From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process

Jaewoong Kim
Department of Applied Data Science
Sungkyunkwan University
Seoul, Republic of Korea
jwoongkim11@g.skku.edu

Minseok Hur
Department of Immersive Media Engineering/Convergence Program for Social Innovation
Sungkyunkwan University
Seoul, Republic of Korea
alexhur3535@skku.edu

Moohong Min
Department of Computer Education/Convergence Program for Social Innovation
Sungkyunkwan University
Seoul, Republic of Korea
iceo@skku.edu

Abstract

Regulatory compliance in the pharmaceutical industry involves navigating complex and voluminous guidelines, often requiring significant amounts of human resources. Recent advancements in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) methods provide promising enhancements to data processing and knowledge management, potentially easing these burdens. However, despite these advancements, conventional Retrieval-Augmented Generation (RAG) methods fall short in this domain due to inherent structural problems. To address these challenges, we introduce the Question and Answer Retrieval Augmented Generation (QA-RAG) framework. This framework enhances the conventional RAG framework. It integrates a dual-track retrieval mechanism tailored to the specific and dynamic nature of pharmaceutical regulations. It utilizes not only the original query but also the answers generated by a fine-tuned LLM, thus providing a more robust foundation for document retrieval. Our experiments demonstrate that QA-RAG outperforms conventional methods in various evaluation metrics including precision, recall, and F1-score. These results underscore QA-RAG's capability to enhance both the accuracy and efficiency of regulatory compliance processes in the pharmaceutical industry. This paper details the structure and efficacy of QA-RAG, emphasizing its potential to revolutionize the regulatory compliance process in the pharmaceutical industry and beyond.

Keywords

Retrieval-Augmented Generation (RAG), Fine-Tuning Large Language Models (LLMs), Information Retrieval Effectiveness, Pharmaceutical Regulatory Compliance

ACM Reference Format:
Jaewoong Kim, Minseok Hur, and Moohong Min. 2025. From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process. In *The 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25), March 31-April 4, 2025, Catania, Italy.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3672608.3707749

1 INTRODUCTION

Recent advancements in Generative AI have significantly enhanced its capabilities in various industries including the pharmaceutical industry being one of the notable areas of focus. Navigating the complex and extensive guidelines provided by agencies such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) is often a daunting and time consuming task for industry players as the sheer volume of guidelines combined with their intricate details can make it challenging for companies to find and apply relevant information quickly. Compliance effort alone can consume up to 25% of a medium or large pharmaceutical site's operational budget [3], necessitating a more efficient method for navigating and interpreting regulatory guidelines.

While large language models (LLMs) can contribute to solving the problem by leveraging the ability to understand and generate texts, they can struggle to access or generate the precise requirements of pharmaceutical-specific regulations. Furthermore, while the retrieval-augmented generation (RAG) framework utilizes the innate knowledge of LLMs and fetches additional information from external sources to generate responses, they may fall short in the regulatory domain as described in Section 3, indicating significant room for improvement. This study introduces the **Question and Answer Retrieval Augmented Generation (QA-RAG)**, a framework that incorporates a dual retrieval mechanism of leveraging both the query and a contextually enriched hypothetical answer generated by a fine-tuned LLM.

2 PROPOSAL

2.1 Previous Works

Conventional RAG frameworks employ a single query to retrieve relevant documents using similarity search. However, relying heavily on specific query phrases can exclude relevant documents, which becomes more prominent with domain-specialized content. To overcome this issue, solutions, including multiquery retrieval [1, 7] and HyDE [5], have been proposed. Multiquery retrieval generates multiple queries with different perspectives using LLMs, and HyDE leverages hypothetical documents generated in response to the query using an instruction-following model to generate a text snippet, which is then used in similarity search for document retrieval. However, mutiquery is unable to capture a wide range of information due to the narrow scope of the user's query, and HyDE often produces very incomplete hypothetical answers in highly specialized domains due to the use of a general LLM.

SAC '25, March 31-April 4, 2025, Catania, Italy

ACM ISBN 979-8-4007-0629-5/25/03
https://doi.org/10.1145/3672608.3707749

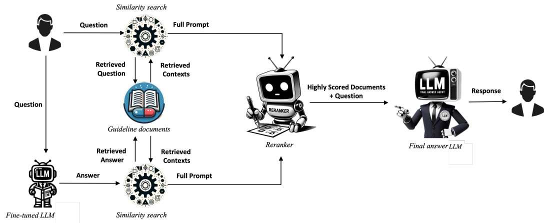

Figure 1: Overall Architecture of the QA-RAG Framework

2.2 Proposed Methodology

To address the limitations of previous works, we leverage a dual-track approach that utilizes both the user’s query and the tailored response generated by the fine-tuned LLM for enhanced accuracy and diversity of document retrieval. For each collected guideline document D_i, we divided them into chunks of 10,000 with an overlap of 2,000 characters between them, which can be denoted as $D_{i,j}$, where j represents the sequence number of each chunk for each document i. Then, we embedded the documents using the LLM-Embedder [9] and employed the BGE reranker [8].

