DEEP LEARNING IN BUSINESS

RAG-BASED Q&A SYSTEM

PRESENTED BY GROUP 1

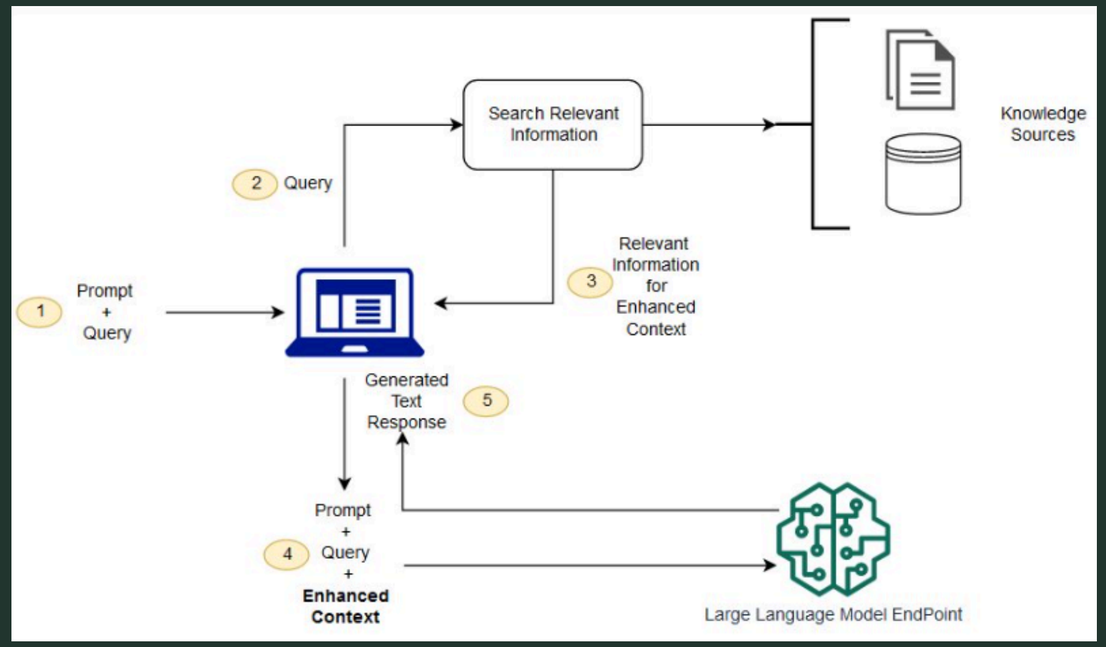
TABLE OF CONTENTS

INTRODUCTION

- The Challenge: Information overload from large, specialized PDF documents.
- The Limitation: Traditional keyword search (Ctrl+F) is ineffective for complex, semantic questions.

--> The Need: A conversational tool to ask questions and receive accurate, fast answers directly from the document's content.

PROPOSED SOLUTION: RAG



Solution: Using the RAG (Retrieval-Augmented Generation) model.

