Final Project

Project: Ask Vietnam: A RAG-based QA System for Vietnamese History

Team 2:

Nguyễn Thanh Tùng (50%)

Trần Quốc Công (50%)

Overview

- Ask Vietnam system is a Retrieval-Augmented Generation (RAG) Question Answering (QA) system designed to answer queries about Vietnam history in Vietnamese.
- It integrates information retrieval with a large language model to provide accurate, context-based responses.

Architecture

- **User Input:** Users enter questions via a Gradio interface.
- **Data Preprocessing:** Loads a Wikipedia dataset, tokenizes text using underthesea, creates LangChain Documents and creates semantic chunks.
- **Retrieval Module:** Semantic (embedding-based) retriever to fetch relevant documents.
- **Generation Module:** Uses the Qwen3-4B-AWQ model with a RAG prompt to generate responses from retrieved context. If retrieved response results are not relevant, LLM can directly generate the answer.
- **Output:** Displays question, answer, and optional context in Gradio.
- **Evaluation:** Retrieval information & Chatbot pipeline

Resource

- Kaggle Notebook
- 2 NVIDIA Tesla T4 GPUs
- Each with 16GB VRAM, totally 32GB

User Query

- The user inputs a question in Vietnamese about Vietnamese history via a Gradio interface.
- An example: “Hãy tóm tắt lịch sử Việt Nam?”

Data Preprocessing

- **Dataset**: A dataset of Vietnamese history (crawled from Wikipedia, stored in vietnam_history_dataset.csv) is loaded using Pandas.
- **Document Creation**: The dataset is converted into Document objects using LangChain, where each document contains a title, content, URL, and metadata (e.g., row_id, source).
- **Text Preprocessing:** Vietnamese text is tokenized and normalized using the underthesea library for better retrieval performance.
- **Chunking**: Semantic chunking with underthesea.sent_tokenize and RecursiveCharacterTextSplitter.

