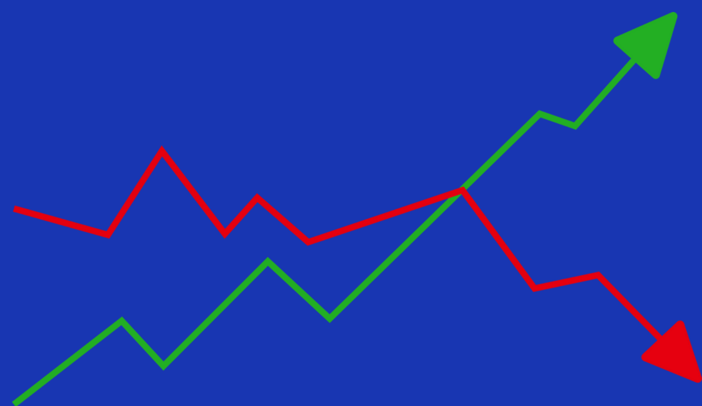
FPT Education

FPT UNIVERSITY

ENHANCING STOCK MARKET INFORMATION RETRIEVAL WITH RETRIEVAL AUGMENTED GENERATION (RAG)

Final Project

SEG301 - Group 5 - AI1801



Members

Phong Huan	Huynh Nhu	Anh Thu	Cong Lenh	Huu Nhan
CE180685	CA182007	CE180615	CE180059	CE181655

FPT Education
FPT UNIVERSITY

Sections

1/ Introduction
2/ Related Work
3/ Data Preparation
4/ Proposed Metholodogy
5/ Experimental Result
6/ Conclusion

FPT Education
FPT UNIVERSITY

Introduction

Current Challenges of LLMs

- Hallucination problem
- Knowledge update difficulty
- Lack of domain specialization

RAG Applications

- **Stock markets** are highly dynamic, requiring continuously updated NLP models
- **Current problem**: Traditional NLP models struggle with complex, rapidly changing financial data.

--> **RAG as a potential solution**: Enables retrieving key financial data and generating more accurate insights.

