

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN CUỐI KỲ  
CHATBOT RAG VỀ CHƯƠNG TRÌNH ĐÀO TẠO  
FITHCMUS - CHƯƠNG TRÌNH CHUẨN**

**Nhóm 5**

20120145	Đường Yến Ngọc
21120198	Nguyễn Thị Lan Anh
21120417	Nguyễn Thị Ngọc Châm
21120426	Huỳnh Phát Đạt

**NHẬP MÔN NGÔN NGỮ HỌC THỐNG KÊ VÀ ỨNG DỤNG  
--GIẢNG VIÊN--  
Lê Thanh Tùng**

## MỤC LỤC

<b>I. Thành viên nhóm .....</b>	3
<b>II. Giới thiệu đề tài .....</b>	4
1. Nội dung đề tài .....	4
2. Phương pháp tiếp cận.....	4
<b>III. Dataset.....</b>	5
1. Thu thập data.....	5
2. Tiền xử lý data .....	6
<b>IV. Xây dựng ứng dụng.....</b>	6
1. Nền tảng hỗ trợ .....	7
2. Xử lý và nạp dữ liệu .....	7
3. Truy xuất thông tin và sinh câu trả lời.....	8
4. Kết nối Front-end .....	9
<b>V. Đánh giá kết quả .....</b>	11
1. Đánh giá front-end.....	11
2. Đánh giá Back-end.....	16
a. Đánh giá Retrieval .....	16
b. Đánh giá Generation .....	19
c. Đánh giá tổng thể mô hình sử dụng để xây dựng chatbot .....	21
3. Các hướng cải tiến .....	22
<b>VI. Demo.....</b>	23
<b>VII. Tham khảo .....</b>	24

## I. Thành viên nhóm

MSSV	Họ và Tên	Phân công	Mức độ hoàn thành
20120145	Đường Yên Ngọc	- Thu thập và xử lý dữ liệu - Khám phá dữ liệu - Đánh giá mô hình - Viết báo cáo	100%
21120198	Nguyễn Thị Lan Anh	- Thu thập và xử lý dữ liệu - Code back-end - Đánh giá mô hình - Viết báo cáo	100%
21120417	Nguyễn Thị Ngọc Châm	- Thu thập và xử lý dữ liệu - Vector embedding - Code back-end - Viết báo cáo	100%
21120426	Huỳnh Phát Đạt	- Thu thập và xử lý dữ liệu - Code front-end - Kiểm thử - Viết báo cáo	100%

## II. Giới thiệu đề tài

### 1. Nội dung đề tài

- Tên đề tài: Xây dựng chatbot RAG hỏi đáp về Chương trình đào tạo FITHCMUS - Chương trình Chuẩn.
- Lý do chọn đề tài: Nhu cầu tìm kiếm thông tin nhanh và chính xác về chương trình đào tạo. Do tài liệu về chương trình đào tạo của các ngành/chuyên ngành được tách ra thành các file riêng lẻ và là file scan, mất nhiều thời gian tra cứu thông tin.
- Mục tiêu đề tài: Xây dựng một hệ thống chatbot RAG có thể truy xuất thông tin chương trình đào tạo và trả lời các câu hỏi bằng ngôn ngữ tự nhiên. Cung cấp công cụ hỗ trợ tra cứu nhanh, chính xác và thân thiện với người dùng.
  - + Input: Một câu hỏi bằng ngôn ngữ tự nhiên của người dùng.  
*Ví dụ:* “Tôi cần tích lũy bao nhiêu tín chỉ bắt buộc chuyên ngành Công nghệ tri thức?”
  - + Output: Câu trả lời có chứa thông tin được chọn lọc và tổng hợp từ các mục dữ liệu trong dataset, cung cấp các địa điểm phù hợp với mô tả, đánh giá, và các thông tin liên quan.  
*Ví dụ:* “Bạn cần tích lũy tối thiểu 16 tín chỉ cho các học phần bắt buộc chuyên ngành Công nghệ tri thức”
- Đối tượng và phạm vi nghiên cứu:
  - + Đối tượng: Chương trình đào tạo của Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên - HCMUS.
  - + Phạm vi: Thông tin ngành/chuyên ngành, các môn học, số tín chỉ, nội dung môn học, và các yêu cầu học phần,...

### 2. Phương pháp tiếp cận

- Retrieval-Augmented Generation - RAG:
  - + Định nghĩa: RAG là một kỹ thuật kết hợp giữa truy xuất thông tin từ cơ sở dữ liệu (Retrieval) và tạo câu trả lời tự nhiên (Generation) dựa trên mô hình ngôn ngữ lớn (Large Language Model - LLM).
  - + Ưu điểm:
    - Khả năng truy xuất thông tin từ dữ liệu lớn.
    - Tạo câu trả lời tự nhiên và linh hoạt, phù hợp với ngữ cảnh.

### III. Dataset

#### 1. Thu thập data

- Nhóm thu thập dữ liệu chương trình đào tạo (Chương trình Chuẩn khóa 2023) từ tài liệu chính thức (Trang web khoa FITHCMUS):

+ Đề cương môn học: Trang web khoa FITHCMUS > Đào tạo > Chương trình đào tạo > Chương trình Chuẩn

(<https://www.fit.hcmus.edu.vn/vn/Default.aspx?tabid=36>)

+ Chương trình đào tạo: Trang web khoa FITHCMUS > Hệ thống sinh viên > CQuy > Chương trình đào tạo > CQuy Khóa tuyển 2023

(<https://www.fit.hcmus.edu.vn/vn/Default.aspx?tabid=289>)

+ Câu hỏi và trả lời mẫu: Trang Q&A FIT

(<https://courses.fit.hcmus.edu.vn/q2a/>)

##### 7. NỘI DUNG CHƯƠNG TRÌNH ĐÀO TẠO

###### 7.1. KIẾN THỨC GIÁO DỤC ĐẠI CƯỜNG

Tích lũy tổng cộng 56 tín chỉ (Không kể Ngoại ngữ, Giáo dục thể chất và Giáo dục quốc phòng – an ninh):

###### 7.1.1. I.ý luận chính trị – Pháp luật

STT	MÃ HỌC PHẦN	TÊN HỌC PHẦN	SỐ TC	SỐ TIẾT	Lý thuyết	Thực hành	Bài tập	Loại học phần	Ghi chú
1	BAA00101	Triết học Mác - Lênin	3	45	0	0	BB		
2	BAA00102	Kinh tế chính trị Mác - Lênin	2	30	0	0	BB		
3	BAA00103	Chủ nghĩa xã hội khoa học	2	30	0	0	BB		
4	BAA00104	Lịch sử Đảng Cộng sản Việt Nam	2	30	0	0	BB		
5	BAA00003	Tư tưởng Hồ Chí Minh	2	30	0	0	BB		
6	BAA00004	Pháp luật đại cương	3	45	0	0	BB		
<b>TỔNG CỘNG</b>			<b>14</b>						

###### 7.1.2. Khoa học xã hội – Kinh tế – Kỹ năng

STT	MÃ HỌC PHẦN	TÊN HỌC PHẦN	SỐ TC	SỐ TIẾT	Lý thuyết	Thực hành	Bài tập	Loại học phần	Ghi chú
1	Chọn 01 học phần (02 tín chỉ) trong các học phần sau:								
	BAA00005	Kinh tế đại cương	2	30	0	0	TC		
	BAA00006	Tâm lý đại cương	2	30	0	0	TC		
	BAA00007	Phương pháp luận sัง tạo	2	30	0	0	TC		

###### Các qui định

Qui định	Category	Modified Date
QD 2380 sửa đổi bổ sung QĐ 1625 K2022	Tài Xuống	17/09/2022
Quy định về việc giá thuê học phí QĐ 2165/QĐ-KHTN, 26/10/2023	Tài Xuống	10/11/2023
Tiêu chuẩn đăng ký KLTN áp dụng cho Khóa 2011-2014	Tài Xuống	11/10/2022
Hướng dẫn cách thức nộp KLTN- Cử nhân tài năng K2015	CNTN	29/09/2021
Quy định chuẩn đầu ra ngoại ngữ CT Cử nhân tài năng Khóa 2023 về sau	CNTN	11/03/2023
DHCQ-Quy chế đào tạo 2021	QCHV	11/10/2022
QD 1175-DHCQ-Quy chế đào tạo 2021	QCHV	11/10/2022
Quy chế đào tạo HCTC 2016	QCHV	11/10/2022
QD 534 - sửa đổi chuẩn TA và trình độ TA đối với DHCQ, ngày 15/06/2020	QĐ TA	11/10/2022
QĐ 11/10/2022	QĐ TA	11/10/2022

*Dữ liệu về chương trình đào tạo được tổng hợp từ trang web chính thức FITHCMUS*

## 2. Tiền xử lý data

- Chuyển đổi dữ liệu thành dạng thuận văn bản:
  - + Format theo cấu trúc chung
  - + Xử lý lỗi chính tả
- Chia thành các file định dạng .txt theo từng chuyên ngành, quy chế đào tạo, mô tả môn học.

