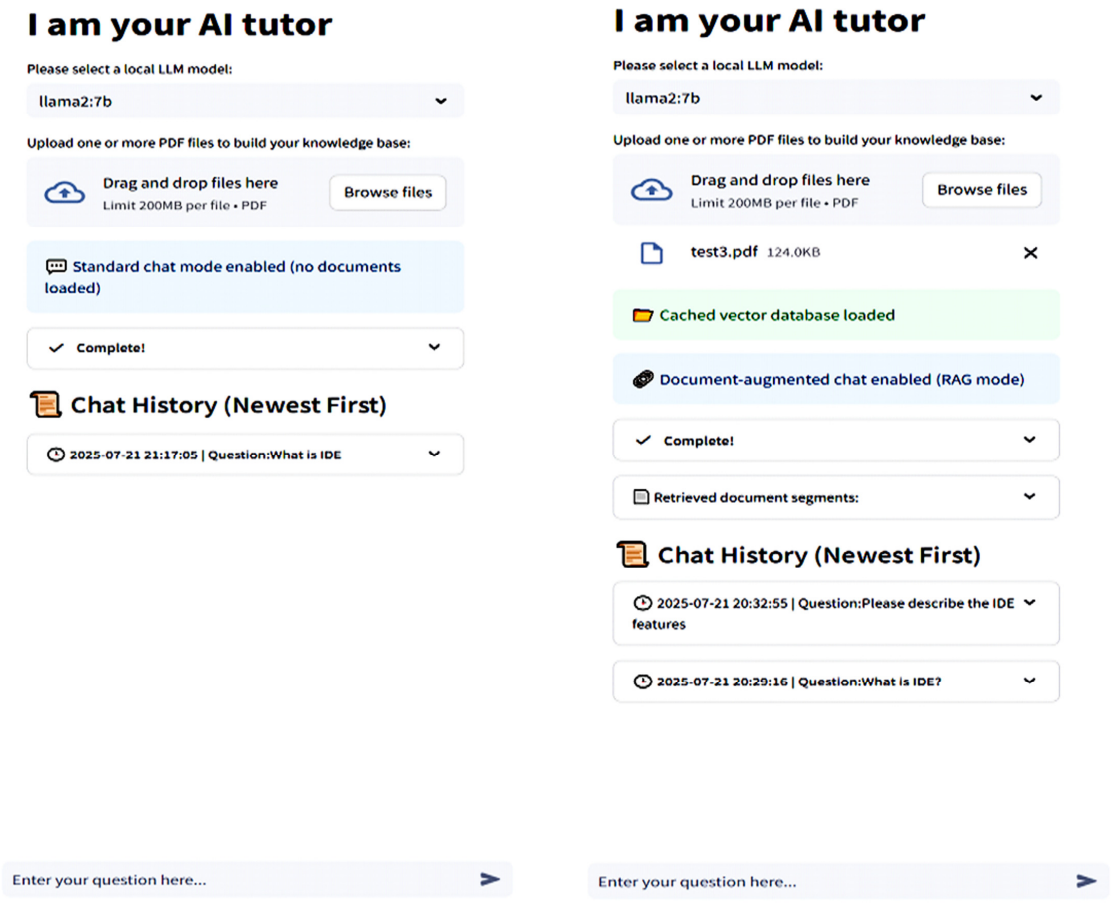
The system incorporates user feedback into semantic embedding and retrieval processes. The embedding model in Ollama is used to vectorize feedback, which is then integrated into the knowledge base for foundational RLHF implementation. The pseudocode is as follows:

```
Function Load_Feedback_Vector_DB(embeddings)
    If feedback log file does not exist then
        Return None
    End if

    Initialize empty list feedback_docs
    Open feedback log file
    For each line in the file do
        Parse the line as JSON data
        Combine "question" and "answer" into a single document
        Append the document to feedback_docs
    End for

    If feedback_docs is empty then
        Return None
    End if

    Try
        Build a FAISS vector database from feedback_docs with given embeddings
        Display success message with number of documents loaded
        Return feedback_db
    Catch exception
        Display error message
        Return None
    End try
End function
```

2.2. System implementation

Building upon the system design described in Section 2.1, this subsection presents the practical implementation and user interface of the proposed AI teaching assistant. The system integrates both a standard LLM response mode and an RAG mode, allowing users to experience how semantic retrieval enhances the accuracy and contextual relevance of generated answers. Fig. 2 displays the interface for the two operational modes of the AI teaching system: the standard LLM response mode without RAG integration, as shown in Fig. 2(a), and the RAG-enhanced LLM response mode incorporating semantic retrieval, as shown in Fig. 2(b).