$$S_{i,j} = \text{Rerank}(Q, D_{i,j}) \quad \text{where } j \in \{1, 2, \ldots, N\} \tag{1}$$

$$D = \{D_{i,j} \mid S_{i,j} \text{ is among the top scores}\} \tag{2}$$

$S_{i,j}$ represents the relevance score assigned to the j-th chunk of the i-th document by the reranker, in relation to the query Q. The highly ranked documents constitute the final document set D, which is used in the final stage of response generation. The generation of the final answer is done using a Final Answer LLM incorporated with a sophisticated few-shot prompting technique [6].

3 RESULTS AND DISCUSSION

3.1 Experiments

We focus on the pharmaceutical domain by collecting 1,263 FDA and 141 ICH (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use) for a total of 1,404 guidelines regarding the pharmaceutical industry. We extract the content of the documents using Nougat [2], a transformer-based OCR (Optical Character Recognition) tool, and divide them into chunks as mentioned previously to obtain a holistic view and to minimize information loss. We then fine-tune ChatGPT 3.5-Turbo and Llama3-8B with 1,681 frequently asked questions and answers sets collected from the FDA website by dividing 85% for *training*, 10% for *validation* and 5% for *testing* to select the optimal fine-tuned LLM. As shown in Figure 2, we train the models for over 3 epochs and select ChatGPT 3.5-Turbo, which achieved the best scores of precision (0.579), recall (0.589), and F1 (0.578). We employ ChatGPT 3.5-Turbo as the Final Answer LLM.

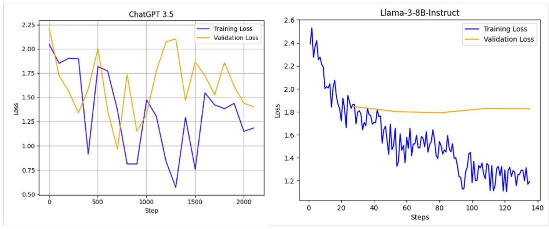

Figure 2: Training and Validation Loss Over Steps

Table 1: Evaluation of Context Retrieval.

Retrieval Method (# of Retrieved Documents)	Context Precision	Context Recall
Question (12) + Hypothetical Answer (12)	**0.717**	**0.328**
Multiquery Questions (24)	0.564	0.269
HyDE with BGE Reranker (24)	0.673	0.283
Only Question (24)	0.556	0.270
Only Hypothetical Answer (24)	0.713	0.295

Table 2: Evaluation of Final Answer Generation.

Retrieval Method (# of Retrieved Documents)	Precision	Recall	F1
Question (12) + Hypothetical Answer (12)	**0.551**	**0.645**	**0.591**
Multiquery Questions (24)	0.532	0.629	0.573
HyDE with BGE Reranker (24)	0.540	0.641	0.582
Only Question (24)	0.540	0.636	0.581
Only Hypothetical Answer (24)	0.539	0.642	0.583

Using the test dataset, we assess the QA-RAG framework’s performance against other baselines in two key areas: context retrieval performance and answer generation performance. We fix the number of retrieved documents to 24 and the number of reranked documents to top 6. The final answer LLM and the prompts are kept consistent throughout the experiments. Retrieval Augmented Generation Assessment (Ragas) framework [4] and Bertscore [10] were selected as the metric for context retrieval and answer generation, respectively.

3.2 Baselines

Proposed Method - Question + Hypothetical Answer: This method represents the QA-RAG framework, which incorporates both the question and hypothetical answer derived from the fine-tuned LLM.

Baseline 1 - Multiquery Questions: Multiquery questioning is widely employed in information retrieval to enhance the breadth and depth of document retrieval. We implemented this by expanding the original question by generating three additional questions using GPT-4. For each of the four total queries, six contextually pertinent documents were retrieved. To extract relevant documents, we applied the reranker for the top six most relevant documents.

Baseline 2 - HyDE with BGE Reranker: Utilizing LLM without additional fine-tuning, a single hypothetical document was created by GPT-3.5 Turbo following the “web search” prompt described in [5]. This document was then used for context retrieval of 24 documents. In contrast to the original HyDE methodology, we opted for the reranker to maintain consistency with other baselines and ensure fair evaluations by selecting the top 6 documents.

Baseline 3 - Only Question: This method represents the conventional RAG framework, which uses only the original user question for retrieving documents. From the 24 documents initially retrieved based on the question, the top six were selected based on their relevance using the reranker.

Table 3: Ablation Study Results of QA-RAG model.