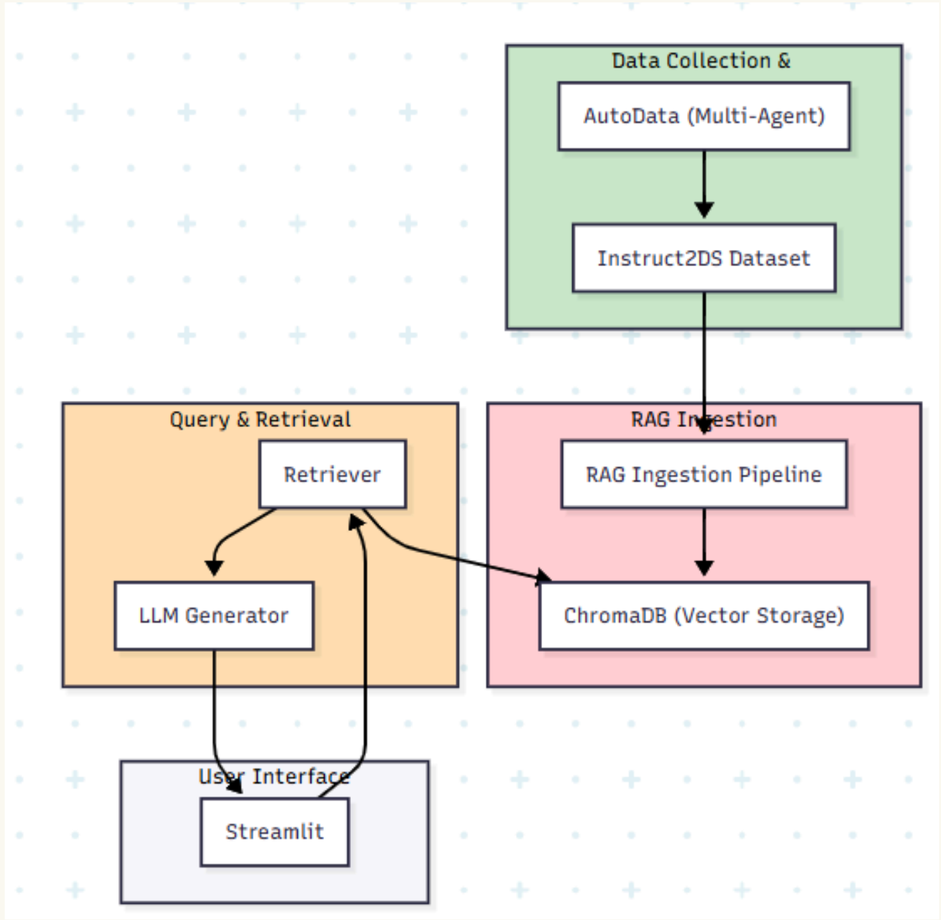
What is RAG? A technique that combines the strengths of:

- **Retrieval:** Searching and fetching the most relevant information from a database (the PDF content).
- **Generation:** Using a Large Language Model (LLM) to generate a natural answer based on the retrieved information.

Why we choose RAG?

Advantages: Ensures the answer is grounded in the document, reduces the risk of the LLM 'making up' information, and allows us to cite the source.

SYSTEM ARCHITECTURE OVERVIEW



FIRST PHASE

Data Ingestion & Embedding Phase

- Collect and prepare data
- Clean and normalize text
- Chunk documents
- Generate vector embeddings
- Store embeddings in a Vector Database
- Initialize the RAG pipeline

SECOND PHASE

Query & Answer Generation Phase

- User submits a query
- Perform semantic retrieval
- Combine retrieved context with the LLM
- Generate a concise, accurate answer

PHASE 1 :

Data Ingestion & Management Phase

Objective: **Build a Knowledge Base from PDFs.**

Load & Clean

Load, clean, and normalize data (from Instruct2DS).

metadata.json

pdf downloader

pdf parser

Text Chunking

Split the text into meaningful "chunks".

text_chunker

Embedding

Convert each chunk into a numerical vector (using an embedding model).

model gte-Qwen2-1.5B-instruct

Vector Store

Store all vectors in ChromaDB.

PHASE 2 : Query & Answer Generation

Objective: Find relevant text chunks from ChromaDB
→ Use those chunks to generate the final answer.

User query

Embed the user query

Query vector

Search ChromaDB (top-k)

Retrieve chunks

Return context chunks

Rag prompt

Build RAG prompt (context + query)

Final answer

Call LLM → generate final answer

PHASE 3: Evaluation and Fine-tuning

Objective: measure the retrieval quality and question answering ability of the RAG system using the current embedding..

Evaluation

- Load embeddings + ChromaDB
- Generate QA pairs
- Compute metrics for each QA
- Pick best config
- Save artifacts