In the standard LLM mode, the system directly generates responses from the local language model based solely on user input. This mode serves as a baseline for evaluating the impact of knowledge augmentation and operates without any external document retrieval. In the RAG-enhanced mode, the system connects to the FAISS-based vector database to retrieve semantically relevant document chunks from the uploaded knowledge base. Retrieved text segments are then combined with the user query to form a context-aware prompt, which is passed to the selected LLM for grounded answer generation. The generated response is presented together with the retrieved source text, allowing users to trace the origin of the answer and assess its accuracy. Additionally, a feedback panel enables users to rate each response as "good" or "needs correction". The feedback is stored in JSONL format and vectorized for use in future sessions, progressively improving system accuracy through RLHF.



(a) AI teaching assistant without RAG mode (b) AI teaching assistant with RAG mode

Fig. 2 AI teaching assistant interface

3. Results and Discussion

This study aims to analyze the effectiveness of LLM responses to subject-based certification questions when integrated with various RAG knowledge base structures. Two open-source language models, LLaMA2 and Qwen2, were randomly selected as the target LLMs for evaluation. Initially, five distinct knowledge base structures (Type0–Type4) were designed.

LLaMA2-7B was used as the primary model, generating responses through an RAG architecture without reinforcement learning. These responses were scored according to the standards of the Level B Computer Hardware Fabrication subject test. Subsequently, the highest-scoring knowledge base structure was further enhanced and defined as Type5. The performance of LLaMA2 and Qwen2 was evaluated with the Type5 knowledge base applied to gauge the effect of knowledge augmentation.

The experimental objectives were as follows:

(1) To compare the performance of different RAG knowledge base structures

A. The same LLM model was used for all comparisons.
B. Each structure answered the same set of 10 questions.
C. The response accuracy (correct answer rate) was recorded for each structure.

(2) To compare the response accuracy of different LLMs

A. Both LLMs had the same parameter size (7B).
B. Each model answered the same 10 questions.
C. Response accuracy was compared between models.

3.1. Knowledge base design

(1) Type0 (no RAG): The LLM answers questions directly without external knowledge support.

(2) Type1 (original test format): Questions are presented in a multiple-choice format resembling the certification test layout. Regarding LLM, the knowledge is unstructured and minimally processed.

Examples
(3) Which of the following is not a type of computer interface? ① IDE ② SCSI ③ VLSI ④ SATA (1) The BASIC language belongs to: ① High-level language ② Low-level language ③ Assembly language ④ Machine language

(3) Type2 (QA format): Each question is paired with its corresponding standard answer in a concise and retrieval-friendly format.

Examples
Q: Which of the following is not a type of computer interface? A: VLSI Q: The BASIC language belongs to which category? A: High-level language.

(4) Type3 (explanatory format): Each item is supplemented with a detailed explanation that provides logical reasoning steps and background information to support understanding.

Examples
IDE, SCSI, and SATA are all types of storage device interfaces, whereas VLSI (or very-large-scale integration) refers to an integrated circuit design technology. BASIC is an easy-to-learn high-level programming language that is commonly used for introductory teaching purposes.

(5) Type4 (QA + explanation format): Corresponding explanatory content is provided for the correct answer, proffering a balance between accuracy and interpretability.

Examples
Q: Which of the following is not a type of computer interface? A: IDE, SCSI, and SATA are all storage device interfaces. VLSI (or very-large-scale integration) is a circuit design technology. Q: The BASIC language belongs to which category? A: BASIC is an easy-to-learn high-level programming language that is commonly used for introductory teaching purposes.