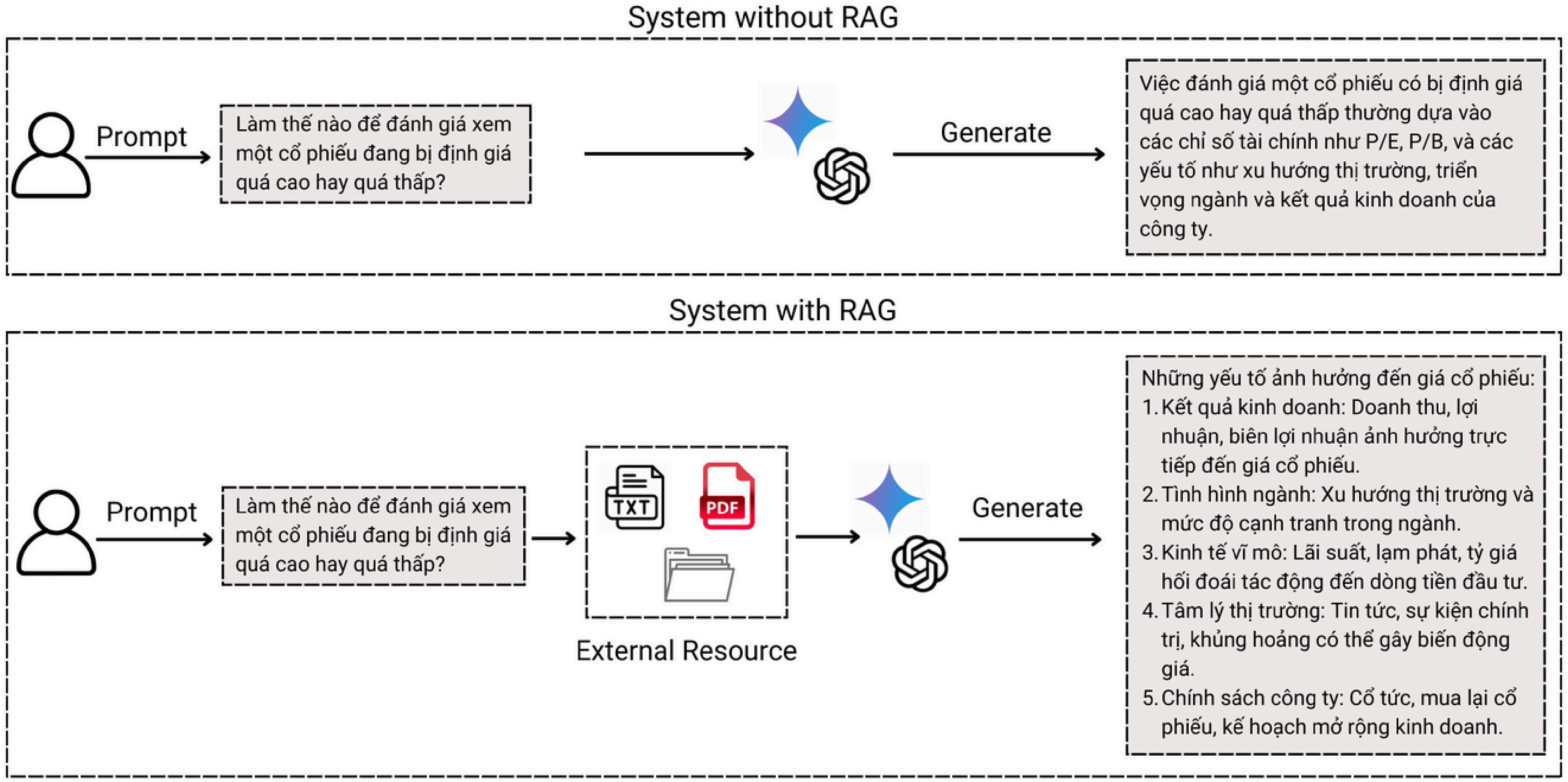


Fig 1. System without RAG and System with RAG

Introduction

Research Objectives

- Apply **RAG** in the stock market domain.
- Use **LaBSE** as the embedding model and **Gemini-1.5-Flash** as the LLM.
- Evaluate performance using **ROUGE Score** and **Latency** metrics.

```
🟢 Nhập câu hỏi của bạn (hoặc gõ 'exit' để thoát): Không chỉ KSV, nhiều cổ phiếu chung ngành cũng tích lũy thị giá liên tục như?

💡 Câu trả lời:
Không chỉ KSV, nhiều cổ phiếu chung ngành cũng tích lũy thị giá liên tục như MSR, HGM, AAH, BMC...


🟢 Nhập câu hỏi của bạn (hoặc gõ 'exit' để thoát): VIX dẫn đầu về khối lượng sang tay với hơn 96 triệu cổ phiếu, cao hơn 3 mã xếp sau là

💡 Câu trả lời:
VIX dẫn đầu về khối lượng sang tay với hơn 96 triệu cổ phiếu, cao hơn 3 mã xếp sau là TCB, HPG và EVF cộng lại.
```

Fig 2. RAG system for Stock Market Information

Related Work

NLP in Stock Market

- Akita et al.: **LSTM** + **news data** → cross-company impact.
- Sidra & Jaydip: **Social sentiment** → better predictions.
- "Smart Trader" **chatbot** → AI-powered stock advice.

RAG in NLP

- Combines **retriever** (external knowledge) + **generator** (LLM)
- Benefits:
  + Reduces hallucinations, outdated info.
  + Used in chatbots, fact-based text generation.

RAG in Finance & Stock Market

- **FinBERT + RAG**: Sentiment-based stock predictions.
- **Stock-Chain framework**: Retrieval + reasoning for market insights.
- **Multi-modal** analytics: Combines stock prices, news, technical charts.
- Financial risk assessment: **LLaMa3.1**, **Gemini-1.5-Flash** → better audit analysis.