### Data 10 items

Name	Last modified	File size
BSMS.txt	Dec 4, 2024	5 KB
CNTT.txt	Dec 14, 2024	47 KB
DKTN.txt	Dec 14, 2024	9 KB
HTTT.txt	Dec 14, 2024	32 KB
KHMT.txt	Dec 14, 2024	67 KB
KTPM.txt	Dec 14, 2024	32 KB
MonHoc.txt	Dec 4, 2024	474 KB
NN.txt	Dec 14, 2024	13 KB
QData.xlsx	Dec 14, 2024	9 KB
TTNT.txt	Dec 14, 2024	33 KB

Chương trình: Đại học Chính quy  
 Ngành: Hệ thống thông tin  
 Khóa tuyển: 2023  
 (Ban hành kèm theo Quyết định số 1712/QĐ - KHTN ngày 07/9/2023  
 của Hiệu trưởng Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM)

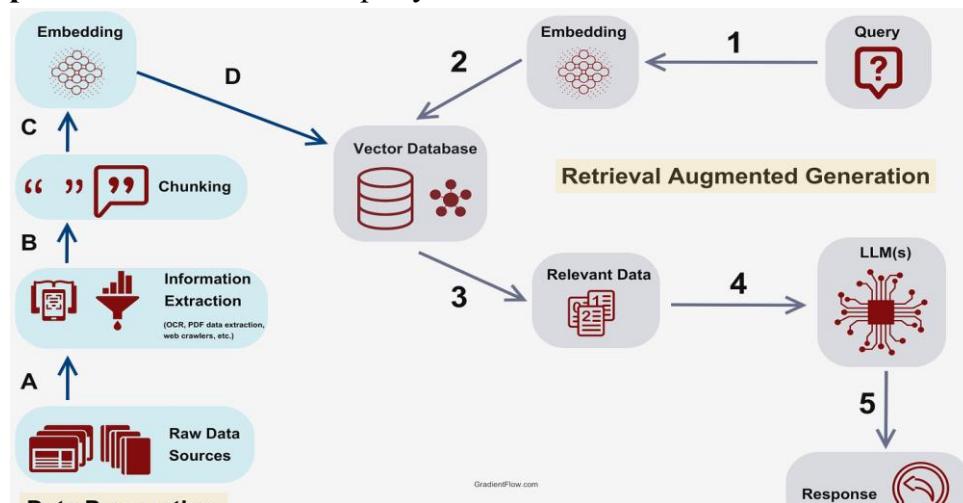
1. Thông tin chung về chương trình đào tạo:  
 1.1. Tên ngành:  
 - Tiếng Việt: Hệ thống thông tin  
 - Tiếng Anh: Information Systems  
 1.2. Mã ngành: 7480104  
 1.3. Trình độ đào tạo: Đại học  
 1.4. Tên chương trình: Cử nhân Hệ thống thông tin  
 1.5. Loại hình đào tạo: Chính quy  
 1.6. Thời gian đào tạo: 4 năm  
 1.7. Tên văn bằng sau khi tốt nghiệp:  
 - Tiếng Việt: Cử nhân Hệ thống thông tin  
 - Tiếng Anh: Bachelor of Science in Information Systems  
 1.8. Ngôn ngữ giảng dạy: tiếng Việt  
 1.9. Nơi đào tạo:  
 - Cơ sở 1: 227 Nguyễn Văn Cừ, P4, Q5, TP. HCM  
 - Cơ sở 2: Phường Linh Trung, Thành phố Thủ Đức, TP. HCM

Dữ liệu được chuyển sang dạng text và lưu vào các file .txt

## IV. Xây dựng ứng dụng

- Nhóm xây dựng ứng dụng hỏi đáp với các bước chính:

- + **Chunking** : Nhiệm vụ chính là chia nhỏ tài liệu (text) thành các đoạn nhỏ hơn (chunks).
- + **Embedding** : Nhúng các chunks thành các vector đại diện cho chunks đó.
- + **VectorDB** : Những vector được nhúng sẽ được lưu vào một vector database để phục vụ cho việc truy vấn dữ liệu sau này.
- + **Retrieval** : Khi hệ thống nhận được truy vấn (query) từ người dùng, hệ thống sẽ nhúng query thành vector và tìm kiếm các chunks có liên quan đến query của người dùng (relevant chunks).
- + **Response Generation**: Gửi query và relevant chunks lên LLM để nhận câu trả lời.



RAG pipeline

## 1. Nền tảng hỗ trợ

- Ngôn ngữ lập trình:
  - + Back-end: Python, thư viện chính Langchain..
  - + Front-end: JavaScript, thư viện Vite React.
- **Mô hình generate:** Vi-Qwen2-3B-RAG, gemini-1.5-flash
- **Mô hình embedding:** LaBSE, multilingual-e5-base
- **Môi trường:** Google Colab, VS Code

## 2. Xử lý và nạp dữ liệu

- **Documents:** Dữ liệu Các tài liệu hoặc dữ liệu ban đầu được đưa vào hệ thống.
- **Chunks:** Các tài liệu được chia nhỏ thành từng đoạn (chunks) để dễ quản lý và xử lý: Áp dụng Recursive chunking, thay đổi 1 chút để đảm bảo các chunk sẽ bắt đầu bằng đề mục (ví dụ: 1. Thông tin chung chương trình đào tạo,...)

Chunk 33:

7.1.3. Toán – Khoa học tự nhiên – Công nghệ – Môi trường

- Tên học phần: Vi tích phân 2B
- Mã học phần: MTH00004
- Số tín chỉ: 3
- Loại học phần: Bắt buộc

- Tên học phần: Toán rời rạc

- Mã học phần: MTH00041

- Số tín chỉ: 3

- Loại học phần: Bắt buộc

Metadata: {'source': '/content/Data/KTPM.txt'}

=====

In 5 chunk đầu tiên

Chunk 34:

7.1.3. Toán – Khoa học tự nhiên – Công nghệ – Môi trường

- Mã học phần: MTH00083
- Số tín chỉ: 1
- Loại học phần: Bắt buộc

- Tên học phần: Xác suất thống kê

- Mã học phần: MTH00008

- Số tín chỉ: 3

- Loại học phần: Bắt buộc

- Tên học phần: Thực hành Vi tích phân 2B

- Mã học phần: MTH00082

- Số tín chỉ: 1

- Loại học phần: Bắt buộc

- Tên học phần: Đại số tuyến tính

- Mã học phần: MTH00007

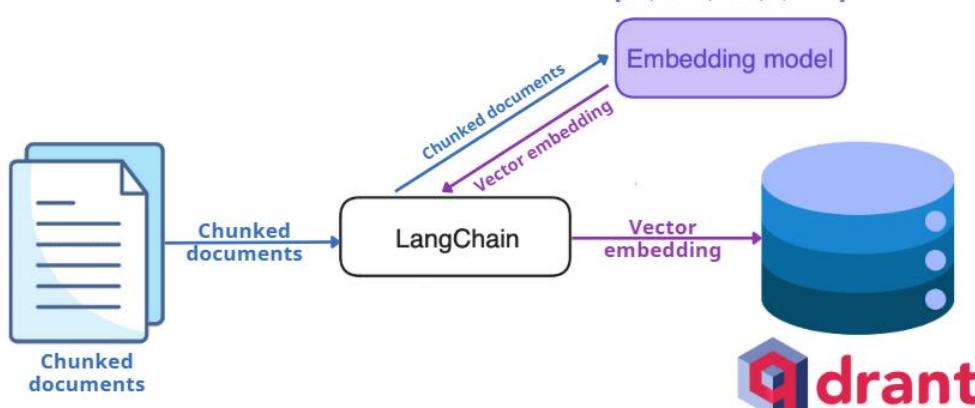
- Số tín chỉ: 3

- Loại học phần: Bắt buộc

### Kết quả sau khi chunking

- **Embeddings:** Sử dụng embedding models trên hugging face với số chiều là 768 (LaBSE, multilingual-e5-base) để chuyển các đoạn dữ liệu này thành thành các vector, sao cho: những dữ liệu tương đồng thì sẽ có vị trí gần nhau trong không gian vector. Ví dụ, hai câu có ý nghĩa tương tự sẽ được ánh xạ thành hai vector sao cho khoảng cách giữa chúng sẽ nhỏ hơn so với các câu không liên quan.
- **Vector Database:** Sau khi dữ liệu được xử lý và chuyển đổi thành vector embeddings, các vector này được lưu trữ cùng với metadata liên quan (như ID tài liệu, title,...)

