A Multiple LLM RAG Platform
Using Answer Templates and Matching-Based Integration

Athiruj Poositaporn[12], Jaehyun Kim[3], Hanmin Jung[12]
[1]University of Science and Technology, [2]Korea Institute of Science and Technology Information,
[3]Sungkyunkwan University
athiruj@kisti.re.kr, danyou7im@gmail.com, jhm@kisti.re.kr

정답 템플릿과 매칭 기반 통합을 이용한 다중 LLM RAG 플랫폼

Athiruj Poositaporn[12], 김재현[3], 정한민[12]
[1]과학기술연합대학원대학교, [2]한국과학기술정보연구원, [3]성균관대학교

Abstract

Retrieval-Augmented Generation (RAG) often depends on a single LLM, which can lead to hallucinations and limited domain-specific knowledge. This study proposes a multiple LLM RAG platform that uses an answer template with matching-based integration to combine outputs from different models. Our experiment showed that using meaning-based matching for word similarity metric produced better outputs than word-based matching. Future work will enable users to choose the answer template and the number and type of LLMs, extend evaluation with large-scale datasets and performance metrics such as F1, ROUGE, and BLEU, apply the approach to diverse query types, and incorporate algorithm-based result verification.

I. Introduction

RAG has emerged as a valuable framework for generating responses grounded in relevant knowledge [1–2]. However, most implementations rely on a single LLM, which could lead to limited domain knowledge and hallucinations [3]. Additionally, recent work by S. Chakraborty et al. (2025) has explored multiple LLMs in a RAG architecture, but it focuses on evaluation rather than methods for integrating outputs into a unified response [4]. To address this, we propose a multiple LLM RAG platform that applies an answer template and matching-based integration. In this study, we experiment on two integration methods: word-based matching (baseline) and meaning-based matching (proposed).

II. Multiple LLM RAG Platform

Figure 3 illustrates how information flows from the user prompt through to the final integrated results. The architecture is composed of four sequential modules:

1. LLM-based Question Answering: A user's query is combined with a 5W1H answer template and passed into the $LLM_{A/B}$ producing $Answer_{A/B}$ [5]. In this study, we use GPT-4o and Gemini 2.5 Flash for the LLM A and B. Each answer contains a set of words for each 5W1H component, as shown in Figure 1.

Q8: What did CERN do in 2012?

Answer$_A$ (GPT-4o):
Who = ["CERN", "Physicists", "ATLAS and CMS teams"]
What = ["Higgs boson", "Particle discovery", "Scientific breakthrough"]
When = ["2012", "July 4", "Early 21st century"]
Where = ["Large Hadron Collider", "Geneva", "Switzerland"]
Why = ["Standard Model validation", "Mass origin", "Fundamental physics"]
How = ["Proton collisions", "High-energy experiments", "Data analysis"]

Answer$_B$ (Gemini 2.5 Flash):
Who = ["CERN scientists", "ATLAS", "CMS"]
What = ["Higgs boson", "Discovery announcement"]
When = ["July 4, 2012"]
Where = ["Geneva", "Switzerland"]
Why = ["Standard Model", "Particle physics"]
How = ["Large Hadron Collider", "Particle collisions"]

Figure 1. An Example of The Answer$_{A/B}$

2. Classification-based Template Filling: $Answer_{A/B}$ are then transformed into six binary word-similarity matrices (one for each 5W1H component) using two methods: (i) word-based matching (baseline) by comparing exact word matches and (ii) meaning-based matching (proposed) by considering semantic similarities using Qwen3-Embedding-0.6B. A cosine of 0.7 is chosen as previously established as effective in semantic similarity evaluation [6]. The $Binary\ Matrices_{P/Q}$ are then classified to determine a relationship type on each 5W1H component using the following conditions:

$$K(M) = \begin{cases} Equal\ Sets & a = m \wedge b = n, \\ Disjoint\ Sets & c = 0 \wedge (m + n > 0), \\ Proper\ Subset & (a = m \wedge b < n) \vee (b = n \wedge a < m), \\ Overlapping\ Sets & otherwise \end{cases} \quad (1)$$

Where $M \in \{0,1\}^{m \times n}$ is the binary matrix of two word sets $p = \{w_1, w_2, \ldots, w_m\}$ *and* $q = \{w_1, w_2, \ldots, w_n\}$ from a given 5W1H component. m,n are sizes of the p *and* q. $c = \sum_{ij} M_{ij}$ is the total count of matches. $a = \sum_i 1\left(\sum_j M_{ij} > 0\right)$ is the number of elements in p matched at least one element in q, while the $b = \sum_j 1\left(\sum_i M_{ij} > 0\right)$ is the reverse. Thus, $k(M)$ classifies the relationship between p *and* q as one of four types: Equal Sets, Disjoint Sets, Proper Subset, and Overlapping Sets. Each relationship type is associated with an integration rule that specifies how the word sets p *and* q are merged across the 5W1H components, as shown in Figure 2. The outputs of this module are the $Fill\text{-}in\ Template_{P/Q}$.