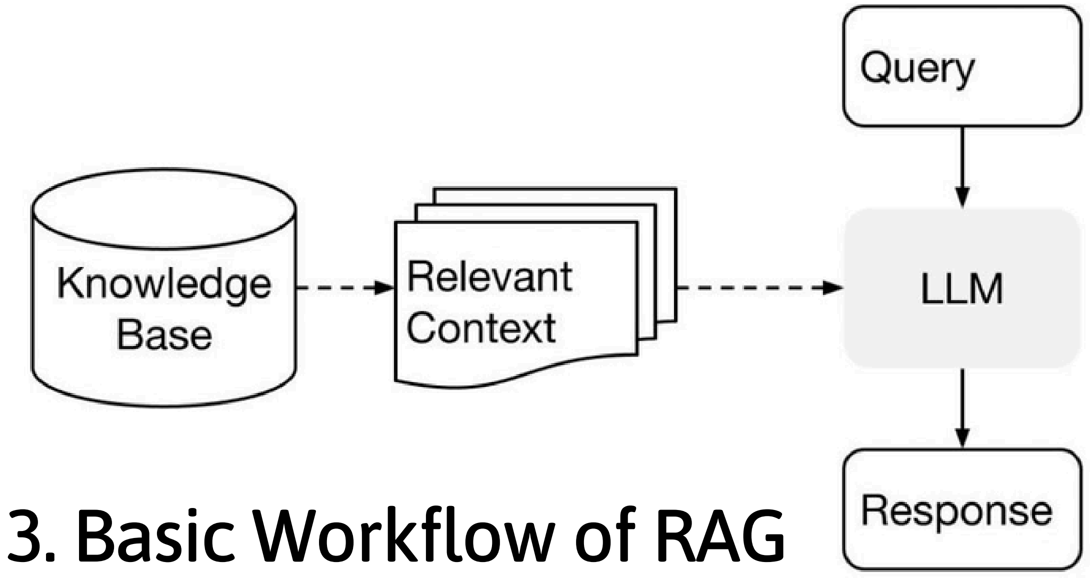


Fig 3. Basic Workflow of RAG

Data Collection

Information

- Source: 📰 **VnExpress** (Vietnamese news platform).
- Dataset: 📄 **30** stock market-related articles.

Content Coverage

- 📈 Market Trends & Index Performance
- 🏢 Corporate Stocks & Sectors
- 🌍 Foreign Investment & Liquidity
- ⚖️ Regulatory & Corporate Actions

Category	Number
Market Trends & Index Performance	8
Corporate Stocks & Sectors	9
Foreign Investment & Liquidity	7
Regulatory & Corporate Actions	6
Total	**30**

Table 1. Number of Main Documents

Data Preprocessing

Vietnamese Text Processing

- Uses **Underthesea** for word tokenization.
- **Normalization**: Remove spaces, URLs, special characters.
- **Lowercasing**: Standardizes text.
- **Stop Words Removal**: Filters out non-informative words.

Stop Words Removal

- Eliminates common words (e.g., conjunctions, prepositions).
- Reduces dimensionality and improves NLP efficiency.