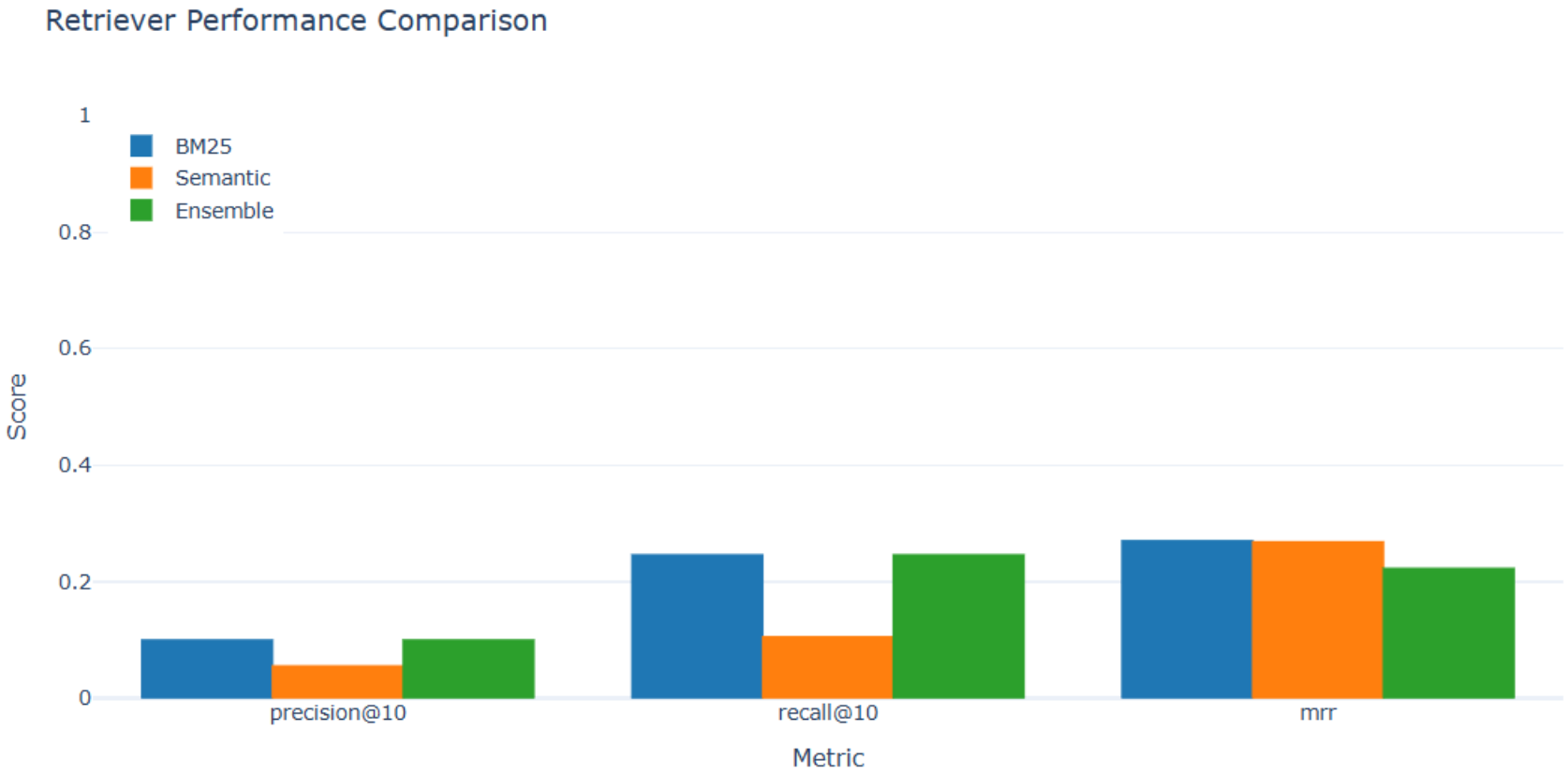
vietnam_history_dataset.csv: https://www.kaggle.com/datasets/tungnguyen1010/vietnam-history-data
Script: https://www.kaggle.com/code/tungnguyen1010/wikipage-crawl

Retrieval Module

- **BM25 Retriever**: Uses the BM25 algorithm for keyword-based retrieval, ranking documents based on term frequency and inverse document frequency.
- **Semantic Retriever:** Utilizes a sentence-transformer model (paraphrase-multilingual-MiniLM-L12-v2) to create embeddings stored in a Chroma vector store, enabling semantic similarity-based retrieval.
- **Ensemble Retriever**: Combines BM25 and Semantic retrievers with weights (0.7 for BM25, 0.3 for Semantic) to balance keyword and semantic search, retrieving the top-k relevant documents (default k=2).

Retrieval Module

- Retrieval test set is created from first 50 chunks.
- The performance of Ensemble Retriever & BM25 Retrieval is better then Semantic Retrieval.



retrieval-testset.json: https://www.kaggle.com/datasets/tungnguyen1010/retrieval-testset