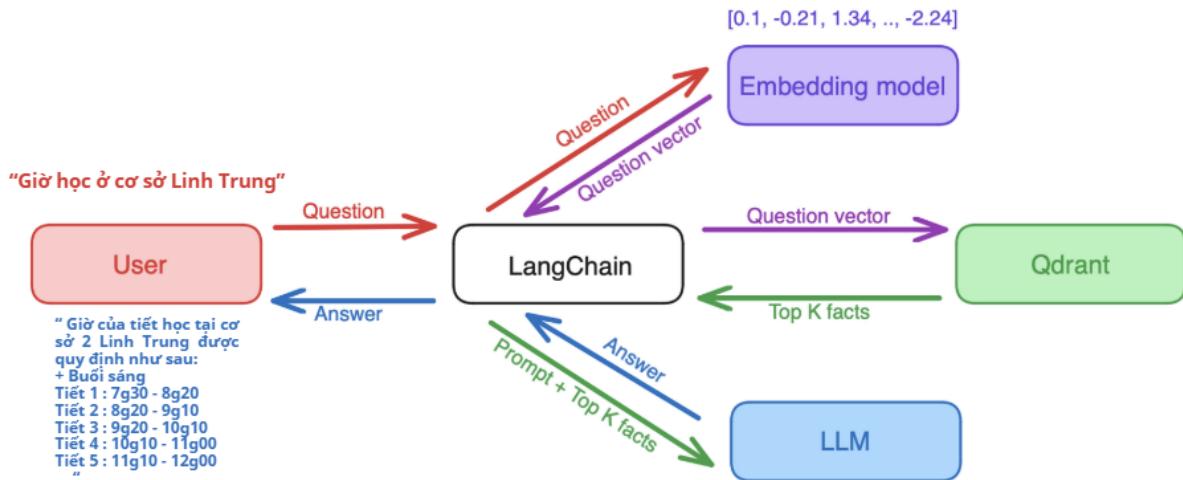
[0.1, -0.21, 1.34, ..., -2.24]



Nhúng các chunks vào vector database Qdrant

### 3. Truy xuất thông tin và sinh câu trả lời

- **Vector Search:** Khi có truy vấn từ người dùng, query vector được tạo và sử dụng để tìm kiếm trong Vector Database để tìm ra top-k vectors gần nhất.
- Kết hợp mô hình ngôn ngữ lớn để truy xuất và sinh câu trả lời tự nhiên trả về cho người dùng:
  - + Sau khi tìm được top-k vectors gần nhất, kết hợp với prompting để LLM có thể dựa vào ngữ cảnh để suy luận ra câu trả lời phù hợp cho người dùng.



*LLM lấy top-k vectors gần nhất qua các bước suy luận và đưa ra câu trả lời*

+ Promt template:

```

<|system|>
'''Thực hiện trả lời câu hỏi từ thông tin có trong ngữ cảnh được cho. Chú ý các yêu cầu sau:
- Câu trả lời phải chính xác, đầy đủ và có liên quan đến câu hỏi.
- Chỉ sử dụng các thông tin có trong ngữ cảnh được cung cấp.
- Nếu có nhiều ngữ cảnh liên quan với câu hỏi thì kết hợp các ngữ cảnh để tổng hợp thông tin.
- Nếu ngữ cảnh không chứa thông tin về câu trả lời thì từ chối trả lời và không suy luận gì th
"Vui lòng liên lạc Khoa Công Nghệ Thông Tin, trường Đại học Khoa Học Tự Nhiên - Đại học Quốc G
Địa chỉ: Phòng I.54, toà nhà I, 227 Nguyễn Văn Cừ, Q.5, TP.HCM
Điện thoại: (028) 62884499
Email: info@fit.hcmus.edu.vn"
</s>
<|user|>
Hãy trả lời câu hỏi sau dựa trên ngữ cảnh sau:

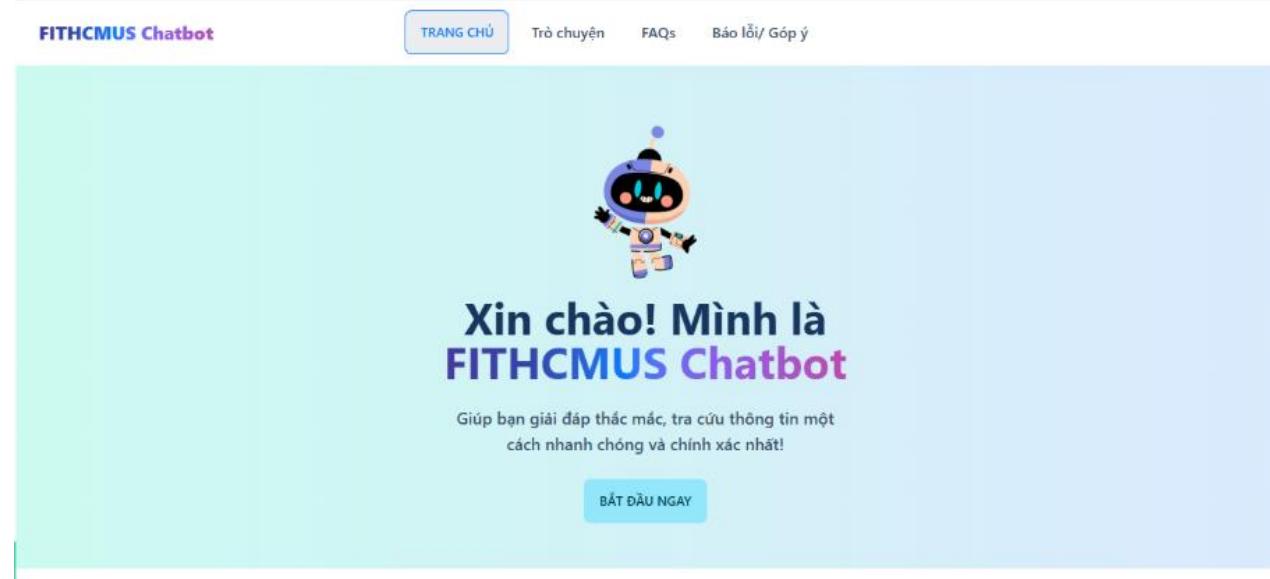
{context}
---

Câu hỏi: {question}
</s>
<|assistant|>
"""
  