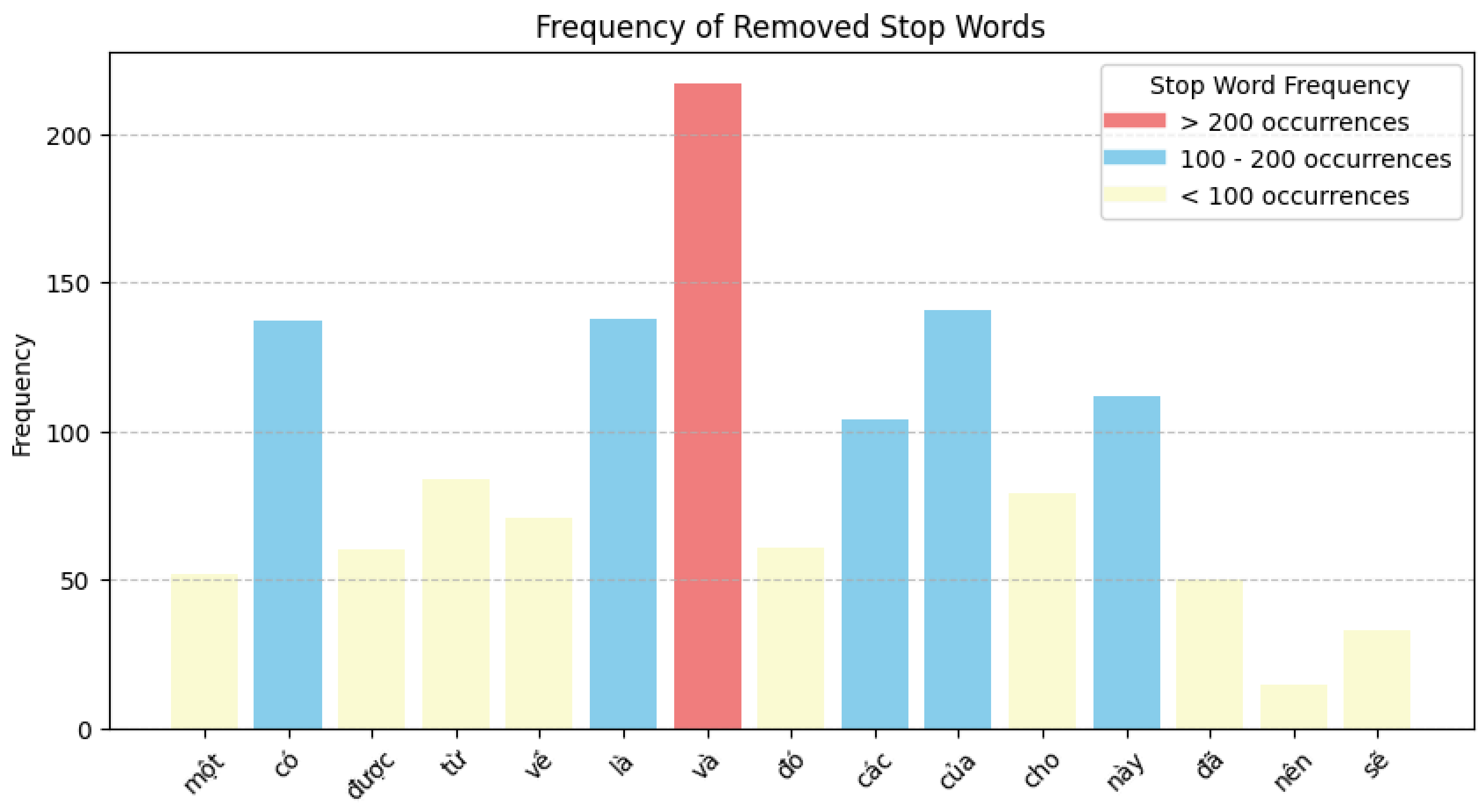


Fig 4. Frequency of Removed Stop Words

Proposed Methodology

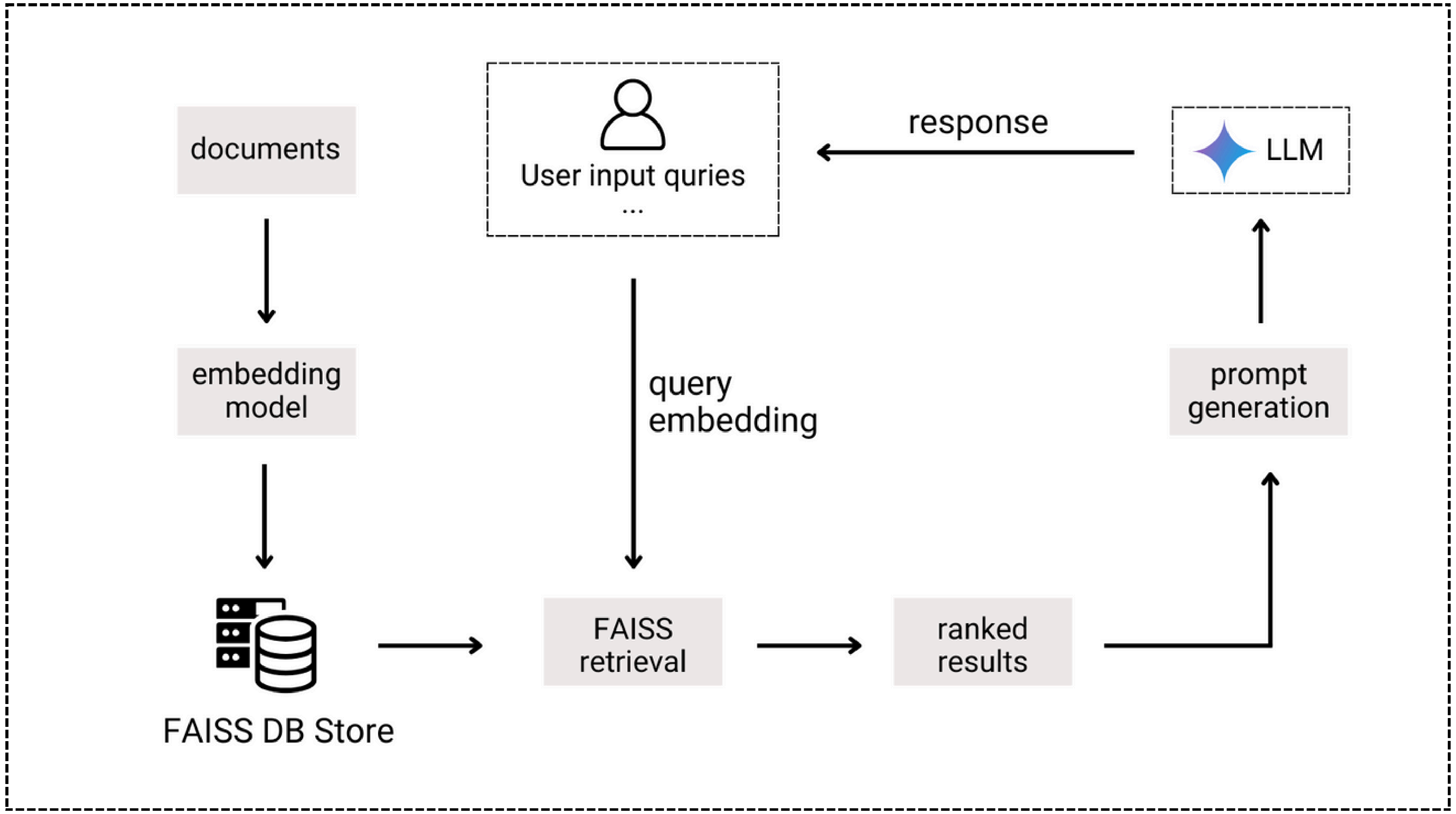


Fig 5. Retrieval-Augmented Generation (RAG) System Architecture

Text Embedding

FPT Education
FPT UNIVERSITY

Purpose

- Convert stock market news into numerical vectors for analysis.

Embedding Model

- **LaBSE** from Sentence-Transformers.
- Chosen for multilingual support, especially **Vietnamese**.

