Xây dựng PDF Chatbot tiếng Việt với RAG

Người thực hiện: Đỗ Đức Vĩnh
Nguyễn Mạnh Tuấn

Các nội dung chính

01 Giới thiệu bài toán

02 Phương pháp thực hiện

03 Kết luận

01

Giới thiệu bài toán

Mục tiêu và ý nghĩa

Mục tiêu: Phát triển một công cụ chat thông minh có khả năng trả lời câu hỏi của người dùng, dựa trên thông tin từ các bản PDF cung cấp.

Ý nghĩa: Giúp người dùng tiết kiệm thời gian, nâng cao hiệu quả học tập và làm việc thông qua việc tìm kiếm thông tin nhanh chóng.

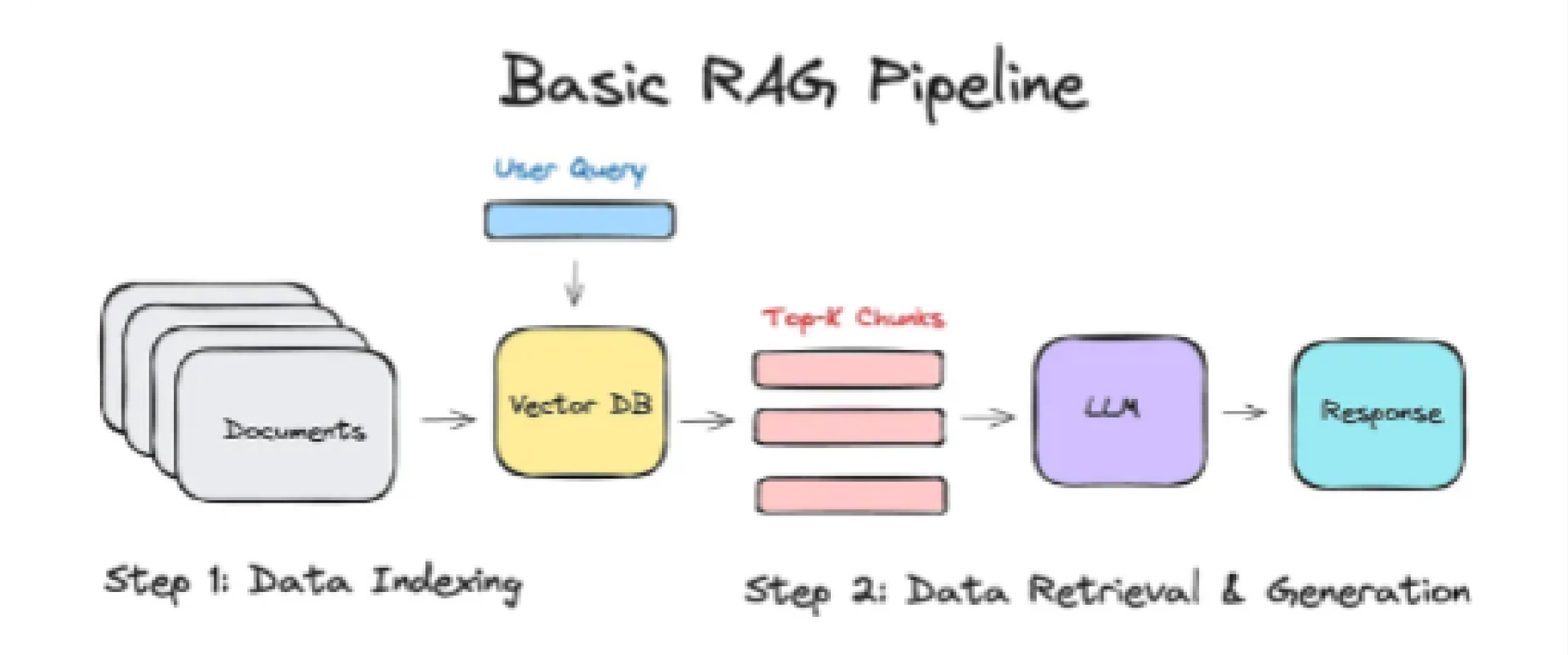
Hoạt động: Người dùng truyền vào các PDF và bắt đầu hỏi, công cụ chat sẽ đưa ra câu trả lời dựa trên thông tin tìm được từ các PDF đó.