```

*Thực hiện prompt để lấy thông tin từ LLM*

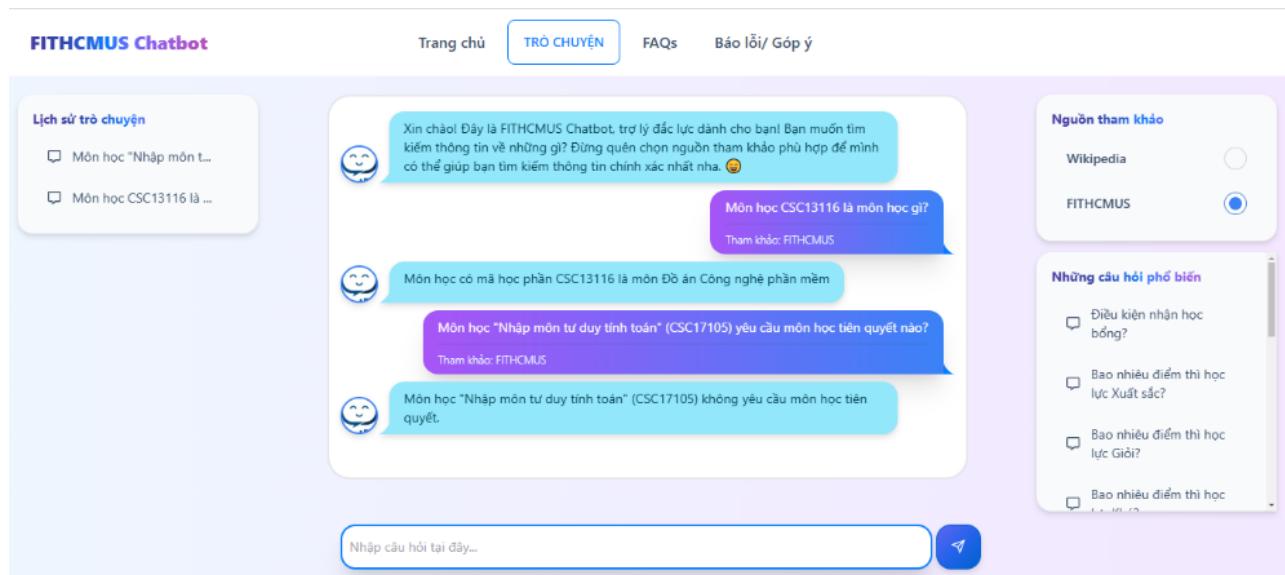
## 4. Kết nối Front-end

- Trang chủ của FITHCMUS Chatbot với local host: <http://localhost:5173/> giới thiệu ngắn gọn về mục đích hoạt động của Chatbot:



*Trang chủ của ứng dụng giới thiệu ngắn gọn về mục đích hoạt động của Chatbot*

- + Muốn thực hiện chatbot thì có thể chọn nút “**BẮT ĐẦU NGAY**” ở giữa màn hình hoặc nút *Trò chuyện* trên thanh Navbar.



*Giao diện hỏi đáp của chatbot*

- Trang hỏi đáp local host: <http://localhost:5173/chat> với chức năng chính là hỏi đáp với chatbot về chủ đề Chương trình đào tạo ngành CNTT (FITHCMUS) khóa 2023 Chương trình Chuẩn.
  - + Người dùng có thể chọn câu hỏi nhanh ở mục câu hỏi phổ biến nằm bên trái hoặc hỏi những câu hỏi mong muốn liên quan đến chủ đề bằng cách nhập vào thanh input ở dưới sau đó nhấn enter hoặc nút gửi bên cạnh ô input.

- Trang FAQ để giới thiệu về chatbot: <http://localhost:5173/faq> sử dụng accordion để giúp người dùng hiểu về cách thức hoạt động của chatbot, chủ đề về chatbot là gì, ... và những câu hỏi khác liên quan

**Những câu hỏi thường gặp (FAQs)**

Chatbot hoạt động như thế nào?

Chatbot hoạt động bằng cách từ câu hỏi của người dùng, sử dụng kỹ thuật tìm văn bản liên quan đến câu hỏi trong bộ dữ liệu đã được vector hóa (text similarity) và lưu trữ thông qua vector database. Giúp lấy ra những đoạn văn bản có liên quan sau đó dùng mô hình ngôn ngữ lớn (LLM) Vietcuna để sinh câu trả lời.

Cách sử dụng chatbot để tra cứu thông tin

Thông tin từ chatbot có đáng tin cậy không?

Vì là một mô hình xác xuất nên thông tin chatbot đưa ra có thể không chính xác ở một số trường hợp, bạn nên kiểm chứng thông tin hoặc liên hệ hỗ trợ nếu cần thiết nhé

Tôi có thể liên hệ hỗ trợ như thế nào?

*Giao diện trang FAQs*

- Cuối cùng là trang báo lỗi / góp ý: <http://localhost:5173/feedback>

**Báo lỗi hoặc góp ý**

Sự đóng góp ý kiến từ các bạn sẽ là sự hỗ trợ đắc lực giúp chúng tôi ngày càng hoàn thiện sản phẩm hơn.

Nêu cải thiện để trải nghiệm người dùng tốt hơn

[huyndphatdat@gmail.com](mailto:huyndphatdat@gmail.com)

**GỬI Ý KIẾN**

*Giao diện trang báo lỗi/góp ý*

- + Ở trang này, người dùng có thể góp ý hoặc báo lỗi bằng cách nhập nội dung và email của người dùng, sau đó chọn nút *Gửi ý kiến* để nhóm có thể tham khảo và sửa đổi dựa trên góp ý của mọi người để hoàn thiện chatbot tốt hơn.

## V. Đánh giá kết quả

### 1. Đánh giá front-end

- Kiểm thử các chức năng của ứng dụng bằng **Manual Testing**. Sử dụng phương pháp **Use Case Testing** để kiểm thử các chức năng của ứng dụng.

#### a. Header.

Xác định use case:

Flow of event	Step	Description
Basic flow	1	Người dùng ở trang chủ, trên thanh header ở phần navigation hiển thị các navigation item để người dùng có thể di chuyển đến các trang khác nhau.
Basic flow	2	Người dùng chọn các navigation item như “Trò chuyện”, “FAQS” hoặc “Báo lỗi / Góp ý”
Basic flow	3	Hệ thống di chuyển đến các trang tương ứng Chat page, FAQ page, Feedback page.
Basic flow	4	Người dùng muốn trở về trang chủ, chọn navigation item “Trang chủ”
Alternate flow	4a	Người dùng muốn trở về trang chủ, chọn logo ở bên trái phần header.
Basic flow & Alternate flow	5	Hệ thống di chuyển màn hình người dùng đến trang chủ.

Xác định scenario:

Scenario name	Starting flow	Alternate flow
Scenario 1 - Di chuyển thành công	Basic flow	
Scenario 2 - Di chuyển thành công	Basic flow	Alternate flow

**b. Trang chủ.**

Xác định use case:

Flow of event	Step	Description
Basic flow	1	Người dùng ở trang chủ
Basic flow	2	Người dùng chọn nút "BẮT ĐẦU NGAY" ở giữa màn hình để chuyển đến trang hỏi đáp với chatbot
Alternate flow	2a	Người dùng chọn nút "Trò chuyện" trên thanh navigation thuộc phần header.
Basic flow & Alternate flow	3	Hệ thống di chuyển màn hình người dùng đến trang trò chuyện với Chatbot.

Xác định scenario:

Scenario name	Starting flow	Alternate flow
Scenario 1 - Di chuyển thành công	Basic flow	
Scenario 2 - Di chuyển thành công	Basic flow	Alternate flow

**c. Chat page.****• Hỏi đáp chatbot.**

Xác định use case:

Flow of event	Step	Description
Basic flow	1	Người dùng ở trang trò chuyện
Basic flow	2	người dùng nhập câu hỏi sau đó chọn nút bên phải màu xanh hoặc "Enter" để thực hiện hỏi đáp.
Alternate flow	2a	Người dùng chọn câu hỏi có sẵn ở mục "Những câu hỏi phổ biến" ở bên phải sau đó chọn nút bên phải màu xanh hoặc "Enter" để thực hiện hỏi đáp.
Alternate flow	2b	Người dùng để khoảng trắng " " nhập vào hệ thống
Basic flow & Alternate flow	3a	Hệ thống sẽ tạo câu trả lời tương ứng câu hỏi mà người dùng đã hỏi.
Alternate flow	3b	Hệ thống không cho phép gửi khoảng trắng

Xác định scenario:

Scenario name	Starting flow	Alternate flow
Scenario 1 - Trả lời thành công	Basic flow	Basic flow 3a
Scenario 2 - Trả lời thành công	Basic flow	Alternate flow 3a
Scenario 3 - Không gửi được	Basic flow	Alternate flow 3b

- **Xem lịch sử trò chuyện.**

Xác định use case.

Flow of event	Step	Description
Basic flow	1	Người dùng ở trang trò chuyện
Basic flow	2	Người dùng thao tác hỏi đáp với cách nhập bàn phím hoặc câu hỏi có sẵn.
Basic flow	3	Hệ thống sẽ tự động tạo ra các lịch sử cho từng câu hỏi.

Xác định scenario:

Scenario name	Starting flow	Alternate flow
Scenario 1 - Xem thành công	Basic flow	

#### d. FAQS page.

Xác định use case:

Flow of event	Step	Description
Basic flow	1	Người dùng ở trang FAQS
Basic flow	2	Người dùng chọn các nút accordion hiển thị câu hỏi về Chatbot.
Basic flow	3	Hệ thống hiện ra những câu trả lời cho các câu hỏi accordion tương ứng.

