Term Project Report

Development of a Domain-Specific Q&A Chatbot for Querying Korean Housing Subscription Guidelines

Team: AML / Member: 이서혁 (2024-25545)

1. Problem Definition

The housing subscription system in South Korea is one of the most critical institutional mechanisms for non-homeowners to acquire their own homes. However, related laws (such as the Housing Act and the Special Act on Public Housing) and guidelines are vast, complex, and subject to frequent amendments. This creates a significant barrier to entry for the general public, making it difficult for them to accurately understand and prepare for the process.

Existing information retrieval methods, such as searching through web portals or manually checking scattered public notices, incur high temporal costs and are prone to misunderstandings due to non-expert information. In particular, there is a lack of systems capable of providing accurate, personalized information—such as calculating subscription score points or verifying specific eligibility requirements—leading to persistent issues of information asymmetry.

2. Research Goals

The objective of this project is to build an accurate and reliable Q&A chatbot system based on housing subscription laws and practical handbooks using **Retrieval-Augmented Generation (RAG)** technology.

The specific sub-goals are as follows:

1. Structure vast and fragmented subscription-related documents (laws, FAQs, operational handbooks) into a Vector Database.
2. Identify user query intent and extract the most relevant supporting documents via **Semantic Search**.

3. Generate accurate answers with minimized **Hallucination** using a Large Language Model (LLM) based on the extracted context.
4. Experimentally verify and optimize hyperparameters (e.g., Chunk Size, Retriever Weights) using the **Ragas** framework.

3. Experimental Data Set

To ensure the provision of authoritative information, this study exclusively utilized official documents published by the Ministry of Land, Infrastructure and Transport and related agencies. The dataset is categorized into guidance materials, legal bases, and operational regulations.

- **Guidance/Reference:** 2024 주택청약 FAQ, 2024년 주택업무편람
- **Legal Basis:** 공공주택 특별법 시행규칙/시행령, 주거기본법 시행령, 주택공급에 관한 규칙
- **Operational Regulations:** 주택분양규정(2024.12.24 개정), 주택분양규정 시행세칙

A total of 8 PDF documents were processed using pdfplumber to extract text. Page-level metadata (source filename, page number) was preserved to ensure the traceability of the answers.

4. Set-up

The hardware and software environment configurations for system construction and experimentation are as follows:

- **Framework:** LangChain, Streamlit (UI)
- **Embedding Model:** OpenAI text-embedding-3-small (1536 dimensions) – Selected for its high expressive power and cost-efficiency.
- **Inference Model (LLM):** GPT-4o-mini – Selected considering inference speed, cost, and Korean language processing capabilities within the RAG pipeline.
- **Vector Store:** FAISS (IndexHNSWFlat) – Applied the HNSW algorithm for fast Approximate Nearest Neighbor (ANN) search on high-dimensional data.
    - Parameters: M=16, efConstruction=200, efSearch=200
- **Evaluation:** Ragas Framework (Metrics: Context Recall, Context Precision)