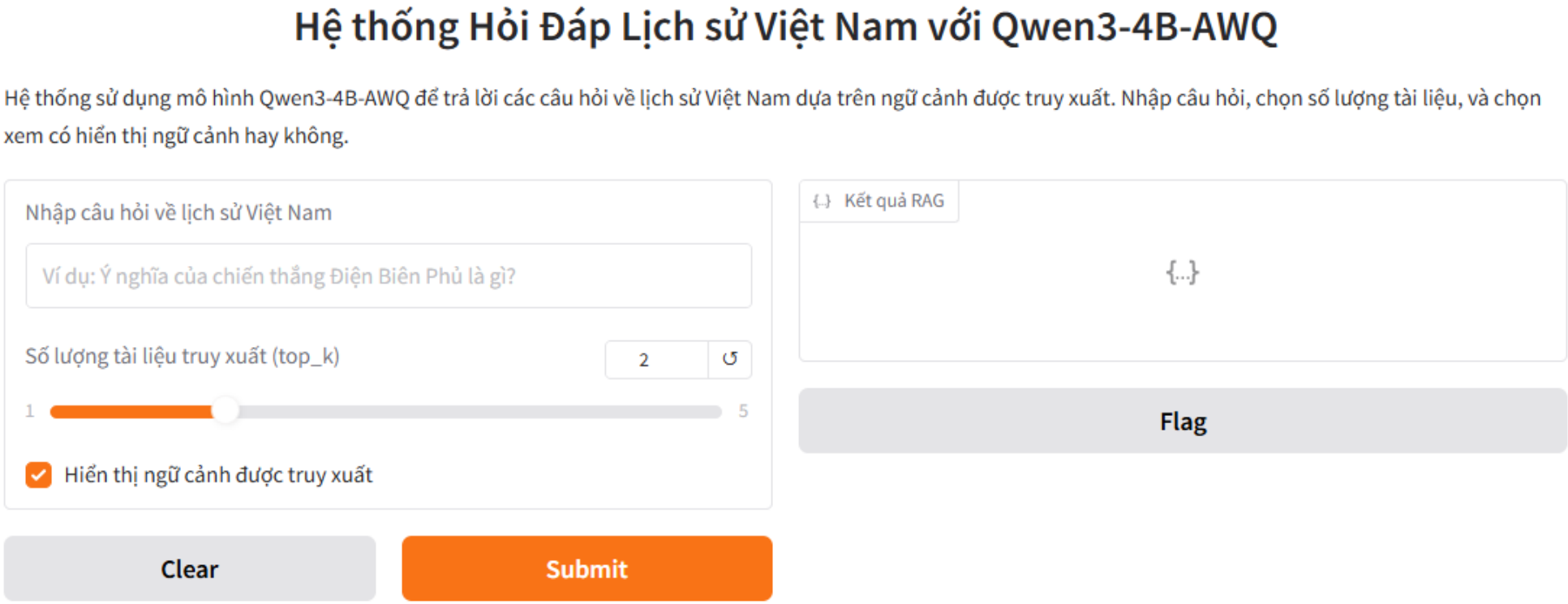
Generation Module

- **LLM (Qwen3-4B-AWQ):** A quantized large language model (Qwen3-4B-AWQ) is used via the vLLM engine for efficient inference. It is configured with sampling parameters (e.g., temperature=0.6, top_p=0.95) and supports reasoning mode.
- **RAG Prompt Template:** A LangChain PromptTemplate formats the retrieved documents (context) and user question into a prompt for the LLM. The prompt instructs the model to prioritize context but use its knowledge if the context is insufficient.
- **Token Management:** The context is truncated to fit within a 7,000-token limit to avoid exceeding the LLM's maximum input length (8,192 tokens).

Output

- The LLM generates a response based on the formatted prompt.
- The Gradio interface displays the question, answer, and optionally the retrieved context in a JSON format.
- Additional validation ensures the question is in Vietnamese using a regex pattern for Vietnamese diacritics.