(6) Type5 (enhanced format): This format is to incorporate additional clarifications based on Type2, error corrections, and formula derivations to render structured knowledge for improved comprehension and reasoning accuracy.

Examples
Q: Which of the following is not a type of computer interface? A: VLSI. Explanation: VLSI (or very-large-scale integration) is a circuit design technology rather than a hardware interface. Q: The BASIC language belongs to which category? A: High-level language. Explanation: BASIC, short for Beginner's All-purpose Symbolic Instruction Code, is designed for simplicity and is commonly used for instructional purposes.

3.2. *Experimental design and response performance analysis*

Ten subject-based questions were developed based on the Level B certification test for computer hardware fabrication, covering logic operations, components, commands, error classification, physical quantities, and instruction usage. For each knowledge base type, content was processed through the RAG system and input into the LLM to generate responses. Each response was subsequently evaluated by domain experts on a per-question basis, with a maximum score of 10 points per item.

Response accuracy was evaluated on a 100-point scale, with 10 standardized questions rated by domain experts (10 points per item). Table 1 presents item-wise and total scores for each knowledge base type. Total scores were distributed as follows: Type0 (no RAG), 48 points; Type1, 65 points; Type2, 66 points; Type3, 49 points; and Type4, 63 points. The integration of Type5—the enhanced knowledge base format—with the LLaMA2 model resulted in a total score of 73 points. Type5 integration with Qwen2 yielded the highest total score of 98 points. Based on these results, Type2 yielded the best performance among the original five knowledge base types (total score: 66). Furthermore, the enhanced Type5 format, which added clarifications and structured information to the Type2 format, considerably improved performance and demonstrated strong compatibility and knowledge application when integrated with Qwen2. Fig. 3 presents a line chart that compares the performance across different knowledge base types.

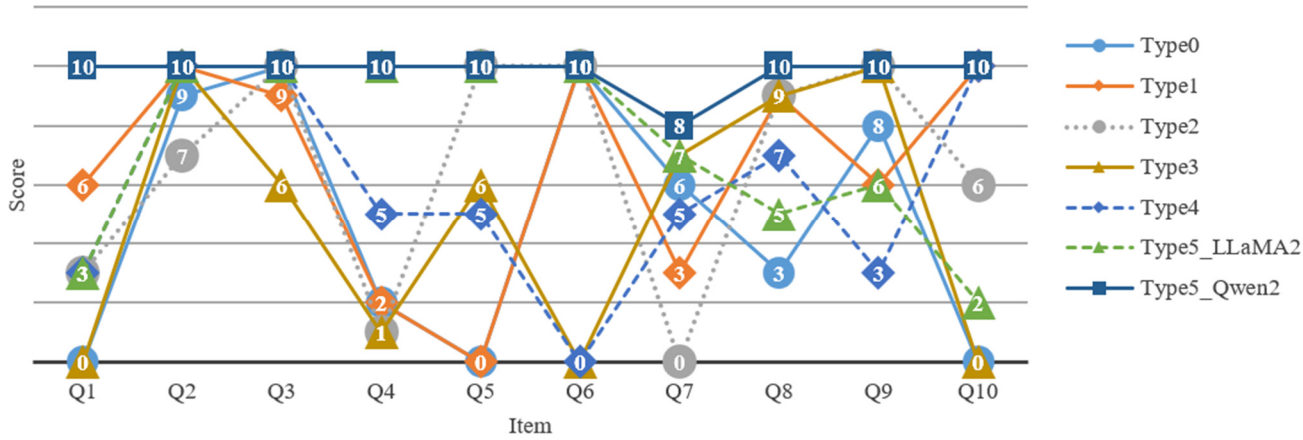


Fig. 3 Line chart illustrating the comparison among different knowledge base types

Table 1 Item-wise and total score comparison across knowledge base types

Item	Type0	Type1	Type2	Type3	Type4	Type5_LLaMA2	Type5_Qwen2
Q1	0	6	3	0	3	3	10
Q2	9	10	7	10	10	10	10
Q3	10	9	10	6	10	10	10
Q4	2	2	1	1	5	10	10
Q5	0	0	10	6	5	10	10
Q6	10	10	10	0	0	10	10
Q7	6	3	0	7	5	7	8
Q8	3	9	9	9	7	5	10
Q9	8	6	10	10	3	6	10
Q10	0	10	6	0	10	2	10
Total score	48	65	66	49	58	73	98

3.3. Statistical analysis

To evaluate the effects of different knowledge base types on LLM-generated responses, scores for the five initial knowledge base structures (Type0–Type4) across 10 questions related to the Level B certification test for computer hardware fabrication were subjected to statistical analysis.