5. Used Methodology

5.1 Data Pipeline & Chunking Strategy

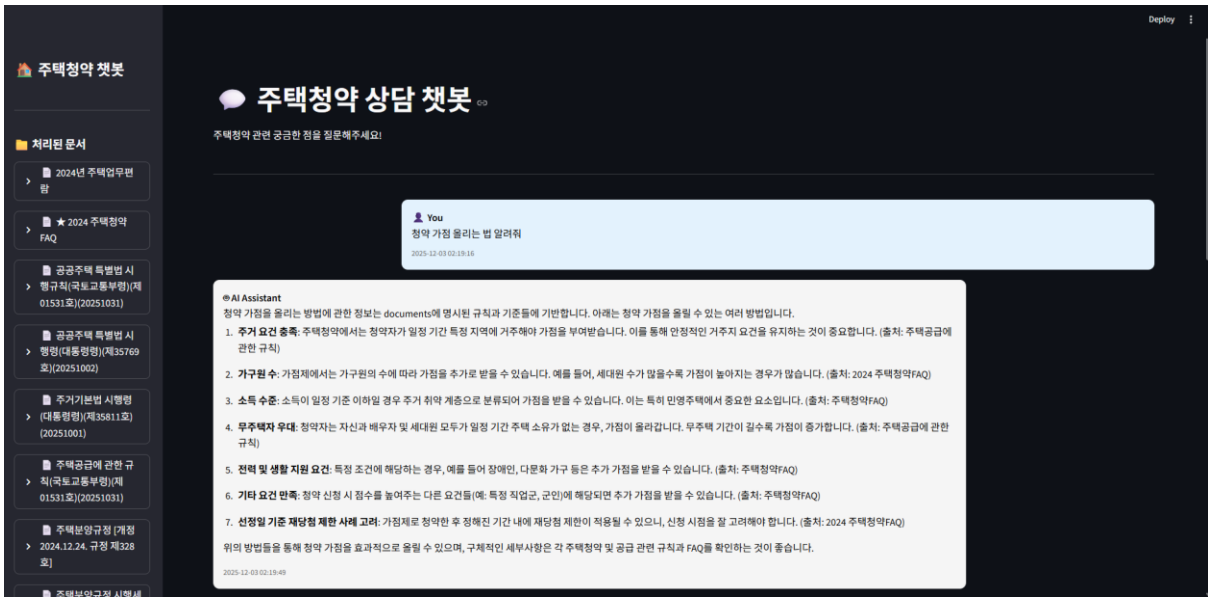
Text extracted from PDF documents was split using RecursiveCharacterTextSplitter. Given the characteristics of legal documents, where sentences are long and context dependency is high, finding the appropriate Chunk Size is crucial. Experiments were conducted with various sizes (700, 1024, 1400 tokens).

5.2 Retrieval System (Retriever & Reranker)

In the initial retrieval stage, a Hybrid Search method combining meaning-based Dense Retriever (FAISS) and keyword-based Sparse Retriever (BM25) was considered. To improve the accuracy of the retrieved documents, a two-stage pipeline was constructed wherein the top-k documents from the first retrieval are re-sorted by relevance to the query using the BGE-M3 Reranker (Cross-Encoder).

5.3 Optimization & UI

- **Batch Processing:** Applied batch processing during embedding generation to prevent OpenAI API token limits and speed degradation.
- **Caching:** Implemented local caching for processed chunks and vector indexes, reducing system initialization time by over 95% upon re-execution.
- **Transparency:** Implemented a Streamlit-based UI that explicitly displays the original text and page number of the referenced documents as 'Sources' for user verification.



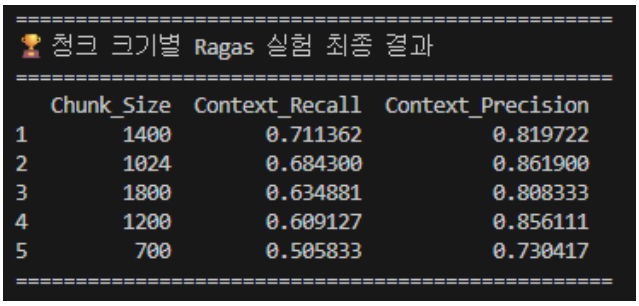
6. Result

Quantitative evaluation was performed based on a 'Question-Keyword-Golden Context' test set (20 items) generated using the Ragas framework.

