# HUST

**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

# Xây dựng trợ lý ảo hỗ trợ tư vấn tuyển sinh HUST


Giảng viên hướng dẫn: PGS.TS Lê Chí Ngọc

Trình bày: Sinh viên Vũ Thị Hằng

MSSV: 20206136


ONE LOVE. ONE FUTURE.

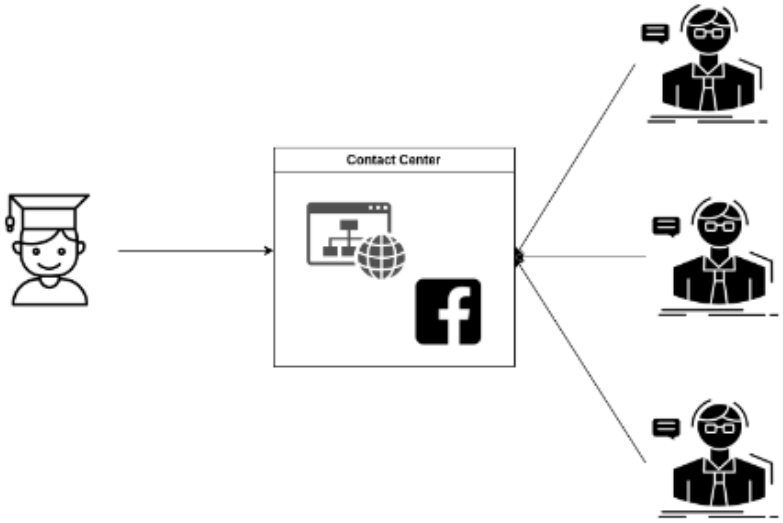
# Mục lục

1. Mục tiêu và nội dung thực hiện
2. Xây dựng bộ dữ liệu
3. Kiến trúc của hệ thống chatbot
4. Đánh giá hệ thống
5. Demo chương trình
6. Kết luận

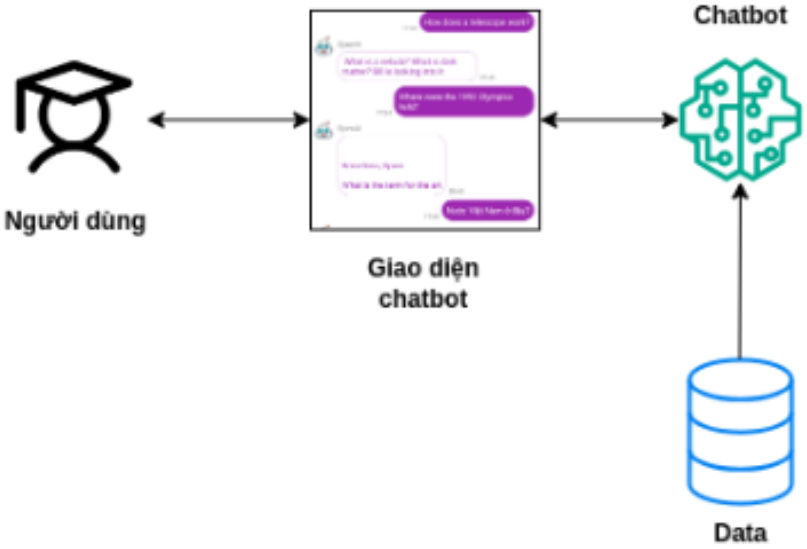
# PHẦN 1: MỤC TIÊU VÀ NỘI DUNG THỰC HIỆN

# Lý do chọn đề tài

- Nhu cầu tư vấn tuyển sinh tăng cao, hệ thống hiện tại còn hạn chế.
- Sự phổ biến của AI và chatbot trong việc tự động hóa quy trình, cải thiện hiệu suất tư vấn.
- Chatbot giúp giảm tải công việc, cung cấp thông tin nhanh chóng, chính xác, và tiện lợi