Retrieval Method (# of Retrieved Documents)	Context Precision	Context Recall
Question (12) + Hypothetical Answer (12)	**0.717**	**0.328**
Only Question (12)	0.559	0.308
Only Hypothetical Answer (12)	0.700	0.259

Baseline 4 - Only Hypothetical Answer: This method relies solely on the fine-tuned LLM's response to fetch documents, deliberately omitting the use of the original question. It was done to observe the performance changes when only the answer is utilized for document retrieval. Similarly, out of the 24 documents retrieved, only the six most relevant documents were selected after reranking.

3.3 Results

Table 1 illustrates the effectiveness of the QA-RAG framework (i.e., "Question + Hypothetical Answer"), achieving the highest context precision (0.717) and context recall (0.328). The effectiveness of fine-tuned LLM is underscored as "Only Hypothetical Answer" surpasses "HyDE with BGE Reranker" in both metrics. The fine-tuned model enhanced the relevance and accuracy of the retrieved documents. The evaluation of final answer generation indicates similar findings. Similarly in Table 2, the QA-RAG framework achieved the highest scores in precision (0.551), recall (0.645), and F1 (0.591), demonstrating the efficacy and high-accuracy contexts in generating precise responses.

3.4 Ablation Study

We conducted an ablation study to further understand each component's individual contributions to the QA-RAG framework. We compare the results of document retrieval using only 12 documents retrieved based on the question (i.e., "Only Question") and hypothetical answer (i.e., "Only Hypothetical Answer"). We use the reranker to narrow down the top 6 documents.

Table 3 illustrates the ablation study results. Focusing on the hypothetical answer component alone, the framework achieved an impressive context precision of 0.700, lower by just 0.017 points than the full framework's performance. Conversely, relying solely on the user's question led to a marked drop in context precision to 0.559. Regarding context recall, the "Only Question" approach achieved a slightly higher score of 0.308 than the "Only Hypothetical Answer" method (0.259). The difference in context precision scores between the "Only Question" (0.559) and "Only Hypothetical Answer" (0.700) – more pronounced than in context recall – highlights the crucial role that hypothetical answers play in enhancing precision, suggesting their significant contribution to the framework's overall accuracy.

4 Conclusion

Our investigation into the QA-RAG framework reveals its effectiveness in merging generative AI and RAG and provides a cornerstone as one of the first instances of applying generative AI within the regulatory compliance domain. We address the significance of applying fine-tuned LLM by validating its strong performance in our experiments. For both context retrieval and answer generation, methods utilizing fine-tuned LLMs ("Question + Hypothetical Answer," "Only Hypothetical Answer") showed consistently high ranking in performance. Also, we underline the importance of a balanced hybrid question-answer approach as shown in the ablation study; the QA-RAG framework effectively merges the two elements and enhances its retrieval accuracy and relevance.

The QA-RAG framework can be implemented to reduce the time and resources needed to navigate complex regulations by streamlining the compliance process within the pharmaceutical industry. It can also be applied in other domains of industry such as legal compliance, financial regulation, and academic research, leading to swifter management of large data and enhanced decision making.

However, like any emerging technology, the long-term implications of the model within various industries will require ongoing evaluation and refinement. The integration of generative AI in highly specialized fields will raise questions about the model's adaptability to nuanced changes in data and industry practices. Thus, future developments should focus on proving the model's sustained effectiveness, ensuring it remains a robust tool in the face of ever-changing landscapes by staying aligned with the evolving generative AI technologies.

Acknowledgments

We acknowledge the use of ChatGPT, developed by OpenAI, for generating figures used in this paper to illustrate the model's design and functionality.

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439139, Development of a Cyber Crisis Response and Resilience Test Evaluation Systems)

This research was also supported by the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2024-RS-2023-00254129) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

References

[1] A. Anand, A. Anand, and V. Setty. 2023. Query understanding in the age of large language models. *arXiv preprint arXiv:2306.16004* (2023).

[2] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418* (2023).

[3] M. Crudeli. 2020. Calculating quality management costs. *Technology Record* (2020).

[4] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint arXiv:2309.15217* (2023).

[5] L. Gao, X. Ma, J. Lin, and J. Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496* (2022).

[6] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350, 6266 (2015), 1332–1338.

[7] L. Wang, N. Yang, and F. Wei. 2023. Query2doc: Query Expansion with Large Language Models. *arXiv preprint arXiv:2303.07678* (2023).

[8] S. Xiao, Z. Liu, P. Zhang, and N. Muennighof. 2023. C-pack: Packaged resources to advance general Chinese embedding. *arXiv preprint arXiv:2309.07597* (2023).

[9] P. Zhang, S. Xiao, Z. Liu, Z. Dou, and J. Y. Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554* (2023).

[10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).