02

Phương pháp thực hiện

Retrival-Augmented Generation

Ý tưởng chính: Truy xuất thông tin từ kho kiến thức, sau đó đưa vào mô hình ngôn ngữ lớn (LLM) nhằm sinh ra những câu trả lời chính xác và tự nhiên.



Lợi ích của RAG

- Giảm thiểu ảo giác: RAG cung cấp cho LLM thêm thông tin từ các nguồn dữ liệu đã truy xuất. LLM sẽ có nền tảng thông tin cụ thể để dựa vào khi sinh ra câu trả lời, thay vì chỉ dựa vào dữ liệu đã được học trước đó).

- Giảm thiểu chi phí: RAG cho phép sử dụng thông tin từ các tài liệu có sẵn mà không cần huấn luyện lại mô hình ngôn ngữ lớn.

- Xử lý tốt các đề tài khó: RAG có khả năng xử lý các câu hỏi đặc thù và phức tạp hơn bằng cách truy xuất thông tin từ các tài liệu chuyên ngành, như tài liệu mật của các công ty...(thứ sẽ không thể có trong dữ liệu huấn luyện của các LLMs).

Phân tích bài toán

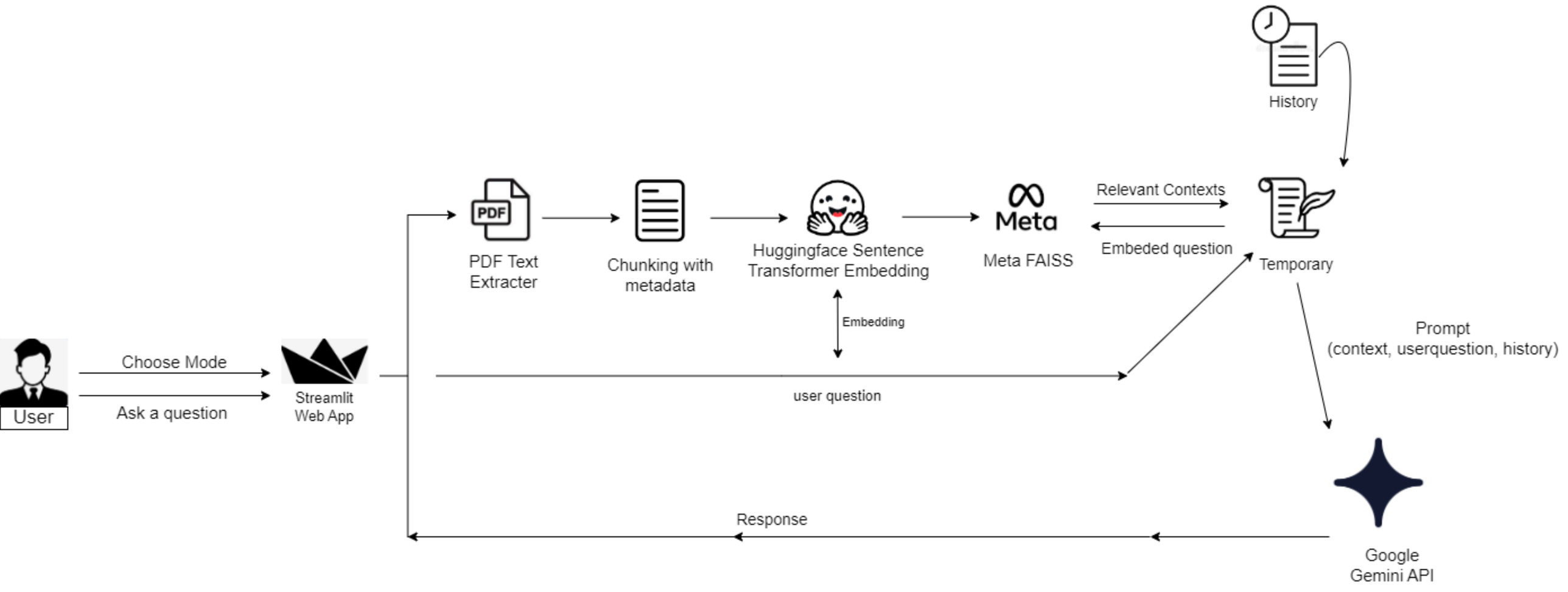
- Đầu vào là PDF, cần chuyển nó sang dạng văn bản để có thể xử lý.

