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

Jaewoong Kim[1] Minseok Hur[2] Moohong Min[3]

[1]Department of Applied Data Science, Sungkyunkwan University
[2]Department of Immersive Media Engineering/Convergence Program for Social Innovation, Sungkyunkwan University
[3]Department of Computer Education/Convergence Program for Social Innovation, Sungkyunkwan University

Motivation

Needs of Efficient Systems for Navigating Pharmaceutical Guidelines

- Navigating through complex and extensive pharmaceutical regulatory guidelines is a **time-consuming** task for industry players
- Up to 25% of the pharmaceutical site's operation budget can be consumed for compliance efforts alone

U.S. Food and Drug Administration (FDA)

European Medicines Agency (EMA)

Need for Enhanced RAG Framework to Meet Compliance Demands

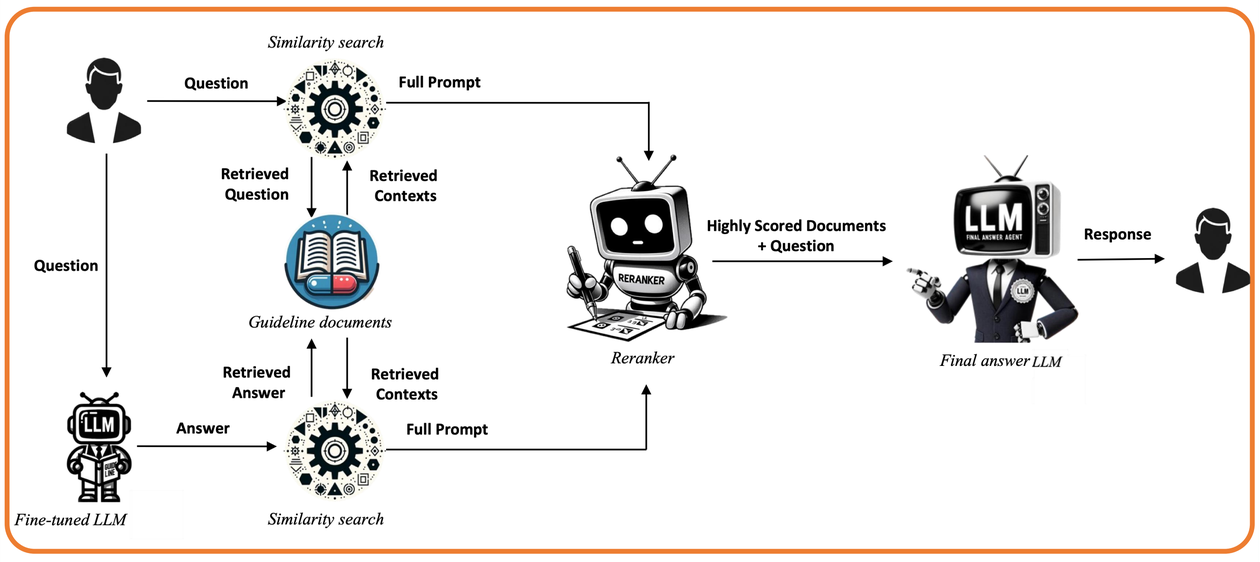
- **Large language models (LLMs)** and **retrieval-augmented generation (RAG)** frameworks **fall short** of generating precise requirements for the pharmaceutical industry
- Conventional RAG performs poorly: context precision (0.556), context recall (0.270)

Dual-track Retrieval for Enhanced Generation

Question and Answer Retrieval Augmented Generation (QA-RAG)

- To meet the needs, we leverage both the **question** and the **hypothetical answer** (generated by a **fine-tuned LLM**) to enhance the retrieval of specific regulatory information
- For retrieval, Facebook AI Similarity Search (FAISS) is employed [1]
- After retrieval, reranking is performed to select relevant documents
- After reranking, the final response is generated with few-shot prompting using question-answer pairs

The QA-RAG Framework



Embedding and Reranking the Pharmaceutical Guidelines

- We compile a total of **1,404** pharmaceutical guideline documents from FDA* (1,263) and ICH** (141) for document retrieval
- We **embed** the documents using the LLM-Embedder [2]
- We **rank** the documents by employing the BGE reranker [3]

$$S_{i,j} = Rerank\left(Q, D_{i,j}\right) \ where\ j \in \{1, 2, \dots, N\} \quad (1)$$

$$D = \left\{D_{i,j} \,\middle|\, S_{i,j}\ is\ among\ the\ top\ scores\right\} \quad (2)$$

- For each guideline document D_i, we divided them into chunks of 10,000 characters with an overlap of 2,000
- $D_{i,j}$ denotes the *j*-th chunk of the document D_i
- $S_{i,j}$ denotes the relevance score assigned to $D_{i,j}$ in relation to the query Q
- D is used in the final stage of response generation

* FDA: Food and Drug Administration
** ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

Final Answer Generation

- ChatGPT-3.5-Turbo is utilized for generating the final answer
  - A prompt with one question-answer pair was given

[1]: D. Danopoulos, C. Kachris, and D. Soudris. 2019. Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud. In Embedded Computer Systems: Architectures, Modeling, and Simulation. 373–386.
[2]: 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).
[3]: 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).

Challenges

Challenges of Conventional RAG Approaches

- Single query is limited in retrieving relevant documents
  → **Multiquery retrieval** and **HyDE** have been proposed
- However, multiquery is still confined to the narrow scope of the original query and HyDE utilizes a general LLM, performing poorly in highly specialized area

→ **QA-RAG complements the two and significantly outperforms them**

Selection of the Fine-tuned LLM

Response Tailored to the Pharmaceutical Realm

- We leverage the **1,681 FAQ dataset** provided by the FDA for fine-tuning LLMs
  - 85% (*training*), 15% (*validation*), 5% (*testing*)

	ChatGPT-3.5-Turbo	ChatGPT-4*	Llama3 8B Instruct**
Precision	**0.579**	0.505	0.516
Recall	0.589	**0.622**	0.582
F1-Score	**0.578**	0.555	0.544

* ChatGPT-4 was not fine-tuned ** Llama3 8B Instruct was fine-tuned using the QLoRA technique

- We select ChatGPT-3.5-Turbo as the fine-tuned LLM

Experimental Results

- The testing portion from the FAQ dataset was used to evaluate the results
  - **Context Precision** assesses the relevance of the retrieved documents to the question
  - **Context Recall** assesses the ability of the retrieval system to gather the information needed to answer the question
- We retrieve 24 documents and select 6 highest-scored documents for the final answer generation

Evaluation of Context Retrieval

Retrieval Method	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

Evaluation of Final Answer Generation

Retrieval Method	Precision	Recall	F1-Score
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

Ablation Study

Retrieval Method	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

Conclusions

QA-RAG Effectively Merges Generative AI and RAG Framework

- We find that applying a fine-tuned LLM significantly validates strong performance
- Our framework can be applied in other domains of industry, such as legal compliance, financial regulation, and academic research

Contact

Code

Jaewoong Kim:
jwoongkim11@g.skku.edu

Minseok Hur:
alexhur3535@skku.edu

Moohong Min:
iceo@skku.edu