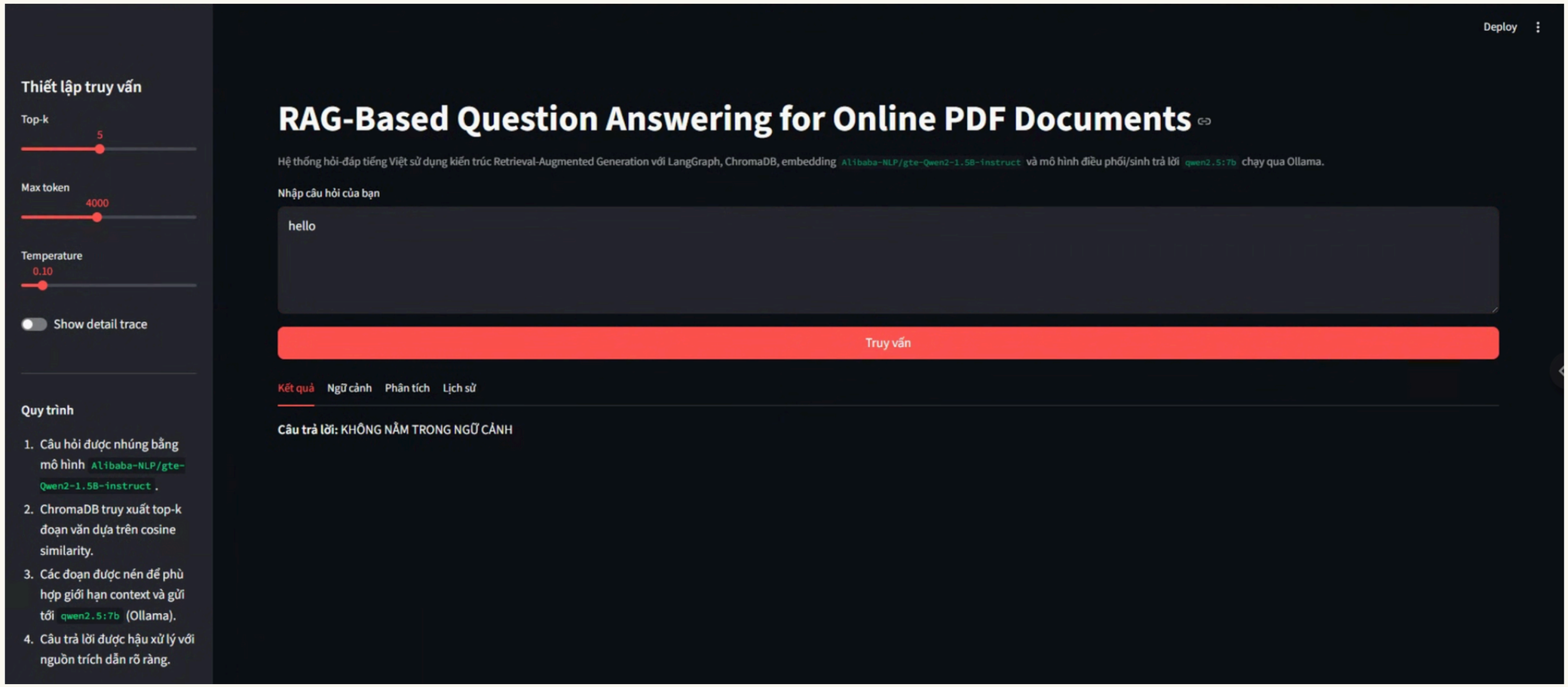
PHASE 3: Evaluation and Fine-tuning

Objective: measure the retrieval quality and question answering ability of the RAG system using the current embedding..

