Retrieval-Augmented Generation (RAG) Pipeline for Financial Data Question Answering

Sigmoix.AI Internship Assessment

1 Objective

The goal of this assessment is to design and implement a Retrieval-Augmented Generation (RAG) pipeline capable of answering natural language queries from financial documents. We used Meta's Q1 2024 financial report as our sole knowledge source.

2 Step 1: Basic RAG Pipeline

Approach

- Extracted and cleaned raw text from PDF using `PyMuPDF`.
- Chunked the text into overlapping windows of 100 words.
- Generated embeddings for each chunk using `all-MiniLM-L6-v2` model.
- Indexed embeddings in FAISS for vector similarity retrieval.
- Used an open-source LLM (e.g., Falcon) to generate contextual answers based on retrieved text.

Test Queries and Answers

- **Q:** What was Meta's revenue in Q1 2024?
  A: $36,455

- **Q:** What were the key financial highlights for Meta in Q1 2024?

  A: Revenue $36,455 $28,645 27%
  Costs and expenses 22,637 21,418 6%
  Income from operations $13,818 $7,227 91%
  Operating margin 38% 25%
  Provision for income taxes $1,814 $1,598 14%
  Effective tax rate 13% 22%
  Net income $12,369 $5,709 117%
  Diluted earnings per share (EPS) $4.71 $2.20 114%

3 Step 2: Structured Data Integration

Approach

- Extracted tabular data using `Camelot`.
- Converted tables to Pandas DataFrames.
- Implemented hybrid retrieval by combining FAISS (for text) with DataFrame lookups (for structured data).

Test Queries and Answers

- **Q:** What was Meta's net income in Q1 2024 compared to Q1 2023?

- **Answer:** $12,369 $5,709

4 Step 3: Query Optimization and Advanced RAG

Enhancements

- Query optimization via LLM rewriting to clarify vague inputs.
- Used cross-encoder reranking to boost top-k relevance.
- Conducted retrieval experiments with different chunk sizes (100 vs 300 tokens).

Evaluation Metrics

- **Question:** What was Meta's revenue in Q1 2024?
- **Answer:** $36,455
- **Retrieval:** Precision = 1.0, Recall = 1.0
- **Generation :** 'ROUGE-L': 0.5714285714285715, 'BLEU': 9.291879812217675e-232

- **Question:** What was Meta's net income in Q1 2024 compared to Q1 2023?
- **Answer:** $12,369 $5,709
- **Retrieval:** Precision = 1.0, Recall = 1.0
- **Generation:** ROUGE-L = 0.44, BLEU = 1.29e-231

Ablation Study

We compared retrieval results for the question **"What was Meta's revenue in Q1 2024?"** with and without reranking.

The effect of reranking on retrieval results was evaluated by comparing the top-5 retrieved chunks with and without reranking.

Without Reranking:

- Meta Reports First Quarter 2024 Results MENLO PARK, Calif. – April 24, 2024 – Me

- believe that this methodology can provide useful supplemental information to hel

- well." First Quarter 2024 Financial Highlights Three Months Ended March 31, % Ch
- marketable securities were $58.12 billion as of March 31, 2024. Free cash flow w
- following table presents our segment information of revenue and income (loss) fr

With Reranking:

- Meta Reports First Quarter 2024 Results MENLO PARK, Calif. – April 24, 2024 – Me
- believe that this methodology can provide useful supplemental information to hel
- marketable securities were $58.12 billion as of March 31, 2024. Free cash flow w
- well." First Quarter 2024 Financial Highlights Three Months Ended March 31, % Ch
- following table presents our segment information of revenue and income (loss) fr

5 Improvement Proposals

Based on the evaluation results and ablation study, several areas for improvement have been identified to enhance the accuracy and reliability of the RAG pipeline:

1. **Better Answer Generation:** The current Flan-T5 model sometimes produces incomplete or generic responses (e.g., "Depreciation and amortization" for operating expenses). Fine-tuning the generator on financial QA datasets or using a larger instruction-tuned LLM can improve factual accuracy.
2. **Enhanced Table Retrieval:** For numeric answers (e.g., revenue, net income, operating expenses), the pipeline should always prioritize structured table lookups over generative outputs to avoid hallucination.
3. **Query-Specific Retrieval:** Incorporate query classification (numeric vs. descriptive) to decide when to use tables, text chunks, or both. This would improve answers for factual numeric queries.
4. **Hybrid Retrieval Improvements:** Integrate BM25 or ColBERT with FAISS to combine sparse and dense retrieval, increasing recall for text-based questions.
5. **Better Evaluation Metrics:** Current BLEU scores are very low due to short, factual answers. Using smoothing functions or alternative metrics like Exact Match (EM) and F1-score would provide better evaluation for short factual outputs.
6. **Reranking Enhancements:** Although reranking improved precision, further improvements could be achieved by training a financial-domain cross-encoder or using MonoT5 as a reranker.
7. **Numeric Reasoning:** Add a component that can perform arithmetic or comparison on table values (e.g., revenue growth calculation) for more complex questions.
8. **Reduce Warning Messages:** Remove duplicate parameters by setting only `max_new_tokens` and not `max_length` to avoid Hugging Face warnings during text generation.

6 Tools Used

- **PyMuPDF (fitz):** For extracting text from the PDF financial report.
- **Camelot:** For extracting tabular data from the PDF and converting it into DataFrames.
- **SentenceTransformers:** For generating dense embeddings of text chunks (`all-MiniLM-L6-v2`).
- **FAISS:** For efficient similarity search and retrieval of relevant chunks.
- **CrossEncoder:** For reranking retrieved chunks based on query relevance (`ms-marco-MiniLM-L-6`
- **Transformers Pipeline:** For query rewriting (`flan-t5-base`) and final answer generation (`flan-t5-large`).
- **Evaluation Libraries:** ROUGE and BLEU for evaluating generated answers, along with Precision@k and Recall@k for retrieval evaluation.
- **Pandas & NumPy:** For handling tabular data and numerical computations.

7 Challenges

- Extracting accurate table formats from multi-column PDFs.
- Managing hallucination in long-form generated answers.
- Balancing retrieval granularity (chunk size vs recall).

8 Conclusion

This project demonstrates the use of open-source RAG tools to build an effective QA pipeline over financial documents. The staged approach (text, table, reranking) and thorough evaluation confirm its ability to answer questions with high accuracy and clarity.