Disjoint sets: Answer the question based on the intersection elements: $\{w_1, \ldots, w_n\}$. Also, give the answer with distinct aspects using the elements: $\{w_1, \ldots, w_n\}$ and another answer using the elements: $\{w_1, \ldots, w_n\}$.
Proper subset: Answer the question based the intersection elements: $\{w_1, \ldots, w_n\}$. Also, give the answer with distinct aspect on the relative complement elements: $\{w_1, \ldots, w_n\}$.
Equal sets: Answer the question based on the elements: $\{w_1, \ldots, w_n\}$.
Overlapping sets: Answer the question based on the intersection elements: $\{w_1, \ldots, w_n\}$. Also, give the answer with distinct aspects using the elements: $\{w_1, \ldots, w_n\}$ and another answer using the elements: $\{w_1, \ldots, w_n\}$.

Figure 2. Classification Relationship Types

3. Template-based Prompt Integration: The user query, the $Fill\text{-}in\ Template_{P/Q}$, and the system prompt are merged to form the $Integrated\ Prompt_{P/Q}$.
4. Integrated Result Generation: A LLMC (Sonnet-4) generate the Integrated ResultP/Q and return to the user interface.

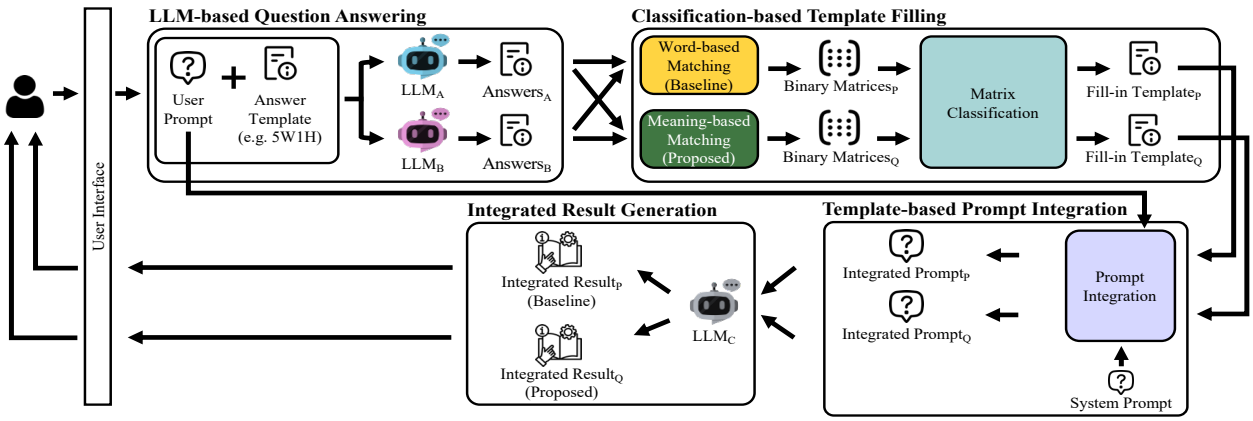

Figure 3. Multiple LLM RAG Platform Architecture

Table 1. Classification Results from Fill-in Template$_{P/Q}$ by Each Question and 5W1H Component

Question	Q1		Q2		Q3		Q4		Q5		Q6		Q7		Q8		Q9		Q10	
Method	WM	MM	WM	MM	WM	MM	WM	MM	WM	MM	WM	MM	WM	MM	WM	MM	WM	MM	WM	MM
Who	D	P	O	E	D	P	P	P	D	O	P	E	D	E	D	P	D	P	O	P
What	D	P	O	O	D	O	D	O	D	O	D	D	D	P	O	P	O	P	O	O
When	D	P	D	E	P	P	D	P	D	P	E	E	D	P	D	P	E	E	D	P
Where	P	E	D	O	D	P	D	D	O	E	D	P	P	P	P	P	D	E	D	P
Why	O	O	D	D	O	O	D	O	D	O	D	O	D	O	D	P	D	O	O	P
How	D	P	O	O	D	E	D	D	D	O	D	D	D	E	D	O	D	D	D	O