Hệ thống contact center



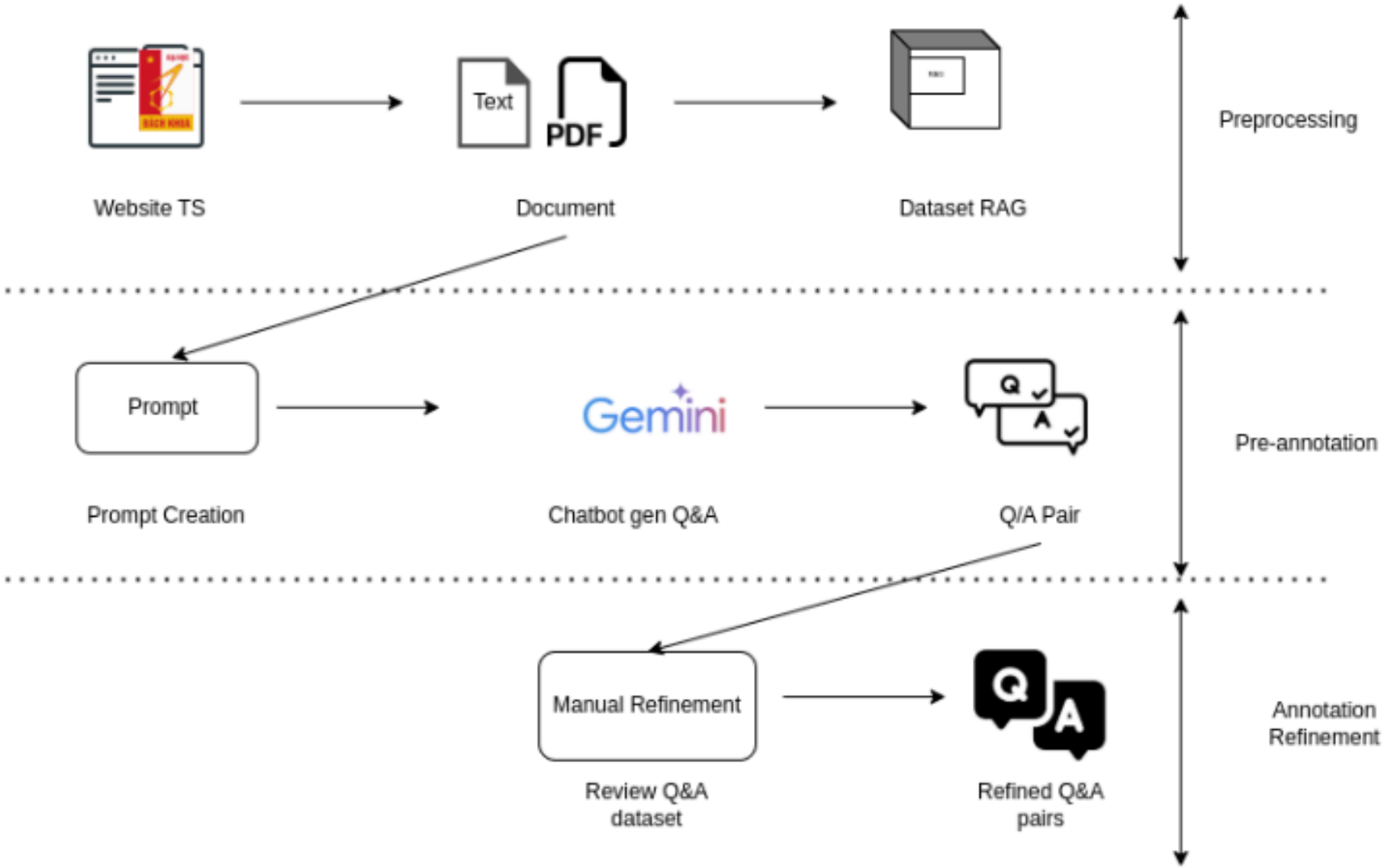

Hệ thống tư vấn TS đề xuất

# Mục tiêu và nhiệm vụ

| Mục tiêu | Nhiệm vụ |
| --- | --- |
| Xây dựng 1 trợ lý ảo hỗ trợ tư vấn tuyển sinh cho phụ huynh, học sinh | Nghiên cứu lý thuyết: Deep Learning, NLP, LLM, Kỹ thuật RAG, Fine-tuning, Pre-trained model,... |
| Xây dựng giao diện chatbot có thể chọn được model sinh câu trả lời | Thử nghiệm và đánh giá hiệu quả của các mô hình trên tập dữ liệu tuyển sinh |
| | Phát triển chatbot sử dụng RAG và Fine-tuning |