Preprocessing & Tokenization

- **Jieba** tool segments text before embedding

Storage & Retrieval

- Embeddings stored in **FAISS** for fast similarity search

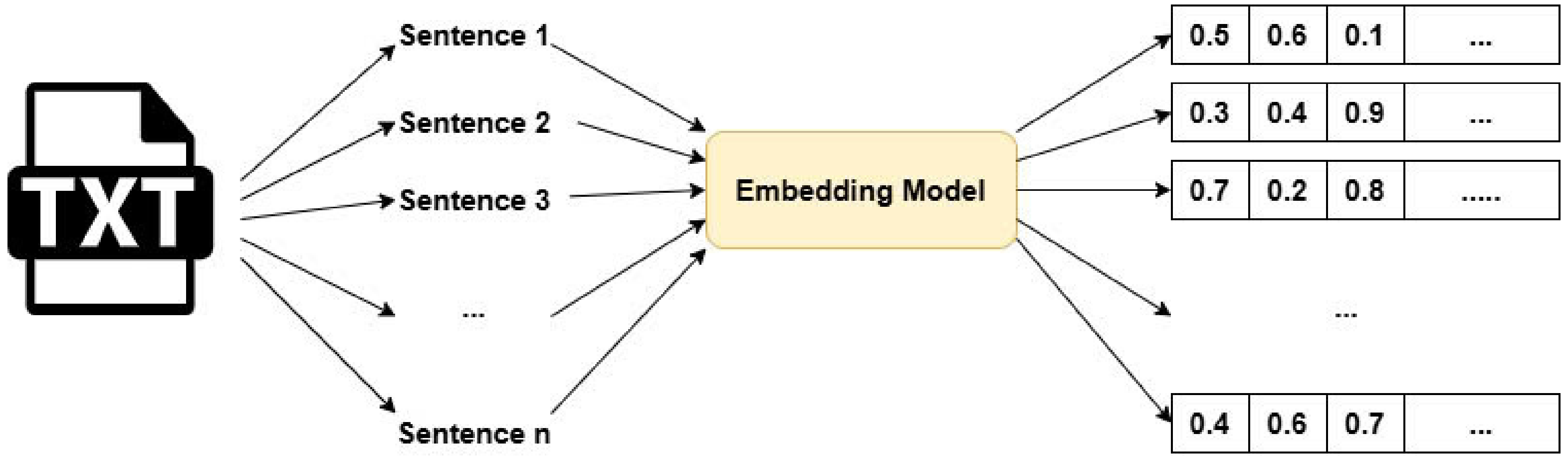


Fig 6. Text Embedding Process: Converting Raw Text into Numerical Representations for Efficient Retrieval