A one-way analysis of variance was conducted of aggregated item scores to determine differences in overall performance between the five knowledge base types. An F-value of 0.102 and p-value of 0.981 were obtained, both of which were substantially higher than the significance threshold of $\alpha = 0.05$. This finding indicates that the differences in response accuracy across the five knowledge base types were nonsignificant. In other words, the accuracy of LLM answers under the original test format (Type0), QA-pair format (Type2), and formats incorporating explanatory content (Type3, Type4) was broadly comparable within the given parameters.

Tukey's honestly significant difference test was performed for post-hoc pairwise comparisons, likewise indicating that no score differences between the initial five knowledge base types reached statistical significance. Thus, no single knowledge base structure was significantly superior to the others in terms of statistical performance. Descriptive statistics suggested that Type2 (total score = 66) and Type1 (total score = 65) performed slightly better than the other formats.

Subsequently, to assess whether the performance of the enhanced knowledge base format (Type5) differed significantly between models, scores from the sixth test set generated by both LLaMA2 and Qwen2 were subjected to a paired samples t-test. The calculation was performed as follows:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{1}$$

where $\bar{d}$ is the mean of the paired differences, s_d is the standard deviation of these differences, and n is the number of paired samples.

An independent samples t-test was performed to compare the performance of the responses generated under the Type2 and enhanced Type5 formats. The test statistic was calculated as follows:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{2}$$

where $\overline{X_1}$ and $\overline{X_2}$ denote the group means, ${s_1}^2$ and ${s_2}^2$ represent the sample variances, and n_1 and n_2 are the sample sizes for each group. These formats were compared to determine whether the enhanced knowledge base yielded significantly different outcomes depending on the model employed.

A two-tailed hypothesis test was employed for all t-test analyses; $\alpha = 0.05$ was the threshold for significance. The t-test results, summarized in Table 2, were interpreted as follows:

(1) LLaMA2 vs. Qwen2 (both with Type5 knowledge base applied): Performance differed significantly between both models ($p = 0.036$), indicating that the model choice meaningfully affected outcomes when the enhanced knowledge base was employed.

(2) LLaMA2 (Type5) vs. LLaMA2 (Type2): No significant difference was observed between both formats ($p = 0.667$), suggesting that the advantages of Type5 over Type2 did not substantially affect the performance of the LLaMA2 model.

(3) Qwen2 (Type5) vs. Qwen2 (Type2): Performance differed significantly between both formats ($p = 0.021$), demonstrating that the enhanced Type5 knowledge base significantly improved the response accuracy of the Qwen2 model.

Table 2 Statistical analysis results (*t*-tests)

Comparison group	*t*-value	*p*-value
LLaMA2 vs. Qwen2 (Type5)	–2.466	0.0358*
LLaMA2 (Type5) vs. Type2_QA	0.437	0.6670
Qwen2 (Type5) vs. Type2_QA	2.530	0.0210*

*significant at $p < 0.05$

3.4. Discussion

This study demonstrated that the knowledge base structure substantially affects the response accuracy of LLMs. A systematic comparison of knowledge base types, ranging from no external knowledge support (Type0) to structurally enhanced knowledge bases (Type5), revealed the effect of knowledge base selection on the accuracy of LLM-provided answers to subject-based questions. Performance differences between the LLaMA2 and Qwen2 models under the enhanced knowledge format were further examined. The key findings are summarized as follows:

(1) Knowledge base design affected answer accuracy

The experimental results revealed a notable performance gap between the zero knowledge condition (Type0) and structured knowledge base formats (Type1 through Type5). Both the original test (Type1) and QA pair (Type2) formats improved the accuracy of model responses. Among the initial five knowledge base types, Type2 achieved the highest score (66 points), likely due to its concise and structured answering format, demonstrating the value of clarity and structure in knowledge representation.