# PHẦN 2: XÂY DỰNG BỘ DỮ LIỆU

# Quy trình xây dựng bộ dữ liệu



# Bộ dữ liệu

## Bộ dữ liệu cho RAG

| | |
|---|---|
| 1.thong_tin_chung | 7 items |
| 2. de_an_tuyen_sinh | 7 items |
| 3. xet_tuyen_tai_nang | 20 items |
| 4. diem_chuan_tuyen_sinh | 22 items |
| 5. ky_thi_danh_gia_tu_duy | 34 items |
| 6. xac_thuc_chung_chi_ngoai_ngu | 2 items |
| 7. huong_nghiep | 9 items |

## Bộ dữ liệu cho Fine-tuning

***Context:*** *Số lượng chương trình đào tạo: 64 chương trình, trong đó: • Số lượng chương trình đại trà (chương trình chuẩn): 36 chương trình; • Số lượng chương trình chất lượng cao - Elitech: 23 chương trình; • Số lượng chương trình PFIEV: 02 chương trình; • Số lượng chương trình liên kết đào tạo quốc tế: 03 chương trình.*

***Question:*** *Trường mình có tổng cộng bao nhiêu chương trình đào tạo và chi tiết từng loại chương trình như thế nào ạ?*

***Answer:*** *Trường mình có tổng cộng 64 chương trình đào tạo, trong đó có 36 chương trình đại trà, 23 chương trình chất lượng cao Elitech, 2 chương trình PFIEV và 3 chương trình liên kết quốc tế em nhé.*

***Context:*** *Những lưu ý khi thanh toán • Trường hợp thí sinh đã hoàn thành thao tác thanh toán, tài khoản đã trừ tiền nhưng hệ thống chưa hiển thị trạng thái Đã thanh toán, thí sinh gửi email tới địa chỉ tuyensinh@hust.edu.vn, cung cấp địa chỉ email, số CCCD/CMND, mật khẩu đăng nhập tài khoản, ảnh chụp minh chứng bị trừ tiền, phương thức thanh toán, số điện thoại,... hoặc liên hệ số hotline 024 3 868 3408 (trong giờ hành chính) để được hỗ trợ.*

# Mô tả dữ liệu

## Bộ dữ liệu cho RAG

**Dung lượng**: 108 file docx, pdf - 34,2 MB

**Nội dung**: Dữ liệu về tuyển sinh Đại học gồm: *thông tin chung, đề án tuyển sinh, xét tuyển tài năng, kỳ thi đánh giá tư duy, điểm chuẩn tuyển sinh, xác thực chứng chỉ ngoại ngữ, hướng nghiệp.*

**Nguồn thu thập**: *https://ts.hust.edu.vn/p/dai-hoc*

**Thời điểm thu thập**: 11/12/2024

## Bộ dữ liệu cho Fine-tuning

**Dung lượng**: 1941 cặp ngữ cảnh, câu hỏi, câu trả lời - 949 KB

**Nội dung**: Các cặp ngữ cảnh, câu hỏi, câu trả lời dựa trên bộ dữ liệu RAG, nội dung là những câu hỏi thường gặp khi hỏi đáp về tuyển sinh.