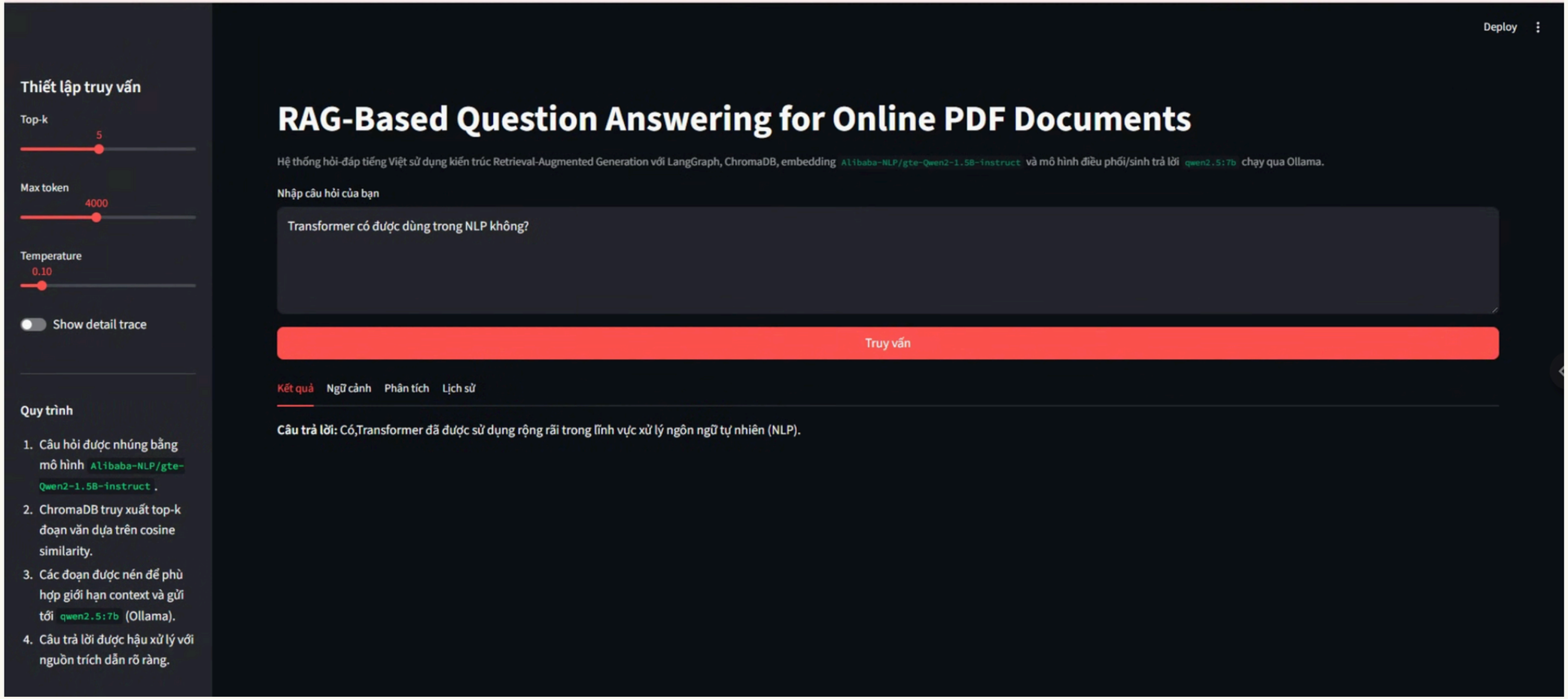
Fine-tuning

1. **Prepare training data**
2. **Fine-tuning**
3. **Save updated embedding model**
4. **Rebuild ChromaDB**
5. **Promote improved model**

RESULT


Deploy
Thiết lập truy vấn
Top-k
5
Max token
4000
Temperature
0.10
Show detail trace
Quy trình
1. Câu hỏi được nhúng bằng mô hình Alibaba-NLP/gte-Qwen2-1.5B-instruct .
2. ChromaDB truy xuất top-k đoạn văn dựa trên cosine similarity.
3. Các đoạn được nén để phù hợp giới hạn context và gửi tới qwen2.5:7b (Ollama).
4. Câu trả lời được hậu xử lý với nguồn trích dẫn rõ ràng.
RAG-Based Question Answering for Online PDF Documents
Hệ thống hỏi-đáp tiếng Việt sử dụng kiến trúc Retrieval-Augmented Generation với LangGraph, ChromaDB, embedding Alibaba-NLP/gte-Qwen2-1.5B-instruct và mô hình điều phối/sinh trả lời qwen2.5:7b chạy qua Ollama.
Nhập câu hỏi của bạn
hello
Truy vấn
Kết quả
Ngữ cảnh
Phân tích
Lịch sử
Câu trả lời: KHÔNG NẰM TRONG NGỮ CẢNH

RESULT


Deploy
Thiết lập truy vấn
Top-k
5
Max token
4000
Temperature
0.10
Show detail trace
Quy trình
1. Câu hỏi được nhúng bằng mô hình Alibaba-NLP/gte-Qwen2-1.5B-instruct.
2. ChromaDB truy xuất top-k đoạn văn dựa trên cosine similarity.
3. Các đoạn được nén để phù hợp giới hạn context và gửi tới qwen2.5:7b (Ollama).
4. Câu trả lời được hậu xử lý với nguồn trích dẫn rõ ràng.
RAG-Based Question Answering for Online PDF Documents
Hệ thống hỏi-đáp tiếng Việt sử dụng kiến trúc Retrieval-Augmented Generation với LangGraph, ChromaDB, embedding Alibaba-NLP/gte-Qwen2-1.5B-instruct và mô hình điều phối/sinh trả lời qwen2.5:7b chạy qua Ollama.
Nhập câu hỏi của bạn
Transformer có được dùng trong NLP không?
Truy vấn
Kết quả
Ngữ cảnh
Phân tích
Lịch sử
Câu trả lời: Có,Transformer đã được sử dụng rộng rãi trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP).

THANK YOU