Xác định scenario:

Scenario name	Starting flow	Alternate flow
Scenario 1 - Xem câu trả lời thành công	Basic flow	

#### e. Feedback page

Xác định use case.

Flow of event	Step	Description
Basic flow	1	Người dùng ở trang Góp ý / Báo lỗi
Basic flow	2	Người dùng nhập nội dung phản hồi
Alternate flow	2a	Người dùng để trống ô nội dung phản hồi
Basic flow	3	Người dùng nhập thông tin email cá nhân của mình
Alternate flow	3a	Người dùng để trống ô thông tin email cá nhân
Basic flow & Alternate flow	3	Người dùng chọn nút “GỬI Ý KIẾN” màu xanh ở bên dưới.
Basic flow	4	Gửi phản hồi thành công. Nhận được thông tin feedback
Alternate flow	4a	Gửi phản hồi thất bại. Không nhận được thông tin feedback

Xác định scenario:

Scenario name	Starting flow	Alternate flow
Scenario 1 - Gửi phản hồi thành công	Basic flow	
Scenario 2 - Gửi phản hồi thất bại	Basic flow	Alternate flow 4a

- Tốc độ truyền tải (Load Testing):** Thông tin của ứng dụng nhanh, nhanh chóng phản hồi tới người dùng sau khi hệ thống nhận được câu hỏi.

## - Kết quả kiểm thử:

Function/Feature ID	Case ID	Test case name	Test step	Expected Result (ER)	Actual Result	Status
<b>Function 01: Header</b>						
UC01	TC01	Điều hướng qua các trang khác nhau trên thanh navigation	Ở trang chủ ứng dụng, chọn nút "Trò chuyện" để đến trang Chat. Chọn nút "FAQS" để đến trang FAQ. Chọn nút "Báo lỗi/Góp ý" để đến trang Feedback.	Điều hướng người dùng đến trang Chat, FAQ, Feedback.	Giống expected result	Pass
UC01	TC02	Trở về trang chủ	Ở trang Chat, trang FAQ hoặc trang Feedback. Có thể trở về trang chủ thay vì chọn nút "Trang chủ" ở trên navigation, người dùng có thể chọn logo ở phía bên trái Header.	Điều hướng người dùng về trang chủ	Giống expected result	Pass
<b>Function 02: Trang chủ</b>						
UC02	TC01	Từ trang chủ điều hướng đến hỏi đáp với chatbot	Ở trang chủ, chọn nút "BẮT ĐẦU NGAY" ở giữa màn hình để chuyển đến trang hỏi đáp với chatbot.	Điều hướng người dùng đến trang Chat.	Giống expected result	Pass
UC02	TC02	Từ trang chủ điều hướng đến hỏi đáp với chatbot	Ở trang chủ, chọn nút "Trò chuyện" trên thanh navigation thuộc phần header.	Điều hướng người dùng đến trang Chat.	Giống expected result	Pass
<b>Function 03: Chat page</b>						
UC03	TC01	Hỏi đáp với Chatbot bằng cách nhập bàn phím	Ở trang Chat, người dùng nhập câu hỏi sau đó chọn nút bên phải màu xanh hoặc "Enter" để thực hiện hỏi đáp.	Chatbot trả lời câu hỏi	Giống expected result	Pass
UC03	TC02	Hỏi đáp với Chatbot bằng câu hỏi có sẵn	Ở trang Chat, người dùng chọn câu hỏi có sẵn ở mục "Những câu hỏi phổ biến" ở bên phải sau đó chọn nút bên phải màu xanh hoặc "Enter" để thực hiện hỏi đáp.	Chatbot trả lời câu hỏi	Giống expected result	Pass
UC03	TC03	Kiểm tra lịch sử trò chuyện	Ở trang Chat, thao tác hỏi đáp với cách nhập bàn phím hoặc câu hỏi có sẵn.	Xuất hiện lịch sử các câu hỏi	Giống expected result	Pass
UC03	TC04	Hỏi đáp thất bại	Ở trang Chat, người dùng để khoảng trắng "" nhập vào hệ thống	Lỗi không cho phép gửi, disable nút gửi	Giống expected result	Pass
<b>Function 04: FAQ page</b>						
UC04	TC01	Xem những câu hỏi thường gặp về Chatbot	Ở trang FAQ, chọn vào các thanh accordion câu hỏi để xem chi tiết câu trả lời cho các câu hỏi về Chatbot	Xuất hiện các câu trả lời	Giống expected result	Pass
<b>Function 05: Feedback page</b>						
UC05	TC01	Gửi feedback	Người dùng nhập nội dung feedback và tên email. Sau đó chọn nút "GỬI Ý KIẾN" màu xanh.	Thông báo: "Gửi phản hồi thành công". Nhận được feedback trên gmail	Giống expected result	Pass
UC05	TC02	Gửi feedback thất bại	Người dùng không nhập nội dung feedback. Sau đó chọn nút "GỬI Ý KIẾN" màu xanh.	Thông báo: "Người dùng phải điền đầy đủ thông tin". Không nhận được feedback	Giống expected result	Pass
UC05	TC03	Gửi feedback thất bại	Người dùng không nhập thông tin email. Sau đó chọn nút "GỬI Ý KIẾN" màu xanh.	Thông báo: "Người dùng phải điền đầy đủ thông tin". Không nhận được feedback	Giống expected result	Pass

## 2. Đánh giá Back-end

- Tập dữ liệu đánh giá:

- + Tập dữ liệu đánh giá gồm các cột: ‘question’ (Câu hỏi về chương trình đào tạo), ‘expected\_output’ (Câu trả lời đúng cho câu hỏi tương ứng) và ‘ground\_truth’ (Nguồn/các nguồn tài liệu liên quan đến câu hỏi)
- + Tập dữ liệu bao gồm 144 mẫu dữ liệu liên quan đến chương trình đào tạo của các ngành/chuyên ngành, quy chế đào tạo, thông tin môn học,...

question	expected_output	ground_truth
Tổng số tín chỉ bắt buộc của ngành Hệ thống thông tin là bao nhiêu?	Tổng số tín chỉ bắt buộc của ngành Hệ thi HTTT.txt	
Tên văn bằng ngành Khoa học máy tính sau khi tốt nghiệp	Tên văn bằng ngành Khoa học máy tính s KHMT.txt	
Sinh viên phải làm gì để đăng ký học phân môn môi học kỹ	Đầu môi học kỹ, sinh viên phải theo dõi tì QuyChe.txt	
Sinh viên phải học môn nào bắt buộc trong chuyên ngành	Sinh viên chuyên ngành Khoa học máy tín KHMT.txt	
Sinh viên phải hoàn thành bao nhiêu tín chí thuộc Kiến thức	Sinh viên phải hoàn thành 82 tín chí thuộc CNTT.txt	
Sinh viên phải đăng ký tối thiểu bao nhiêu tín chí trong họ	Dối với chương trình Tiên tiến, Liên kết, C QuyChe.txt	
Sinh viên có thể kéo dài thời gian học tập tối đa bao lâu?	Sinh viên được phép kéo dài thêm không QuyChe.txt	
Sinh viên chuyên ngành Khoa học máy tính cần tích lũy bao nhiêu tín chí	Sinh viên chuyên ngành Khoa học máy tín KHMT.txt	
Sinh viên cần tích lũy ít nhất bao nhiêu tín chí từ học phân	Sinh viên cần tích lũy ít nhất 16 tín chí từ KHMT.txt	
Sinh viên cần tích lũy bao nhiêu tín chí cho phần tự chọn c	Sinh viên cần tích lũy tối thiểu 8 tín chí ch KTPM.txt	
Sinh viên cần làm gì để được hoãn thi trong trường hợp b	Trong trường hợp đột xuất và có lý do ch QuyChe.txt	
Sinh viên ngành Kỹ thuật phần mềm cần hoàn thành bao nhiêu tín chí	Sinh viên ngành Kỹ thuật phần mềm cần t KTPM.txt	
Sinh viên cần đạt điều kiện gì để đăng ký thực tập dự án	Sinh viên cần đạt 4 học phần Anh văn tro DKTN.txt	
Sinh viên cần đáp ứng những điều kiện nào để được công nhận	Sinh viên được xét và công nhận tốt nghi QuyChe.txt	
Những học phần nào được xem là bắt buộc trong chương	Những học phần bắt buộc là học phần chưa đượ QuyChe.txt	
Học phần bắt buộc là		
Ngành Khoa học máy tính có tổng cộng bao nhiêu tín chí	Ngành Khoa học máy tính có tổng cộng 1 KHMT.txt	
Ngành Công nghệ thông tin được giảng dạy bằng ngôn ngữ	Ngành Công nghệ thông tin được giảng d CNTT.txt	
Nếu không đạt học phần bắt buộc, sinh viên phải thực hiện	Sinh viên không đạt một học phần bắt bu QuyChe.txt	
Mục tiêu học phần của khóa học Thị giác robot bao gồm	Mục tiêu học phần của khóa học Thị giác MonHoc.txt	
Mục tiêu học phần CSC17104 là gì?	Mục tiêu học phần CSC17104 bao gồm: S MonHoc.txt	
Mục tiêu của khóa học 'Thực hành Hóa đại cương 2' là gì?	Sau khi hoàn thành thành công học phần MonHoc.txt	
Một tín chí trong chương trình đào tạo đại học tương ứng	Một tín chí tương đương với 15 tiết học I QuyChe.txt	
Một tiết học được tính bằng bao nhiêu phút	Một tiết học được tính bằng 50 phút gián QuyChe.txt	
Một năm học đại học thường có bao nhiêu học kỳ và thời	Một năm học của Trường được tổ chức r QuyChe.txt	
Môn học CSC16112 yêu cầu bao nhiêu tín chỉ và có những	Môn học CSC16112 - Chuyên đề Xử lý ản MonHoc.txt	
Môn học CSC14008 Phương pháp nghiên cứu khoa học cu	Môn học CSC14008 Phương pháp nghiên MonHoc.txt	
Môn học CSC13116 là môn học gì?	Môn học CSC13116 là Đề án Công nghệ c MonHoc.txt	
Môn học CSC11103 - Thiết kế mang vêu cầu bao nhiêu tín	Môn học CSC11103 - Thiết kế mang vêu c MonHoc.txt	