- Ta cần đưa văn bản thông tin vào LLM để sinh ra câu trả lời, mà đầu vào của LLM cần giảm thiểu càng ít token càng tốt (giảm chi phí) => **tách văn bản gốc thành các đoạn nhỏ hơn, tìm kiếm 2 đoạn văn bản liên quan nhất đến truy vấn.**

- Người dùng có thể muốn hỏi thông tin về câu hỏi trước, và cũng cần xem nguồn của các câu trả lời. => **Thêm hiển thị context, nguồn, lưu trữ lịch sử chat.**

- Người dùng có thể muốn hỏi đáp thông thường để so sánh thông tin bên ngoài tài liệu với thông tin vừa tìm được. => **Thiết kế 2 chế độ chat**

Hệ thống RAG



Hình minh họa hệ thống RAG

Tạo vector database

PDF (text format)

PyPDF Loader

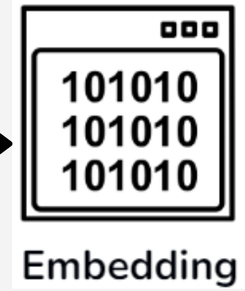
check if exist

chunk_size = 512

overlap = 256

halong_embedding

768-dimensional dense vectors



Indexing and storing with FAISS

FAISS

PyPDF Loader

Original text

Tìm kiếm tương đồng ngữ nghĩa

Cách thức: Sử dụng các mô hình embedding (chuyển đổi text thành vector) sau đó so sánh mức độ tương đồng giữa vector query và vector context, nếu khoảng cách càng gần nhau thì chúng càng liên quan đến nhau (Euclidean Distance).