(2) Knowledge base enhancement substantially improved model performance

The Type2 format was enhanced with supplementary explanations, formula derivations, and clarification of commonly confused concepts to form the enhanced Type5 knowledge base, increasing the total score for LLaMA2 from 66 to 73 points. Additionally, under this condition, Qwen2 achieved an impressive 98 points. These findings indicate that structured augmentation of knowledge base content can effectively reduce misinterpretation and answer deviation in LLMs.

(3) LLMs incorporating the same knowledge base differed significantly in performance

When the same Type5 enhanced knowledge base was integrated into both models, Qwen2 significantly outperformed LLaMA2 ($p = 0.0358$), suggesting that Qwen2 possesses superior language reasoning, structural comprehension, and knowledge extraction capabilities.

(4) The Type5 format provided significant benefits over Type2 for Qwen2

Qwen2 model performance improved significantly when integrated with the Type5 knowledge base compared with the Type2 knowledge base ($p = 0.0210$), indicating that Qwen2 was able to effectively absorb and apply additional information embedded in the enhanced knowledge base. This result reflects its strong knowledge integration and use capacities.

Overall, the experimental results indicate that structured and semantically oriented knowledge base construction is especially instrumental in improving the accuracy of LLM responses to subject-based questions. Model-specific differences in the effects of knowledge base selection were observed, with Qwen2 achieving near-perfect performance and outperforming LLaMA2 in overall accuracy when combined with the enhanced knowledge base. Although the knowledge base structure played a decisive role in enhancing response accuracy, the results also highlight the intrinsic capability differences among LLMs. When both models were integrated with the same Type5 knowledge base, Qwen2 achieved a substantially higher score (98 points) than LLaMA2 (73 points). This indicates that the underlying model architecture, parameter optimization, and pretraining data quality of LLMs critically influence their reasoning and inference ability, even when external knowledge is uniformly provided. Thus, the performance gain cannot solely be attributed to the RAG structure itself; the LLM's internal semantic understanding and contextual alignment mechanisms equally contributed to improved factual accuracy.

4. Conclusions

The large-scale knowledge embedded in LLMs complicates efforts to ensure that these models retrieve information most relevant to a given query, often causing hallucination phenomena in generated responses. The present study addresses this concern through the application of a RAG system to LLM responses regarding the Level B curriculum for computer hardware repair. A RAG-based LLM chatbot was developed to enhance answer accuracy and reliability in technical certification learning while reducing teachers' instructional workload. This study presents a replicable methodology for evaluating LLMs in domain-specific educational settings, contributing novel insights into how knowledge base structure and model choice interact to affect response accuracy. The findings provide a foundation for the future development of intelligent teaching systems to support vocational certification and other structured learning domains. The study makes the following key conclusions.

(1) Knowledge base design significantly affects answer accuracy: The Type2 (QA format) knowledge base outperformed other nonaugmented designs, likely because its structured question and answer pairs enabled efficient semantic retrieval and accurate LLM responses.

(2) Knowledge base enhancement improves LLM performance: The Type5 knowledge base, an enhanced version of the Type2 format with added clarifications and formula explanations, yielded total accuracy scores of 73 and 98 for the LLaMA2 and Qwen2 models, respectively, highlighting the effectiveness of semantic enrichment.

(3) LLM selection plays a decisive role in determining outcome quality: Even under the same RAG and knowledge structure, intrinsic model capacity and architecture significantly affected response precision. Qwen2's superior contextual reasoning and inference capabilities highlight that the LLM itself contributes substantially to performance enhancement, beyond the effects of knowledge base design alone.

(4) Statistical evidence supports the benefits of knowledge enhancement: Qwen2 performance improved significantly when integrated with the Type5 format compared with the Type2 format ($p = 0.021$), demonstrating its capacity to effectively absorb and apply enriched semantic input.