Information Retrieval

FPT Education
FPT UNIVERSITY

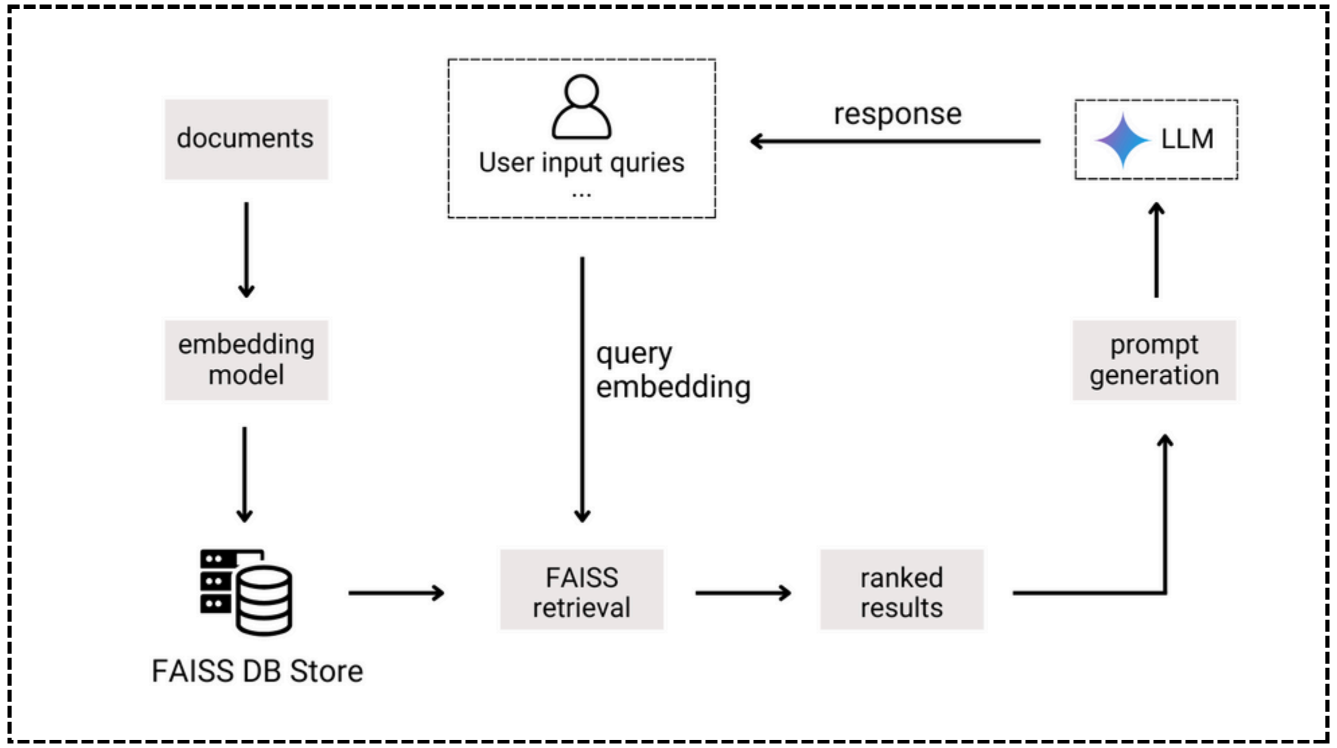
Retrieval Process

- 1 Embed query using **LaBSE**.
- 2 **FAISS search** → Finds top-k nearest documents using L2 distance.

$$d(\mathbf{q}, \mathbf{d}) = \sqrt{\sum_{i=1}^{n} (q_i - d_i)^2}$$

- 3 **BM25 ranking** → Re-ranks results for keyword relevance.

$$BM25(D, Q) = \sum_{t \in Q} IDF(t) \cdot \frac{f(t, D) \cdot (k_1 + 1)}{f(t, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgD}}\right)}$$