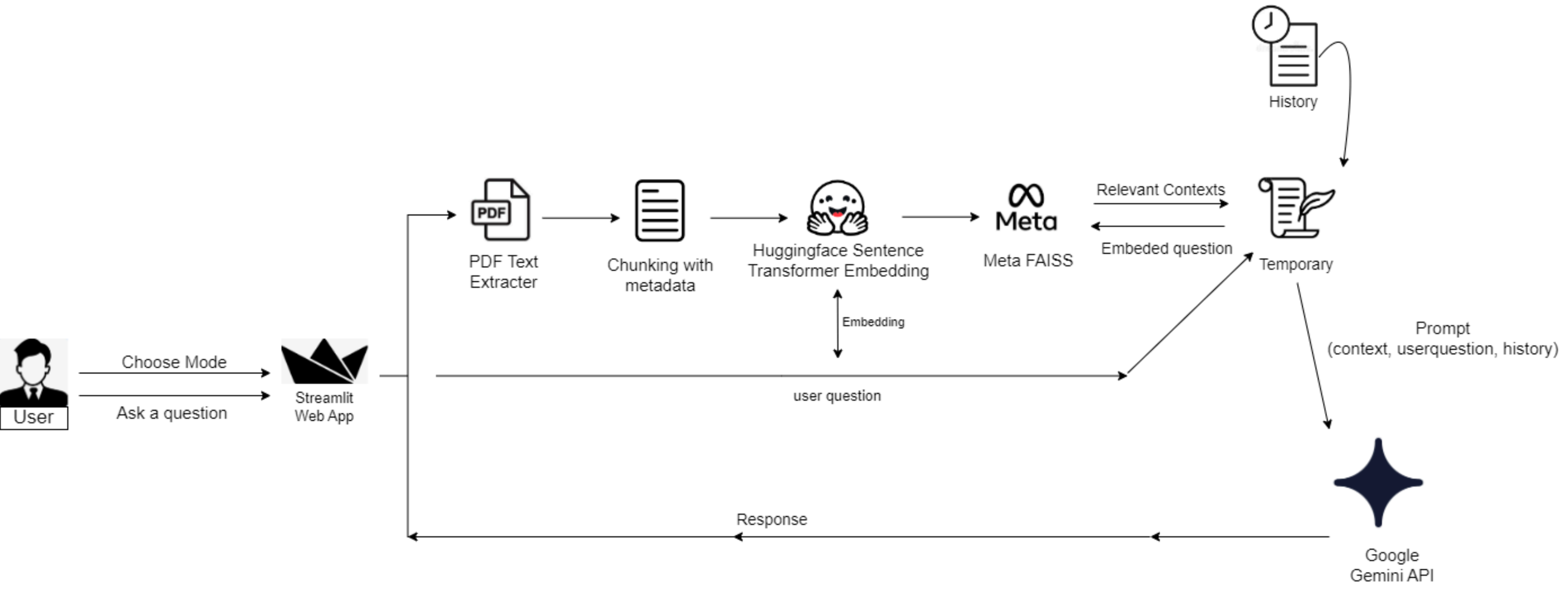
Mô hình embedding sử dụng

Mô hình embedding được sử dụng trong đề tài là hieuu/halong_embedding. Đây là một mô hình sentence transformers được anh Hiếu Ngô - Senior AI của MOMO tinh chỉnh trên một bộ data chất lượng để phục vụ cho tác vụ embed văn bản tiếng việt.

```
1 df_spearman_styled
```

	STS-B	STS12	STS13	STS14	STS15	STS16	STS-Sickr	Mean
VoVanPhuc/sup-SimCSE-VietNamese-phobert-base	**81.430000**	**76.510000**	79.190000	**74.910000**	81.720000	76.570000	76.450000	**78.111429**
keepitreal/vietnamese-sbert	80.160000	69.080000	**80.990000**	73.670000	**82.810000**	74.300000	73.400000	76.344286
nampham1106/bkcare-embedding	79.880000	72.660000	78.350000	70.740000	77.610000	75.140000	**77.380000**	75.965714
intfloat/multilingual-e5-base	76.750000	69.200000	67.260000	65.730000	79.350000	**77.530000**	72.400000	72.602857
hiieu/halong_embedding	74.540000	62.750000	71.410000	65.510000	78.660000	75.350000	70.860000	71.297143
bkai-foundation-models/vietnamese-bi-encoder	72.160000	63.860000	71.820000	66.200000	78.620000	74.240000	70.870000	71.110000