**Nguồn thu thập**: *https://aistudio.google.com/prompts/new_chat*

**Thời điểm thu thập**: 11/12/2024

# Thống kê bộ dữ liệu

## Bộ dữ liệu cho RAG

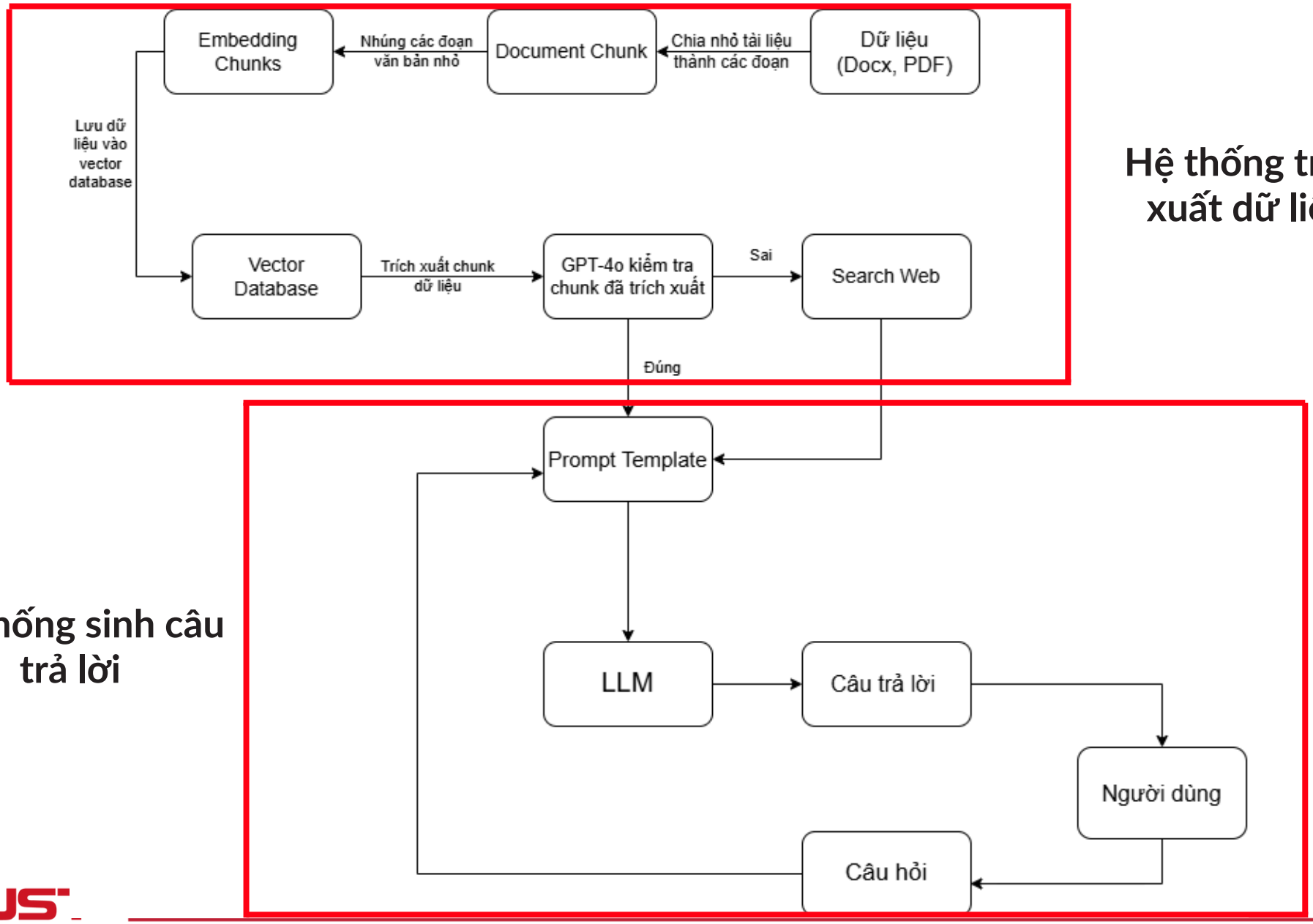
| Tên folder | Số lượng file DOCX | Số lượng file PDF |
|---|---|---|
| Thông tin chung | 5 | 2 |
| Đề án tuyển sinh | 2 | 5 |
| Xét tuyển tài năng | 18 | 2 |
| Điểm chuẩn tuyển sinh | 22 | 0 |
| Kỳ thi đánh giá tư duy | 32 | 1 |
| Xác thực chứng chỉ ngoại ngữ | 2 | 0 |
| Hướng nghiệp | 9 | 0 |

## Bộ dữ liệu cho Fine-tuning

| | Train set | Test set | % |
|---|---|---|---|
| Số lượng câu hỏi | 1746 | 194 | 80% |
| Số lượng câu trả lời | 1746 | 194 | 20% |

# PHẦN 3: KIẾN TRÚC HỆ THỐNG CHATBOT

# Kiến trúc hệ thống




Hệ thống trích xuất dữ liệu

Hệ thống sinh câu trả lời

# Hệ thống truy xuất dữ liệu

## Phân loại miền trích xuất

- Dữ liệu được chia thành 7 miền: ***Thông tin chung, Đề án tuyển sinh, Xét tuyển tài năng, Điểm chuẩn tuyển sinh, Kỳ thi đánh giá tư duy, Xác thực chứng chỉ ngoại ngữ, Hướng nghiệp***