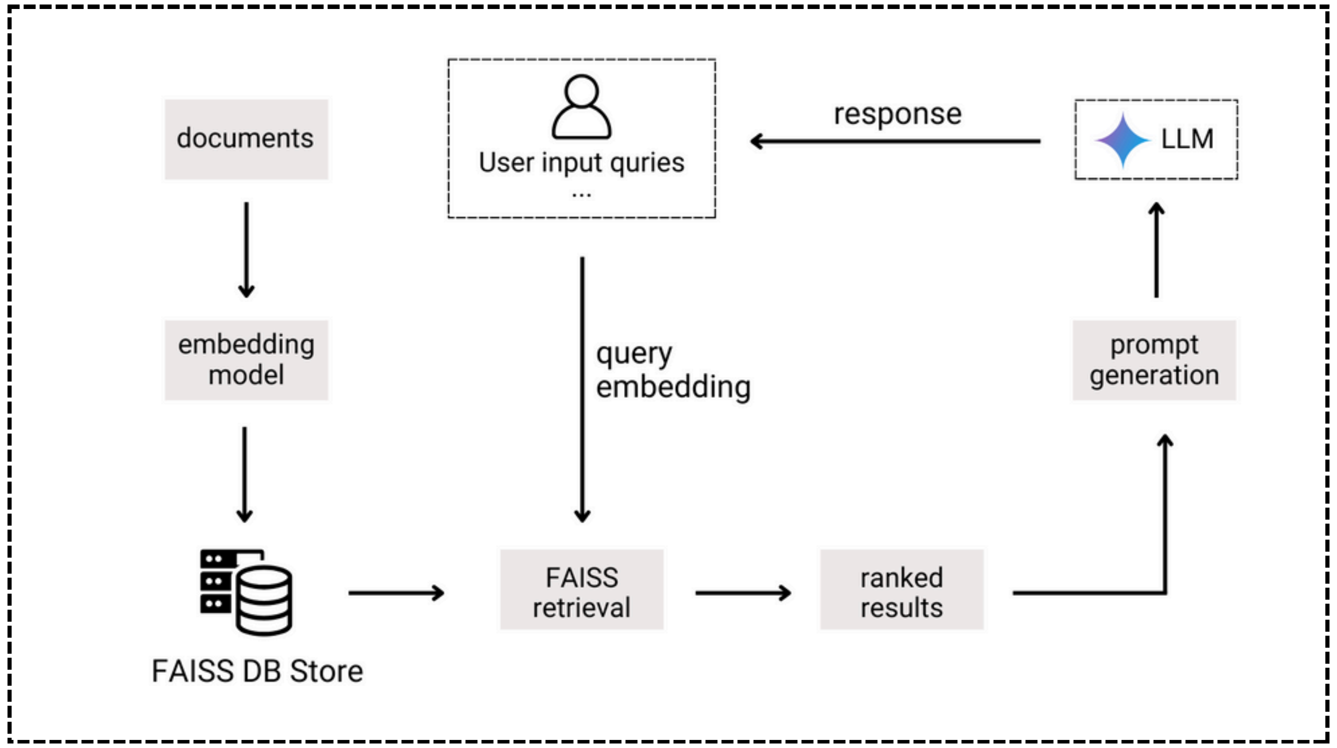
Information Generation

Model Using

- **Gemini-1.5-Flash**: Optimized for real-time responses.
- Low latency & efficient long-context processing.

Process

- 1️⃣ Retrieve relevant documents.
- 2️⃣ Craft structured prompt (clear instruction + user query + retrieved texts).
- 3️⃣ Generate response grounded in factual stock market information.



FPT Education
FPT UNIVERSITY

Evaluation Metrics

ROUGE Score

- Measures response quality 📝
- **ROUGE-1**: Overlap of single words (unigrams).

$$\text{ROUGE-1} = \frac{\sum_{w \in W} \min(\text{count}_{\text{gen}}(w), \text{count}_{\text{ref}}(w))}{\sum_{w \in W} \text{count}_{\text{ref}}(w)}$$

- **ROUGE-2**: Overlap of two-word sequences (bigrams).

$$\text{ROUGE-2} = \frac{\sum_{b \in B} \min(\text{count}_{\text{gen}}(b), \text{count}_{\text{ref}}(b))}{\sum_{b \in B} \text{count}_{\text{ref}}(b)}$$

- **ROUGE-L**: Measures longest common subsequence (LCS) for sentence structure and fluency

$$\text{ROUGE-L} = \frac{LCS(\text{gen}, \text{ref})}{\text{length}(\text{ref})}$$

Latency

- Measures response speed ⏳
- Time from **query submission** → **response generation**.
- Crucial for real-time stock market insights

Results

TABLE II
ROUGE SCORE RESULTS

ROUGE Metric	Precision	Recall	F-measure
ROUGE-1	0.9616	0.9597	0.9555
ROUGE-2	0.9374	0.9356	0.9314
ROUGE-L	0.9496	0.9478	0.9436

TABLE III
LATENCY RESULTS

Latency Metric	Time (seconds)
Average Latency	1.1626
Minimum Latency	0.917
Maximum Latency	2.455

Conclusion

Project Focus

- Evaluated a **Retrieval-Augmented Generation (RAG)** system for **stock market** queries

Key Findings

- **High ROUGE Scores** → Accurate & fluent financial responses.
- **Low Latency** → Fast query processing for real-time insights.

Implications

- The system provides **reliable and timely** financial information

Thank you
for your attention!