Word-based Matching (WM) | Meaning-based Matching (MM) | Equal Sets (E) | Proper Subset (P) | Overlapping Sets (O) | Disjoint Sets (D)

Table 2. A Generated Result Example of Word-based and Meaning-based Matching Methods

Q8: What did CERN do in 2012?
Word-based Matching: In 2012, **CERN, along with Physicists and the ATLAS and CMS teams, and separately CERN scientists with ATLAS and CMS**, achieved a major milestone with the Higgs boson—a Particle discovery regarded as a Scientific breakthrough and marked by a Discovery announcement. **The key moment occurred on July 4 in the Early 21st century, with July 4, 2012 as the exact date. The work took place in Geneva, Switzerland, specifically at the Large Hadron Collider. It mattered for Standard Model validation, explaining Mass origin, and advancing Fundamental physics, as well as addressing crucial questions in the Standard Model and Particle physics. The result was achieved through Proton collisions, High-energy experiments, and Data analysis, powered by the Large Hadron Collider enabling Particle collisions.**
Meaning-based Matching: In the Early 21st century, specifically on July 4, 2012, CERN, together with Physicists from the ATLAS and CMS teams — with a distinct contribution from CMS — announced a Scientific breakthrough: the Particle discovery of the Higgs boson. **This occurred in Geneva, Switzerland, at the Large Hadron Collider, where Proton collisions and other High-energy experiments were conducted. To achieve Standard Model validation and advance Fundamental physics, particularly clarifying Mass origin, the teams relied on extensive Data analysis of the resulting Particle collisions.**

III. Integrated Results Comparison between Word-based and Meaning-based Matching Methods

In the experiment, we asked 10 questions to the platform and the results of in Table 1 showed that word-based matching (WM) produced 41 disjoint sets, 2 equal sets, 6 proper subsets, and 11 overlapping sets. In contrast, meaning-based matching (MM) produced 6 disjoint sets, 11 equal sets, 25 proper subsets, and 18 overlapping sets. The dominance of disjoint sets in WM forced the model to responses long and wordy, while MM's higher rate of equal sets, proper subsets, and overlapping sets reduced redundancy and produced shorter, natural answers.

As an evident shown in Table 2. The WM output included repeated details, creating a wordy and less natural response. In contrast, the MM output was more concise, combining overlapping meanings into smoother text while still covering the same key information.

IV. Conclusion

This study proposed a multiple LLM RAG platform that combines multiple models using answer templates and matching-based integration. The experiments showed that meaning-based matching gave better results than word-based matching by producing shorter and clearer answers.

For future work, we will allow users to choose different answer templates and select how many and which LLMs to include. We also plan to conduct large-scale experiments on diverse datasets and evaluate the platform using performance metrics such as F1, ROUGE, and BLEU. Furthermore, we aim to apply the approach to a wider range of query types to test its adaptability across various scenarios. Finally, an algorithm-based verification step for the integrated result will be added, since the current process still requires manual checking.

Acknowledgment

This work was supported by UST Young Scientist+ Research Program 2024 through the University of Science and Technology. (No. 2024YS12)

References

[1] A. Poositaporn and H. Jung, "On RAG Framework Combined with Query Decomposition for Enhancing Client Engagement," *in Proceedings of KSII Spring Conference*, 2025.

[2] A. Poositaporn, H. Jung, and D. Lee, "A Study of a SQL-Generating RAG System for Improving Client engagement," *in Proceedings of The 17th International Conference on Future Information & Communication Engineering*, 2025.

[3] P. Feldman, J. R. Foulds, and S. Pan, "RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots," *arXiv preprint arXiv:2403.01193*, 2024.

[4] S. Chakraborty et al., "A scalable framework for evaluating multiple language models through cross-domain generation and hallucination detection," *Scientific Reports*, 2025.

[5] Y. Cao et al., "5W1H extraction with large language models," *in Proceedings of Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1–8, 2024.

[6] T. J. Cann et al., "Using semantic similarity and text embedding to measure the social media echo of strategic communications," *arXiv preprint arXiv:2303.16694*, 2023.