User Interface



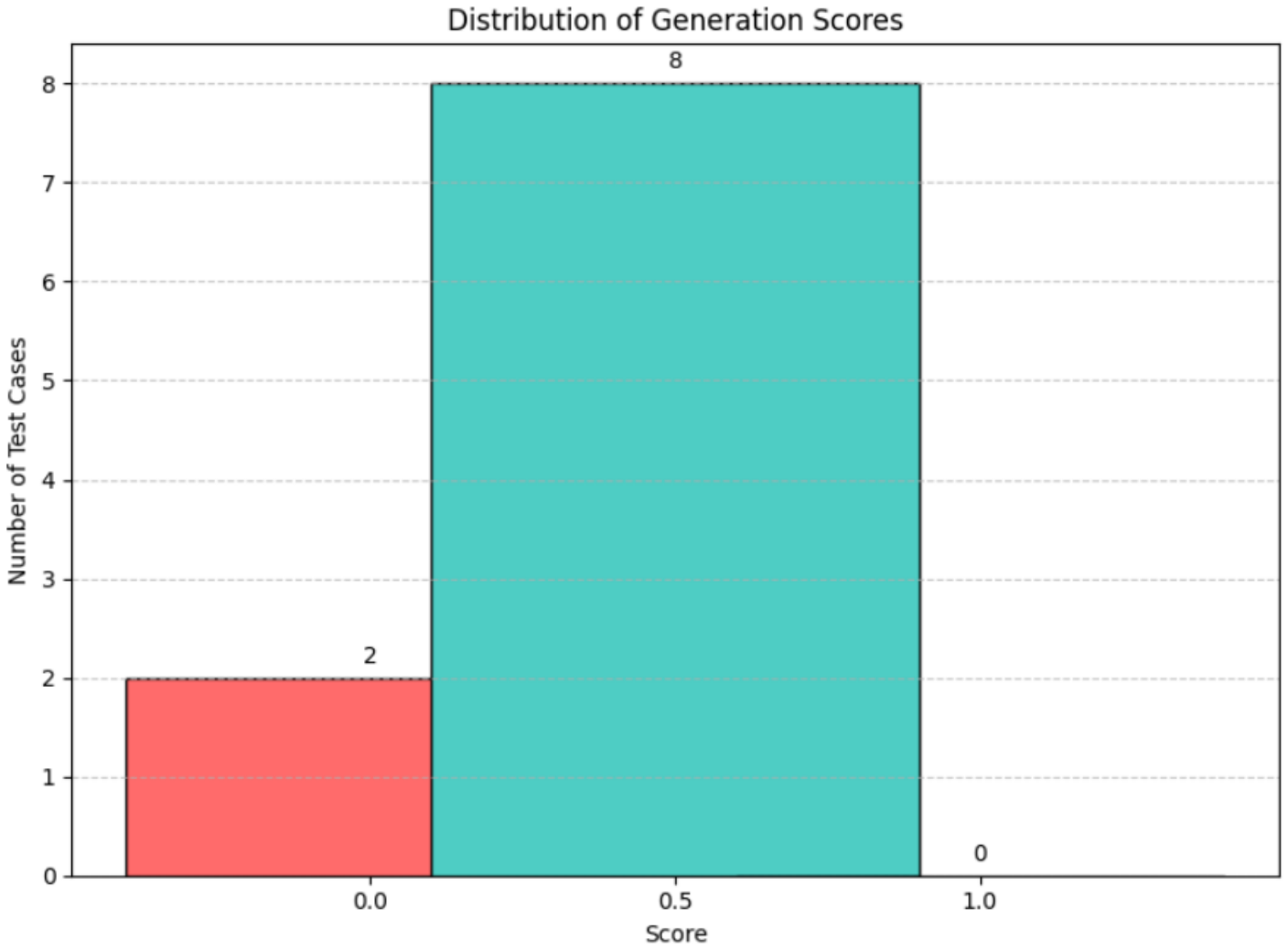
- Built with Gradio, allowing users to input questions, adjust the number of retrieved documents (top_k), and toggle context display.
- The interface is themed and includes a title and description for user guidance.

Evaluation

- **Input:** A question and reference answer (ground truth) from testset, and a generated answer from rag_chain.
- **Keyword Extraction:** Uses word_tokenize to split answers into words and converts them to lowercase for case-insensitive comparison.
- **Overlap Calculation:** Computes the ratio of common keywords between the ground truth and generated answers.
- **Scoring:**
  - Score = 1.0 if >80% of reference answer (ground truth) are in the generated answer.
  - Score = 0.5 if >30% but ≤80%.
  - Score = 0.0 otherwise.