(5) RAG integration mitigates hallucination and improves domain alignment: Compared with the zero-knowledge baseline (Type0), all RAG-enabled designs demonstrated improved response accuracy, validating the effectiveness of RAG-based approaches in aligning LLM outputs with trusted domain knowledge.

Acknowledgments

This work was supported by the Ministry of Education Teaching Practice Research Program (Taiwan) under grant no. MOE-114-TPREE-0222-002Y1.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Workforce Development Agency, Skill Evaluation Center, "Agency Vision," https://www.wdasec.gov.tw/cp.aspx?n=786A0E9AF937C82C, accessed in 2024.
[2] L. S. Li, H. T. Li and S. C. Chen, "A Review and Outlook on the Certification System in Technical and Vocational Education," Bimonthly Journal of Educational Data and Research, vol. 93, pp. 31-52, 2010. (In Chinese)
[3] Workforce Development Agency, Skill Evaluation Center, "Test Reference Materials," https://techbank.wdasec.gov.tw/owInform/TestReferData.aspx, accessed in 2024.
[4] Workforce Development Agency, Skill Evaluation Center, "Regulations for Technician Skill Certification and Issuance,"

https://www.wdasec.gov.tw/News_Content.aspx?n=4D833E26864BB926&sms=1BE761BDBCE7C913&s=91A7C6BD6D520027, accessed in 2024.
[5] Workforce Development Agency, Skill Evaluation Center, "Statistics on Technician Skill Certification from 2011 to 2023: Registrations, Test Attendance, and Pass Rates," https://www.wdasec.gov.tw/News.aspx?n=5941D5DCC3DD7DDA&sms=CA0630966F34DA45, accessed in 2024.
[6] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, "A Survey on RAG with LLMs," Procedia Computer Science, vol. 246, pp. 3781-3790, 2024.
[7] G. Budakoglu and H. Emekci, "Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning, and Their Synergistic Fusion for Enhanced Performance," IEEE Access, vol. 13, pp. 30936-30951, 2025.
[8] R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. S. Bitterman, et al., "Retrieval-Augmented Generation for Generative Artificial Intelligence in Health Care," NPJ Health Systems, vol. 2, article no. 2, 2025.
[9] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," Journal of Empirical Legal Studies, vol. 22, no. 2, pp. 216-242, 2025.
[10] A. Bora and H. Cuayáhuitl, "Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications," Machine Learning and Knowledge Extraction, vol. 6, no. 4, pp. 2355-2374, 2024.
[11] S. Wang, W. Fan, Y. Feng, S. Lin, X. Ma, S. Wang, et al., "Knowledge Graph Retrieval-Augmented Generation for LLM-based Recommendation," Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 27152-27168, 2025.
[12] M. Wang, C. Wang, J. Chen, B. Wang, W. Wang, X. Ma, et al., "A Lightweight Knowledge Graph-Driven Question Answering System for Field-Based Mineral Resource Survey," Applied Computing and Geosciences, vol. 27, article no. 100268, 2025.
[13] S. Knollmeyer, O. Caymazer, and D. Grossmann, "Document GraphRAG: Knowledge Graph Enhanced Retrieval Augmented Generation for Document Question Answering Within the Manufacturing Domain," Electronics, vol. 14, no. 11, article no. 2102, 2025.
[14] J. Nie, X. Liu, Y. Tu, G. Zhao, D. Wu, and Y. Tang, "Design and Implementation of a Medical Question Answering System Based on Retrieval-Augmented Generation," International Journal of Biology and Life Sciences, vol. 11, no. 3, pp. 1-6, 2025.
[15] C. Woesle, L. Fischer-Brandies, and R. Buettner, "A Systematic Literature Review of Hallucinations in Large Language Models," IEEE Access, vol. 13, pp. 148231-148253, 2025.
[16] W. Zhang and J. Zhang, "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," Mathematics, vol. 13, no. 5, article no. 856, 2025.
[17] M. A. M. Abdelghafour, M. Mabrouk, and Z. Taha, "Hallucination Mitigation Techniques in Large Language Models," International Journal of Intelligent Computing and Information Sciences, vol. 24, no. 4, pp. 73-81, 2024.