*Tập dữ liệu hỏi đáp để đánh giá mô hình*

### a. Đánh giá Retrieval

- Nhóm đã thử các mô hình embedding, kết hợp các phương pháp chunking khác nhau để so sánh và lựa chọn mô hình tối ưu. Cụ thể như sau:

+ Mô hình embedding: LaBSE, multilingual-e5-base

+ Phương pháp chunking:

- Recursive chunking truyền thống: Dữ liệu lớn được chia thành các phần nhỏ với kích thước cố định (chunk\_size), các chunks sẽ có một phần dữ liệu chồng lấn nhau để giữ được ngữ cảnh (overlap)
- Bổ sung để mục: Tương tự Recursive chunking nhưng các chunks sẽ được bổ sung thêm để mục nếu chunk đó bị cắt ngay đoạn không có đề mục, điều này giúp giữ ngữ cảnh tốt hơn

+ Nhóm dùng các độ đo để đánh giá như: Precision@5, Recall@5, Mean Reciprocal Rank (MRR), MAP@5, nDCG@5

- Các collection được lưu theo tên model và chunksize, lần lượt là:
  - ITUS\_e5\_R\_600: Chunk\_size 600 – overlap 200
  - ITUS\_e5\_R\_800: Chunk\_size 800 – overlap 200
  - ITUS\_e5\_R\_1000: Chunk\_size 1000 – overlap 200
  - ITUS\_e5\_600: Chunk\_size 600 – overlap 200
  - ITUS\_e5\_800: Chunk\_size 800 – overlap 200
  - ITUS\_e5\_1000: Chunk\_size 1000 – overlap 200
  - ITUS\_LaBSE\_R\_600: Chunk\_size 600 – overlap 200
  - ITUS\_LaBSE\_R\_800: Chunk\_size 800 – overlap 200
  - ITUS\_LaBSE\_R\_1000: Chunk\_size 1000 – overlap 200
  - ITUS\_LaBSE\_600: Chunk\_size 600 – overlap 200
  - ITUS\_LaBSE\_800: Chunk\_size 800 – overlap 200
  - ITUS\_LaBSE\_1000: Chunk\_size 1000 – overlap 200

- Kết quả đánh giá được ghi trong bảng sau:

Collection	Phương pháp chunking	Precision@5	Recall @5	MRR @5	MAP @5	nDCG @5
<b>multilingual-e5-base</b>						
ITUS_e5_R_600	Recursive	0.7750	<b>0.9838</b>	0.9450	0.7911	0.8361
ITUS_e5_R_800	Recursive	0.7708	0.9792	0.9277	0.8183	0.8562
ITUS_e5_R_1000	Recursive	0.7264	0.9676	0.8983	0.7170	0.7726
ITUS_e5_600	Bổ sung đề mục	<b>0.8417</b>	0.9630	<b>0.9653</b>	<b>0.8524</b>	<b>0.8730</b>
ITUS_e5_800	Bổ sung đề mục	0.8236	0.9595	0.9528	0.8345	0.8548
ITUS_e5_1000	Bổ sung đề mục	0.7806	0.9676	0.8980	0.7691	0.8137
<b>LaBSE</b>						
ITUS_LaBSE_R_600	Recursive	0.5750	<b>0.9329</b>	<b>0.8220</b>	0.6521	0.7168
ITUS_LaBSE_R_800	Recursive	0.5694	0.9062	0.7755	0.6538	0.7103
ITUS_LaBSE_R_1000	Recursive	0.5611	0.8981	0.7684	0.6322	0.6904
ITUS_LaBSE_600	Bổ sung đề mục	<b>0.6333</b>	0.8935	0.8124	0.6667	0.7175
ITUS_LaBSE_800	Bổ sung đề mục	0.6264	0.9132	0.8017	<b>0.6785</b>	<b>0.7300</b>
ITUS_LaBSE_1000	Bổ sung đề mục	0.5847	0.9074	0.7799	0.6470	0.7019

- Nhận xét chung:

- Nhìn chung, với cùng một mô hình embedding và phương pháp chunking giống nhau, các độ đo có xu hướng giảm khi chunk size tăng, điều này cho thấy chunk size càng lớn thì khả năng bị mất ngữ cảnh càng lớn.
- Mô hình multilingual-e5-base luôn cho ra kết quả cao hơn mô hình LaBSE cho thấy mô hình multilingual-e5-base có thể tốt hơn về mặt xác định ngữ nghĩa.

- Cụ thể:

+ Precision@5:

- Phương pháp **Bổ sung đè mục** luôn có Precision@5 cao hơn so với phương pháp **Recursive**, đặc biệt là khi sử dụng **chunksize 600** và **800**.
- Kết quả **Precision@5** cao nhất đạt được với **ITUS\_e5\_600 - Bổ sung đè mục** (0.8417), cho thấy phương pháp này có khả năng trả về các tài liệu chính xác hơn trong top 5.

+ Recall@5:

- **Recall@5** của phương pháp **Recursive** nhìn chung cao, đạt gần 1 (0.8981 đến 0.9838), cho thấy rằng hầu hết các tài liệu đúng đều được tìm thấy trong top 5 của kết quả trả về.
- **Bổ sung đè mục** có **Recall@5** thấp hơn một chút (0.8935 đến 0.9676), nhưng vẫn duy trì mức độ rất cao, chỉ có sự thay đổi nhỏ theo kích thước chunk.
- Điều này cho thấy phương pháp **Recursive** có khả năng phủ rộng hơn trong việc tìm các tài liệu đúng.

+ MRR@5 (Mean Reciprocal Rank):

- **MRR@5** là một chỉ số đo lường vị trí của tài liệu đúng đầu tiên. Trong hầu hết các thử nghiệm, **MRR@5** cho thấy phương pháp **Bổ sung đè mục** có giá trị cao hơn **Recursive** trên cùng chunk size.
- Kết quả **MRR** cao nhất thuộc về **ITUS\_e5\_600 - Bổ sung đè mục**, với MRR@5 là 0.9653.

+ MAP@5 (Mean Average Precision):

- **MAP@5** đo lường độ chính xác trung bình của các tài liệu đúng trong top 5. Phương pháp **Bổ sung đè mục** đạt được kết quả MAP cao hơn so với **Recursive**.
- Kết quả **MAP@5** cao nhất thuộc về **ITUS\_e5\_600 - Bổ sung đè mục** (0.8524), cho thấy phương pháp này có độ chính xác tổng thể tốt hơn trong việc trả về các tài liệu đúng.