- Dữ liệu tại mỗi miền được chia thành các chunk có độ dài **450 tokens và chunk_overlap = 50**

- Dữ liệu sau khi phân loại miền và chia chunk sẽ được lưu vào **ChromaDB** để quản lý các vector

# Hệ thống truy xuất dữ liệu

## Lấy dữ liệu từ database

- Lấy ra ***top 3 vector*** có khoảng cách gần nhất với câu hỏi
- Với mỗi chunk thứ i được lấy ra, lấy thêm ***vector thứ i-1 và i+1*** để đưa vào context của LLM

**Ví dụ**

Trường mình là trường gì?

Trường mình là Đại học Bách Khoa Hà Nội, một trong những trường đại học hàng đầu tại Việt Nam, chuyên đào tạo kỹ sư và các chuyên gia trong nhiều lĩnh vực khác nhau. Nếu bạn cần thêm thông tin về các chương trình đào tạo hay tuyển sinh, hãy cho tôi biết nhé!

**Nguồn tham khảo:**

**Q&A Database**

Vector 1 (Score: 0.8259)

Trường mình có tổng cộng bao nhiêu chương trình đào tạo và chi tiết từng loại chương trình như thế nào ạ? Trường mình có tổng cộng 64 chương trình đào tạo, trong đó có 36 chương trình đại trà, 23 chươ...

Vector 2 (Score: 0.8214)

Tin học của trường ở đâu ạ? Thầy Ngô Văn Hiền là phó giáo sư của bộ môn Tin học ở trường mình em nhé. 1. Tuyển sinh đào tạo vừa làm vừa học đối tượng tốt nghiệp THPT 1.1. Đối tượng, điều kiện tuyển si...

Vector 3 (Score: 0.8169)

trường mình em nhé. 1.3. Phương thức tuyển sinh (thi tuyển, xét tuyển hoặc kết hợp thi tuyển và xét tuyển): xét tuyển Phương thức 1: Xét tuyển dựa trên kết quả học tập THPT Trường mình có những phương...

# Hệ thống truy xuất dữ liệu

## Đánh giá nội dung trích xuất

Sử dụng mô hình gpt-4o để đánh giá các vector được trích xuất có đúng với câu hỏi của người dùng hay không.

- **Nếu đúng**, lấy các vector đã trích xuất để làm ngữ cảnh cho LLM
- **Nếu sai**, thực hiện tìm kiếm trên web và đưa ra URL thông tin cho người dùng

# Hệ thống truy xuất dữ liệu

## Tìm kiếm web trích xuất url

- Sử dụng **API Google Custom Search** để tìm kiếm thông tin.
- Chỉ thực hiện tìm kiếm trên website : ***https://ts.hust.edu.vn/*** để tránh việc tìm kiếm thông tin không chính thống

## Mô hình GPT-4o-mini

GPT-4o-mini là phiên bản nhẹ gọn hơn của GPT-4o, tối ưu về tốc độ và hiệu quả. Phù hợp với những tài nguyên giới hạn như thiết bị nhỏ hoặc triển khai nhanh.

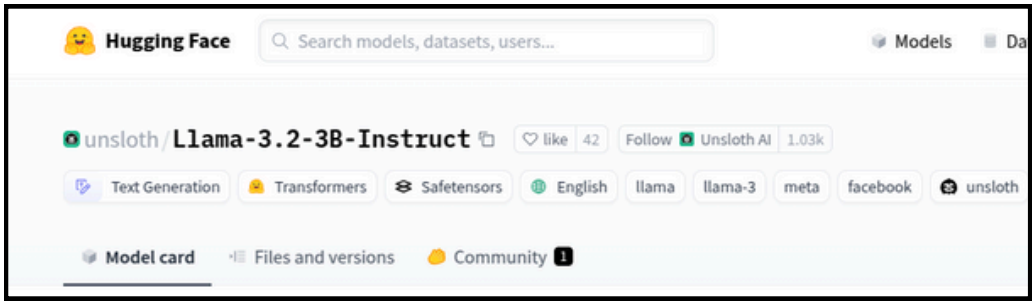
**Đặc Điểm Nổi Bật**

- Hiệu quả cao: Tối ưu hiệu suất về sinh câu trả lời nhanh chóng.
- Chi phí thấp: Tiết kiệm tài nguyên hơn so với các mô hình lớn.
- Dễ triển khai: Có thể sử dụng trong nhiều bài toán thực tế.