6.1 Chunk Size Experiment

The comparison of Context Recall according to chunk size is as follows:

- **Chunk Size 1400:** Recall 0.7114 (Best), Precision 0.8197
- **Chunk Size 1024:** Recall 0.6843, Precision 0.8619
- **Chunk Size 700:** Recall 0.5085, Precision 0.7304



🏆 청크 크기별 Ragas 실험 최종 결과

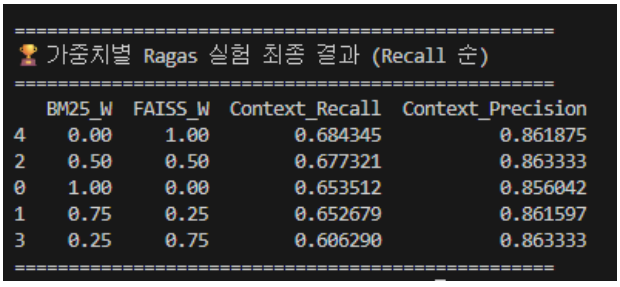
	Chunk_Size	Context_Recall	Context_Precision
1	1400	0.711362	0.819722
2	1024	0.684300	0.861900
3	1800	0.634881	0.808333
4	1200	0.609127	0.856111
5	700	0.505833	0.730417

The results indicate that a chunk size of 1400 yielded the highest Recall. As the chunk size decreased (e.g., 700), the context was severed, leading to a drastic drop in retrieval performance.

6.2 Retriever Weight Experiment

The experiment on adjusting the weights of the Ensemble Retriever (BM25 vs. FAISS) yielded the following:

- **FAISS Only (Weight 1.0):** Recall 0.6843 (Best)
- **Hybrid (0.5 : 0.5):** Recall 0.6773
- **BM25 Only (Weight 1.0):** Recall 0.6555



🏆 가중치별 Ragas 실험 최종 결과 (Recall 순)

	BM25_W	FAISS_W	Context_Recall	Context_Precision
4	0.00	1.00	0.684345	0.861875
2	0.50	0.50	0.677321	0.863333
0	1.00	0.00	0.653512	0.856042
1	0.75	0.25	0.652679	0.861597
3	0.25	0.75	0.606290	0.863333

Contrary to common assumptions, the single-use model of semantic-based search

(FAISS) showed superior performance over keyword matching (BM25) in this specific domain.

7. Analysis

7.1 Impact of Chunk Size

Legal and regulatory documents regarding housing subscriptions utilize long-winded sentences to describe complex preconditions or exception clauses (e.g., "However, this shall not apply in cases of..."). Consequently, setting a small chunk size (under 700 tokens) caused conditions and conclusions to be separated into different chunks, damaging semantic completeness. The 1400-token setting was analyzed to have best preserved this context.

7.2 Characteristics of Search Algorithms

The experiment showed that the FAISS-only model outperformed the BM25 hybrid model. This is attributed to the nature of user queries; questions often require a semantic interpretation of a situation (e.g., "Is a father's spouse recognized as a dependent in case of remarriage?") rather than simple keyword matching (e.g., "Subscription account conditions"). The embedding model was more effective in capturing these contextual similarities.

8. Novelty

1. **Domain-Specific Data Pipeline:** Established a preprocessing and chunk optimization process specialized for complex Korean legal/regulatory documents, rather than a generic RAG pipeline.
2. **Optimization based on Quantitative Experiments:** Objectively verified and optimized system hyperparameters using the **LLM-as-a-Judge** method via the Ragas framework, rather than relying on intuition.
3. **Practical UX Implementation:** Built a UI that goes beyond a simple chatbot by providing original source evidence and metadata, securing the reliability essential for legal information services.

9. Future Direction(s)

To address the limitations of this study and further improve performance, the following future research directions are proposed:

1. **Query Routing System:** Introduce a routing module to classify general questions versus those requiring RAG, thereby increasing efficiency.
2. **Dynamic Hybrid Search:** While FAISS-only was superior in this study, cases requiring specific proper noun searches (e.g., specific article numbers) exist. Research is needed on an adaptive search system that dynamically adjusts BM25 weights based on query type.
3. **Multi-turn Context Awareness:** The current system focuses on single-turn question processing. Improvements are needed to reflect conversation history in **Query Rewriting** to handle follow-up questions containing pronouns (e.g., "Then what about *that* condition?").