Hệ thống RAG

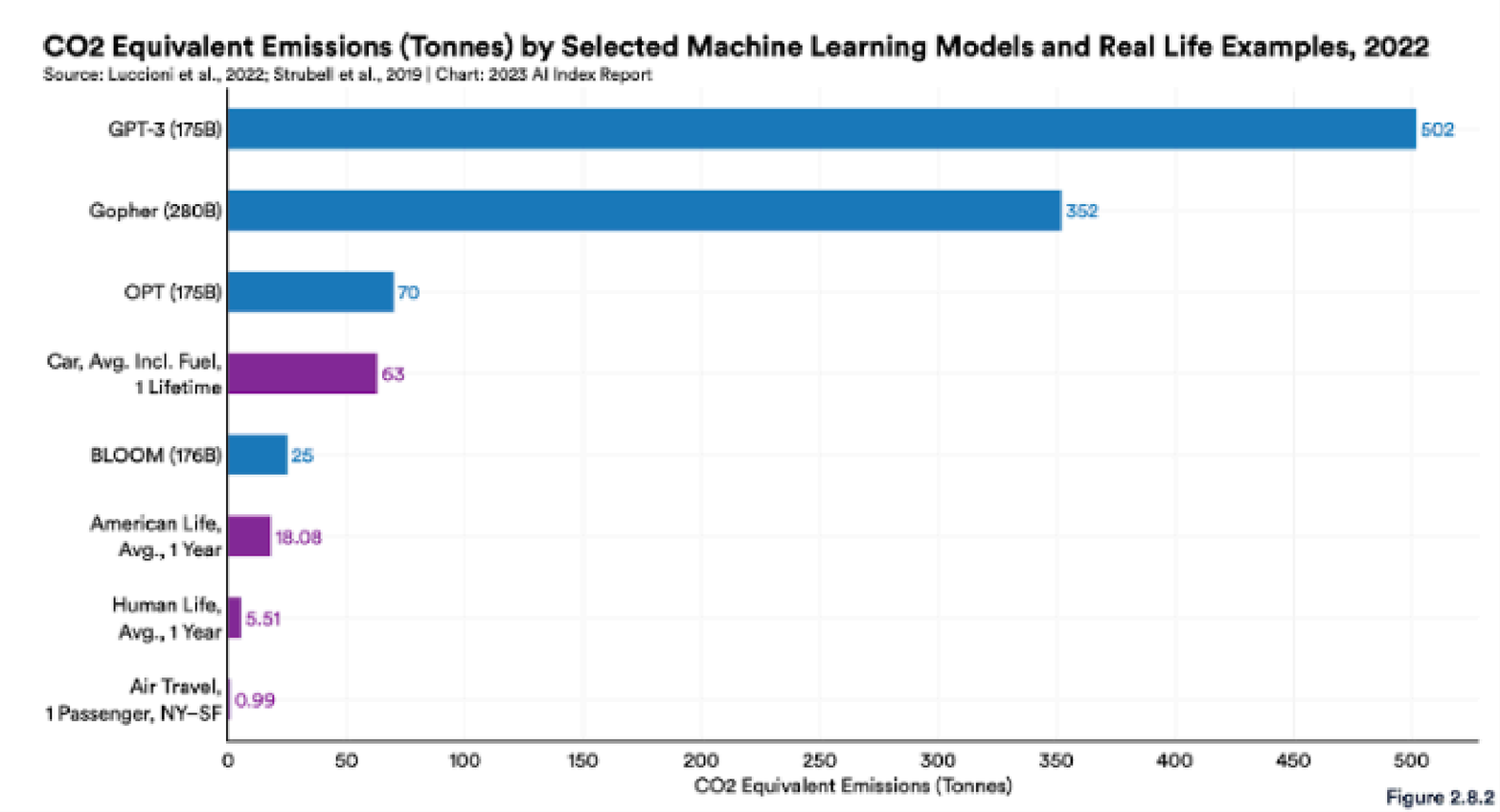


Hình minh họa hệ thống RAG

Mô hình ngôn ngữ lớn

API Gemimi-1.5-flash:

- Tốc độ phản hồi nhanh, hiểu biết ngôn ngữ tự nhiên tốt.
- Tiết kiệm tài nguyên, vì không đủ tài nguyên, cơ sở hạ tầng để vận hành LLM.
- Dễ dàng tích hợp, có hỗ trợ phiên bản miễn phí.



Demo

Hướng phát triển

- Tối ưu hóa hiệu suất để xử lý khối lượng lớn tài liệu và truy vấn đồng thời
- Mở rộng hỗ trợ cho các định dạng tài liệu khác ngoài PDF
- Cải thiện chất lượng phản hồi
- Tạo trang web hoặc app cho phép người dùng đăng nhập vào để có thể lưu trữ và bảo mật các tài liệu tải lên và lịch sử trò chuyện lâu dài

CẢM ƠN MỌI NGƯỜI!