GPT-4o mini

# Hệ thống sinh câu trả lời

## Fine - tuning mô hình



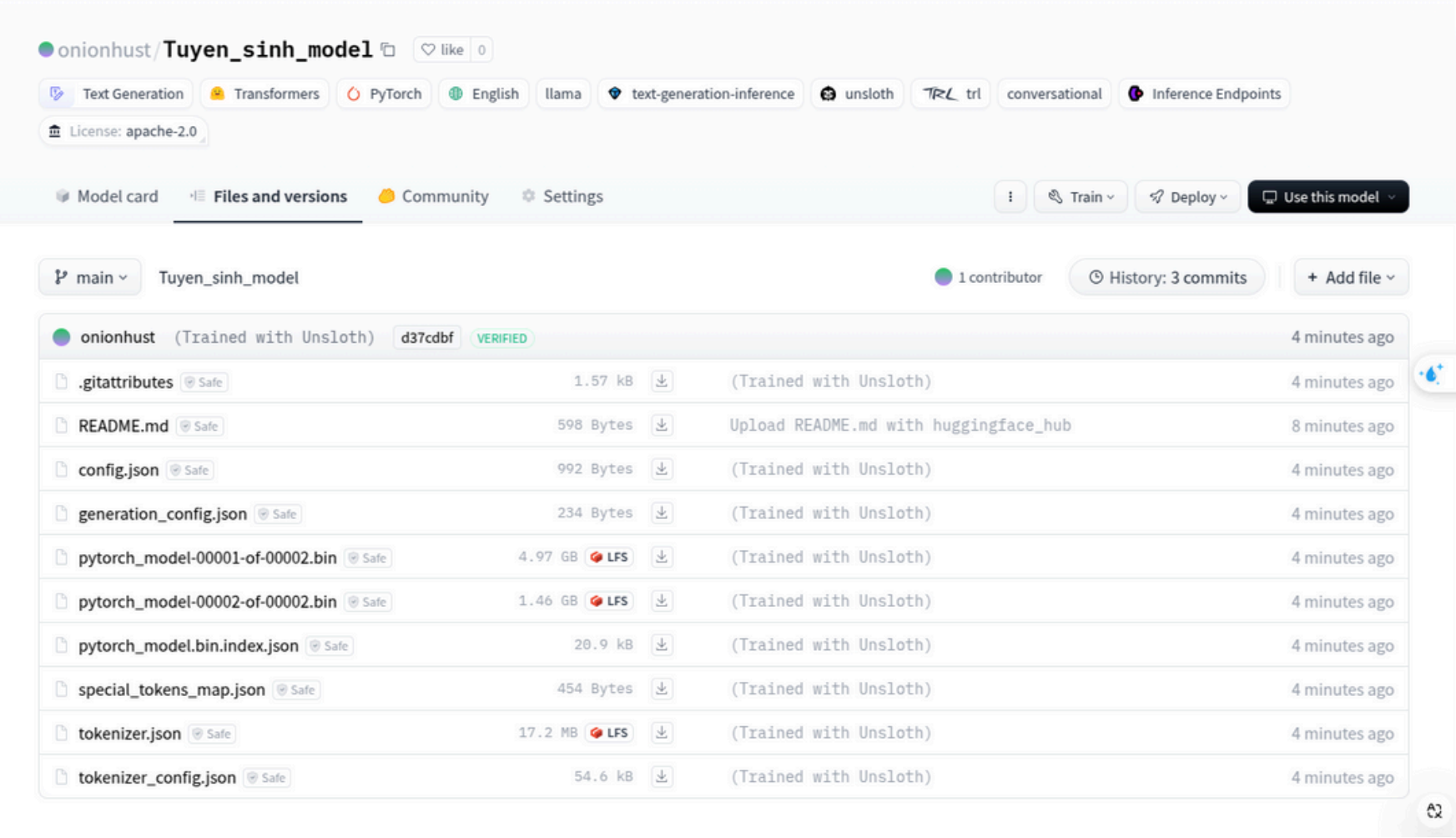

**Base model**

```
model = FastLanguageModel.get_peft_model(
    model,
    r=16,
    target_modules=[
        "q_proj",
        "k_proj",
        "v_proj",
        "o_proj",
        "gate_proj",
        "up_proj",
        "down_proj",
    ],
    lora_alpha=16,
    lora_dropout=0,
    bias="none",

    use_gradient_checkpointing="unsloth",
    random_state=3407,
    use_rslora=False,
    loftq_config=None,
)

Unsloth 2024.12.12 patched 28 layers with 28 QKV layers, 28 O layers and 28 MLP layers.
```