Evaluation

- Score Distribution: {0.0: 2, 0.5: 8, 1.0: 0}
- Observed latency time: averagely 18 second/question



LLMs & Industry Practices 02 – LLMs 02

Evaluation

```
Adding requests: 100% |████████████████████| 1/1 [00:00<00:00, 59.67it/s]
Processed prompts: 100% |████████████████████| 1/1 [00:10<00:00, 10.97s/it, est. speed input: 238.28 toks/s, output: 36.38 toks/s]

Question: Hai Bà Trưng khởi nghĩa trong hoàn cảnh nào và kết quả ra sao?

Grouth Truth: Hai Bà Trưng, tức Trưng Trắc và Trưng Nhị, khởi nghĩa vào năm 40 SCN chống lại ách đô hộ của nhà Hán. Nguyên nhân trực ti
ếp là do thái thú Tô Định giết Thi Sách, chồng Trưng Trắc. Sau khi giành thắng lợi ban đầu, Hai Bà Trưng lên làm vua, xây dựng triều đi
nh đóng đô ở Mê Linh. Tuy nhiên, đến năm 43 SCN, Mã Viện dẫn quân Hán sang đàn áp, Hai Bà Trưng thất bại và tuẫn tiết trên sông Hát. Cu
ộc khởi nghĩa của Hai Bà Trưng dù không thành công lâu dài nhưng đã khẳng định ý chí độc lập, tinh thần quật cường của người Việt.

Generated answer: Hai Bà Trưng khởi nghĩa xảy ra vào thời kỳ nhà Hán, cụ thể là vào khoảng năm 40 trước Công nguyên, thuộc thời kỳ cuối
của triều đại Hán. Đây là một cuộc khởi nghĩa nổi bật trong lịch sử Việt Nam, do hai anh em là Trưng Trắc và Trưng Nhị lãnh đạo. Hai Bà
Trưng là người dân tộc Âu Dương (nay là vùng hiện nay thuộc tỉnh Hà Nội và Hải Dương), sống ở vùng đồng bằng sông Hồng.

**Hoàn cảnh:**
Khi thời Hán còn yếu, các vua Hán đã mở rộng ảnh hưởng đến vùng Việt Nam. Vua Hán Thái Tông (tên gọi khác là Hán Thái Tông) đã phát độn
g chiến tranh xâm lược vùng Việt Nam, bắt giữ và cưỡng bức người dân địa phương. Người dân ở vùng đồng bằng sông Hồng bị đói khổ, đói n
ghèo, và chịu nhiều bất công. Hai Bà Trưng là những người chịu nhiều khổ sở, chịu nhiều áp bức, và họ đã lên tiếng phản đối.

**Kết quả:**
Cuộc khởi nghĩa của Hai Bà Trưng đã giành được sự ủng hộ mạnh mẽ từ người dân địa phương. Họ đã đánh đuổi các quan chức Hán, giải phóng
vùng đất của mình. Cuộc khởi nghĩa đã diễn ra trong thời gian ngắn nhưng có tính chất cách mạng, thể hiện tinh thần yêu nước và sự kiên
quyết chống lại áp bức, bóc lột của chính quyền Hán.

Cuộc khởi nghĩa của Hai Bà Trưng được coi là một trong những cuộc khởi nghĩa cách mạng tiêu biểu trong lịch sử Việt Nam, thể hiện tinh
thần yêu nước và ý chí kiên cường của nhân dân ta. Dù không kéo dài lâu, nhưng nó đã để lại dấu ấn sâu sắc trong lòng dân tộc, được ghi
nhớ và tôn vinh trong lịch sử.

Score: 0.5
```

Query: Hai Bà Trưng khởi nghĩa trong hoàn cảnh nào và kết quả ra sao? -> score: 0.5

Additional Information (experiments)

- Database: 2GB data in pdf format -> latency problem (averagely > 90 second/question) (root cause from Retrieval Information), may need effective data preprocessing.
- The Agent RAG
  - Use an open-source API search <DuckDuckGoSearchRun()
  - LLM uses for reasoning and tool selection.

  -> Got latency problem (averagely 150 second/question)

Limitations

- Latency time: averagely 18 second/question -> big room for improvement.
- Database is small (1000 Wiki files, 12.21 MB).