+ nDCG@5:

- **nDCG@5** (Normalized Discounted Cumulative Gain) đo lường độ phù hợp của danh sách gợi ý, có tính đến vị trí của tài liệu đúng. Phương pháp **Bổ sung đè mục** lại một lần nữa thể hiện kết quả vượt trội.
- **ITUS\_e5\_600 + Bổ sung đè mục** có nDCG@5 cao nhất là 0.8730, cho thấy danh sách gợi ý của phương pháp này không chỉ trả về tài liệu đúng mà còn sắp xếp chúng ở các vị trí cao trong top 5.

- **Tổng kết:**

• **Phương pháp "Recursive":**

- Có **Recall@5** rất cao, cho thấy phương pháp này có khả năng tìm thấy hầu hết các tài liệu đúng.
- Tuy nhiên, **Precision@5, MAP@5**, và **nDCG@5** thấp hơn so với **Bổ sung đề mục**, đặc biệt khi **chunksize** tăng lên, cho thấy hiệu quả giảm sút khi xử lý nhiều thông tin hơn.
- **MRR@5** của phương pháp này cao, đặc biệt với **chunksize 600** và **800**, cho thấy tài liệu đúng đầu tiên xuất hiện khá sớm trong kết quả.

• **Phương pháp "Bổ sung đề mục":**

- Cung cấp kết quả xuất sắc nhất trong các chỉ số chính như **Precision@5, MAP@5**, và **nDCG@5**. Phương pháp này thường trả về các tài liệu đúng với độ chính xác cao và sắp xếp chúng tốt trong top 5, đặc biệt là với **chunksize 600** và **800**.
- **Recall@5** của phương pháp này thấp hơn một chút so với **Recursive**, nhưng vẫn ở mức rất cao, cho thấy khả năng tìm ra các tài liệu đúng là rất tốt.

→ Dựa trên kết quả đánh giá, **ITUS\_e5\_600 - Bổ sung đề mục** có kết quả tổng thể cao nhất, nên nhóm sẽ chọn collection này để sử dụng cho retrieval:

- + Mô hình embedding sử dụng: multilingual-e5-base
- + Chunk size – overlap: 600 – 200
- + Phương pháp chunking: Bổ sung đề mục

**b. Đánh giá Generation**

- Nhóm đã cài đặt RAG\_pipeline trên 2 mô hình Vi-Qwen2-3B-RAG và gemini-1.5-flash, sau đó sử dụng các độ đo đánh giá để so sánh hiệu quả của 2 mô hình này.
- Các độ đo đánh giá đã sử dụng:

- **BLEU (Bilingual Evaluation Understudy):** Đánh giá chất lượng văn bản sinh tự động so với văn bản tham chiếu (reference). Đo lường mức độ tương đồng giữa câu sinh ra và câu tham chiếu dựa trên n-grams.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Đo lường mức độ trùng lặp giữa n-grams của câu sinh ra và câu tham chiếu, đặc biệt phổ biến trong bài toán tóm tắt văn bản.
  - **ROUGE-1:** Đo lường mức độ trùng lặp giữa các đơn từ (unigrams).
  - **ROUGE-2:** Đo lường mức độ trùng lặp giữa các cặp từ liên tiếp (bigrams).
  - **ROUGE-L:** Đo lường chuỗi con chung dài nhất (Longest Common Subsequence - LCS).
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** Đánh giá chất lượng dựa trên sự kết hợp của unigrams, bao gồm việc kiểm tra trùng khớp (match) đồng nghĩa, gốc từ (stemming), và thứ tự từ.

- **SBERT Cosine Similarity (Sentence-BERT):** Đo độ tương đồng ngữ nghĩa giữa câu sinh ra và câu tham chiếu thông qua **cosine similarity** giữa vector biểu diễn hai câu.
  - Model Sentence-BERT sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 được dùng để mã hóa (encode) các câu.

- Kết quả đánh giá được ghi trong bảng sau:

Mô hình	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	SBERT Similarity
gemini-1.5-flash	0.6022	0.8253	0.7593	0.7705	0.7655	0.8808
Vi-Qwen2-3B-RAG	0.0614	0.2397	0.1992	0.2156	0.3339	0.8032

- Nhận xét và so sánh 2 mô hình:

- **BLEU**
  - **gemini-1.5-flash (0.6022):** Đây là một kết quả khá tốt, cho thấy mô hình có khả năng tạo ra câu trả lời gần với câu chuẩn, đặc biệt khi tính đến các n-gram phổ biến.
  - **Vi-Qwen2-3B-RAG (0.0614):** Chỉ số này rất thấp, cho thấy mô hình không tạo ra câu trả lời gần với câu chuẩn từ góc độ n-gram. Điều này có thể là do mô hình thiếu sự tương đồng chặt chẽ về từ vựng hoặc cấu trúc.
- **ROUGE (ROUGE-1, ROUGE-2, ROUGE-L)**
  - **gemini-1.5-flash** có điểm số cao ở cả ba chỉ số **ROUGE-1 (0.8253)**, **ROUGE-2 (0.7593)**, và **ROUGE-L (0.7705)**. Điều này cho thấy mô hình này có khả năng trùng khớp tốt với câu trả lời chuẩn về cả từ vựng và cấu trúc câu.
  - **Vi-Qwen2-3B-RAG** có điểm số rất thấp ở tất cả các chỉ số ROUGE
- **METEOR**
  - **gemini-1.5-flash (0.7655)** có điểm số cao, phản ánh sự tương đồng ngữ nghĩa và cấu trúc ngữ pháp mạnh mẽ, với khả năng kết hợp các yếu tố như từ đồng nghĩa và cấu trúc câu chính xác.
  - **Vi-Qwen2-3B-RAG (0.3339)** có điểm thấp, cho thấy mô hình này không đồng bộ tốt về mặt ngữ nghĩa và cấu trúc với câu trả lời chuẩn. Điều này có thể phản ánh việc mô hình thiếu khả năng hiểu ngữ nghĩa sâu sắc.
- **SBERT Similarity**
  - **gemini-1.5-flash (0.8808):** Điểm số này rất cao, cho thấy mô hình có khả năng hiểu và tái tạo ngữ nghĩa chính xác trong câu trả lời. Đây là một chỉ số mạnh mẽ trong việc đánh giá khả năng hiểu sâu về nội dung.
  - **Vi-Qwen2-3B-RAG (0.8032):** Mặc dù thấp hơn, nhưng đây vẫn là một chỉ số khá tốt. Mô hình có khả năng nhận diện được sự tương đồng ngữ nghĩa, mặc dù không bằng **gemini-1.5-flash**.

### - Tổng kết và Nhận xét

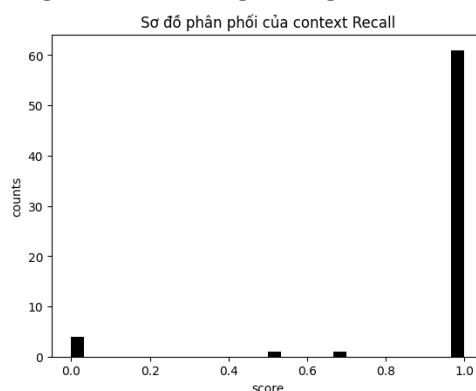
- **gemini-1.5-flash** thể hiện sự vượt trội rõ rệt ở hầu hết các chỉ số, từ **BLEU**, **ROUGE**, **METEOR** cho đến **SBERT Similarity**. Điều này cho thấy mô hình này tạo ra câu trả lời có sự trùng khớp cao về từ vựng, cấu trúc câu và ngữ nghĩa với câu trả lời chuẩn.
- **Vi-Qwen2-3B-RAG** có điểm số thấp, đặc biệt là ở các chỉ số ROUGE và BLEU. Mặc dù vẫn có sự tương đồng ngữ nghĩa (SBERT Similarity), mô hình này không thể tái tạo câu trả lời chính xác hoặc giữ vững cấu trúc tốt như **gemini-1.5-flash**. Điều này có thể chỉ ra rằng mô hình **Vi-Qwen2-3B-RAG** cần cải thiện khả năng sinh văn bản, đặc biệt là trong việc duy trì mối liên kết và sự chính xác với câu trả lời chuẩn.

→ Dựa trên kết quả đánh giá generate, nhóm chọn mô hình **gemini-1.5-flash** để dùng cho việc sinh văn bản.