**LoRA Adapter**

# Cài đặt

## Fine - tuning mô hình



# PHẦN 4: ĐÁNH GIÁ HỆ THỐNG

# Phương pháp đánh giá

## Chỉ số đánh giá

$$\text{Precision}_{LCS} = \frac{LCS(X,Y)}{|X|}[7]$$

$$\text{Recall}_{LCS} = \frac{LCS(X,Y)}{|Y|}[7]$$

$$\text{F1}_{LCS} = \frac{(1+\beta^2)\cdot \text{Precision}_{LCS}\cdot \text{Recall}_{LCS}}{(\beta^2\cdot \text{Precision}_{LCS}) + \text{Recall}_{LCS}}[7]$$

*Trong đó*:

- $LCS(X,Y)$ là độ dài của dãy con chung dài nhất giữa văn bản được tạo $X$ và văn bản tham chiếu$Y$
- $|X|$ là chiều dài của tạo ra văn bản và $|Y|$ là độ dài của văn bản tham chiếu.
- Thông số $\beta$ được đặt thành 1, làm cho điểm $F1$ trở thành giá trị trung bình hài hòa của độ chính xác và khả năng thu hồi

# Đánh giá hệ thống

## Trích xuất dữ liệu

| Miền | Số câu trích đúng | Độ chính xác Top 1 |
|---|---|---|
| Thông tin chung | 45/50 | 90% |
| Đề án tuyển sinh | 30/50 | 60% |
| Xét tuyển tài năng | 38/50 | 76% |
| Điểm chuẩn tuyển sinh | 30/50 | 60% |
| Kỳ thi đánh giá tư duy | 43/50 | 86% |
| Xác thực chứng chỉ ngoại ngữ | 38/50 | 76% |
| Hướng nghiệp | 10/10 | 100% |
| **Trung bình** | **234/310** | **75.48%** |

## Sinh dữ liệu

| Model | Precision | Recall | F1 |
|---|---|---|---|
| ChatGPT4o | 0.6691 | 0.7493 | 0.7194 |
| RAG + GPT-4o-mini | 0.6782 | 0.8395 | 0.7273 |
| RAG + Fine-tuned LLama-3-8B-Instruct | 0.7973 | 0.7554 | 0.7327 |

| Mô hình | Số câu trả lời đúng | Phần trăm |
|---|---|---|
| ChatGPT4o | 38/50 | 76% |
| RAG + GPT-4o-mini | 44/50 | 88% |
| RAG + Fine-tuned LLama-3-8B-Instruct | 46/50 | 92% |

# Nhận xét

## Hệ thống trích xuất dữ liệu

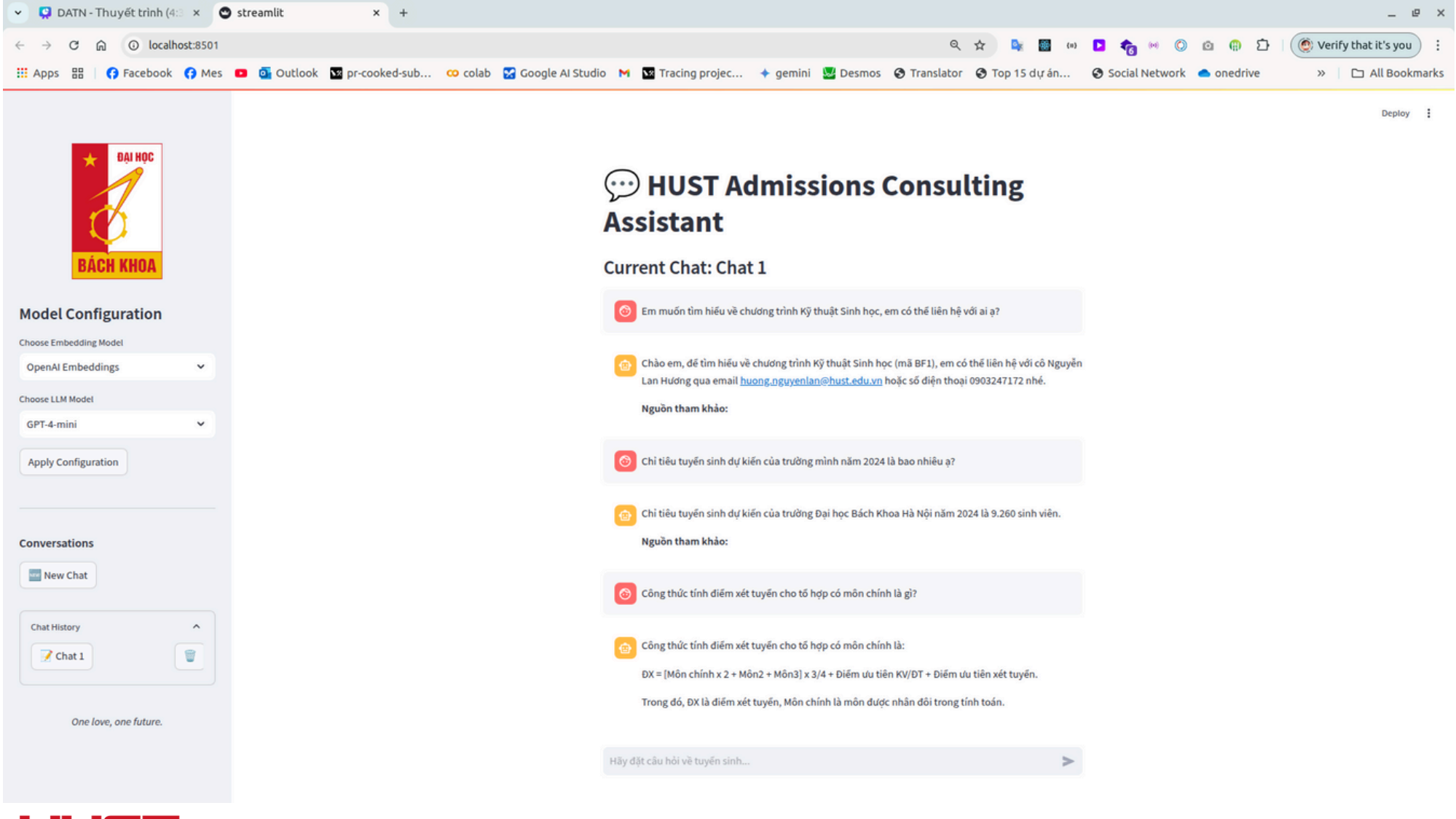
- Độ chính xác: Mức trung bình, hiệu suất ổn định nhưng vẫn có thể cải thiện thêm
- Miền có độ chính xác cao là ***Hướng nghiệp*** và ***Thông tin chung***, miền có độ trích xuất thấp: ***Điểm chuẩn tuyển sinh, Đề án tuyển sinh***

## Hệ thống sinh câu trả lời

- Mô hình sử dụng RAG có hiệu suất tốt hơn so với ChatGPT-4o đơn thuần, chứng tỏ việc kết hợp truy xuất thông tin giúp cải thiện độ chính xác của câu trả lời.
- Fine-tuning tiếp tục giúp nâng cao hiệu suất, thể hiện qua kết quả tốt nhất của mô hình ***RAG + Fine-tuned LLaMA-3-8B-Instruct.***

# PHẦN 5: DEMO CHƯƠNG TRÌNH

# Giao diện chatbot



# PHẦN 6: KẾT LUẬN

# Hướng phát triển trong tương lai

- Tích hợp thêm chức năng phân tích dự báo và khuyến nghị
- Thêm các tiêu chí đánh giá mô hình bằng việc sử dụng mô hình thứ 3 chuyên cho việc đánh giá câu trả lời của LLM
- Tăng độ chính xác trong việc trả lời câu hỏi
- Cập nhật những mô hình LLM thường xuyên

HUST

# THANK YOU !

hust.edu.vn fb.com/dhbkhn