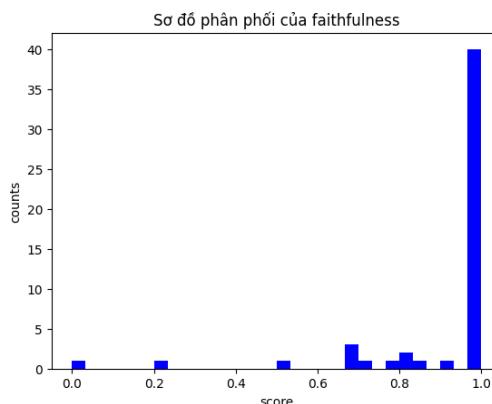
### c. Đánh giá tổng thể mô hình sử dụng để xây dựng chatbot

- Mô hình cuối cùng nhóm lựa chọn để xây dựng chatbot là ‘gemini-flask-1.5’ kết hợp với collection ITUS\_e5\_600 được lưu trong Qdrant. Kết quả đánh giá tổng thể như sau:

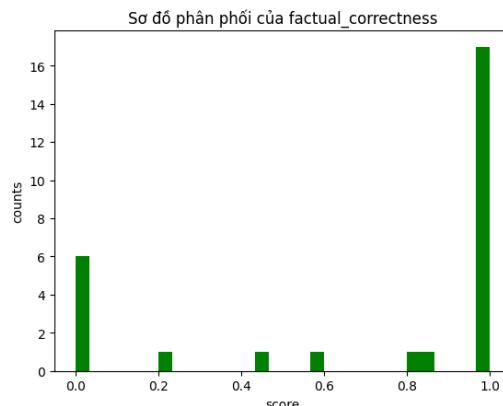
- **Context Recall:** Với chỉ số **0.9279**, mô hình sử dụng ngữ cảnh truy xuất rất hiệu quả, gần như 93% câu trả lời sinh ra đều bám sát vào tài liệu truy xuất. Điều này cho thấy mô hình đã thành công trong việc tận dụng thông tin từ tài liệu nguồn.



- **Faithfulness:** Mô hình khá trung thực với các tài liệu truy xuất, với **0.9144**, cho thấy rằng mô hình hầu như không bị hiện tượng "hallucination" (bịa đặt) thông tin. Mô hình tuân thủ đúng ngữ cảnh và không đưa vào các chi tiết không chính xác.



- **Factual Correctness:** Đây là chỉ số khá quan trọng, phản ánh độ chính xác thực tế của câu trả lời. Mặc dù **0.7107** cho thấy rằng hơn 71% câu trả lời mô hình đưa ra là chính xác, vẫn còn khoảng **28%** câu trả lời chưa đúng.



### 3. Các hướng cải tiến

- Để ứng dụng hỏi đáp hoàn thiện hơn, nhằm đáp ứng nhu cầu người dùng, nhóm đề xuất một số hướng cải tiến cho ứng dụng như:

- Tăng cường chất lượng dữ liệu đầu vào:
  - Mở rộng tập dữ liệu đào tạo
  - Loại bỏ nhiễu
- Tối ưu hóa mô hình RAG:
  - Cải thiện cách truy vấn như: reranking, cải thiện cấu trúc chia chunks, bổ sung metadata,...
  - Cải thiện prompt nếu có thể.
- Kiểm tra lại bộ câu hỏi kiểm tra, bổ sung các độ đo đánh giá khác để tăng tính tin cậy.
- Tối ưu hóa thời gian phản hồi và tăng khả năng chịu tải

## VI. Demo

- Nhóm mở server bằng file back-end trên Google Colab, đồng thời chạy front-end trên VS Code. Chi tiết cách chạy front-end:

- + Mở folder ‘front-end’ bằng VS Code
- + Chạy lệnh “*npm install*” để cài đặt toàn bộ package vào máy.

**PS D:\chatbot> npm install**

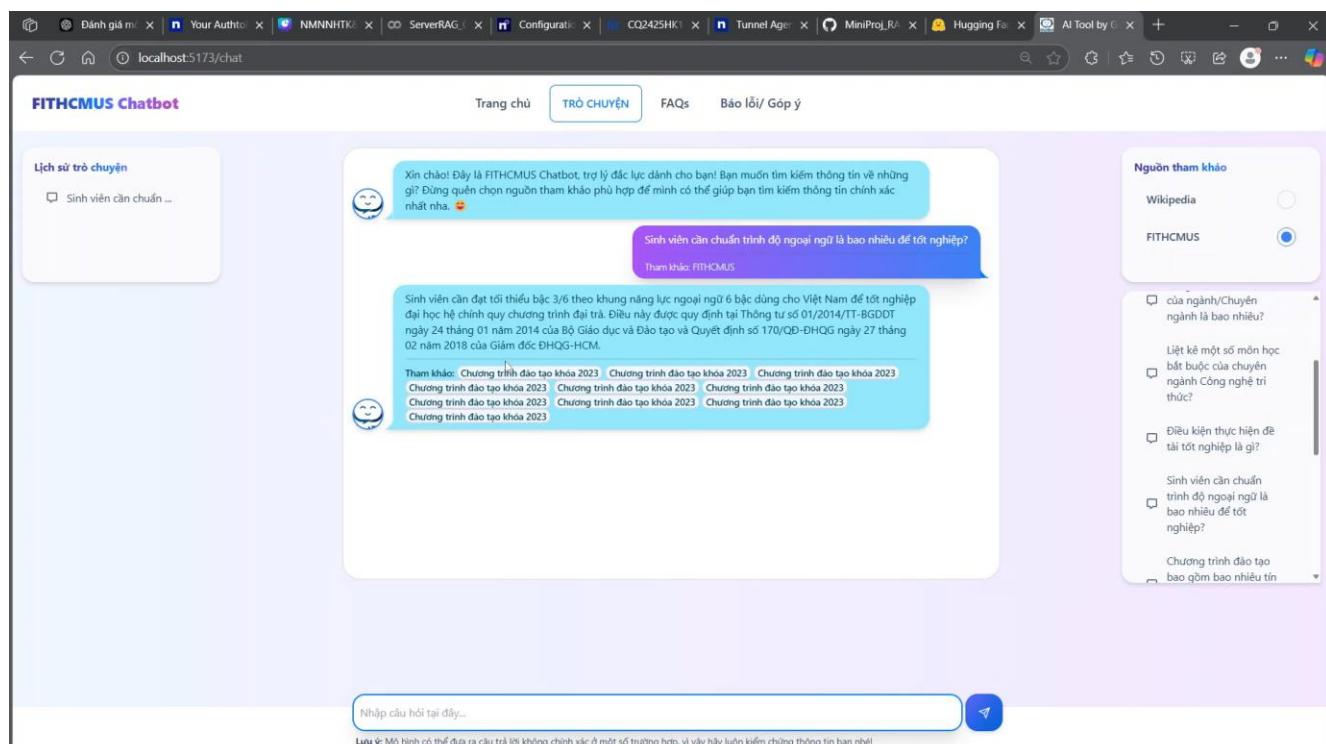
- + Tiếp đến chạy lệnh “*npm run dev*” để chạy chương trình giao diện.

**D:\chatbot> npm run dev**

- + Sau khi chạy sẽ hiển thị lệnh Vite React như sau:

```
VITE v5.4.11 ready in 3110 ms
→ Local: http://localhost:5173/
→ Network: use --host to expose
→ press h + enter to show help
```

- Người dùng truy cập địa chỉ <http://localhost:5173/> để đi đến giao diện web hỏi đáp với chatbot.



*Kết quả hỏi đáp với chatbot*

- Xem chi tiết source code và video demo tại đây:

<https://drive.google.com/drive/folders/1PPkMPZAdiEVXduaSvCBjh3QLbIx9A9pe?usp=sharing>

## VII. Tham khảo

- [1] [ChatGPT Series 5: Tìm hiểu về Retrieval Augmented Generation \(RAG\)](#)
- [2] [Build a Retrieval Augmented Generation \(RAG\) App: Part 1 | LangChain](#)
- [3] [Build a Retrieval Augmented Generation \(RAG\) App: Part 2 | LangChain](#)
- [4] Evaluation of Retrieval-Augmented Generation: A Survey - Hao Yu and Aoran Gan and Kai Zhang and Shiwei Tong and Qi Liu and Zhaofeng Liu
- [5] [RAG Evaluation - Hugging Face Open-Source AI Cookbook](#)