TRƯỜNG ĐẠI HỌC NÔNG LÂM TPHCM

**KHOA CÔNG NGHỆ THÔNG TIN**

**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**Độc lập – Tự do – Hạnh phúc**

TP.Hồ Chí Minh, Ngày 24 tháng 09 năm 2024

**ĐỀ CƯƠNG CHI TIẾT KHÓA LUẬN TỐT NGHIỆP**

1. **Tên đề tài**

Nghiên cứu mô hình ngôn ngữ lớn (Large Language Models) và xây dựng hệ thống chatbot truy vấn dữ liệu dành cho tiếng việt.

1. **Sinh viên thực hiện**

Sinh viên 1:

* MSSV: 21130448
* Tên: Cao Thành Nam
* Lớp: DH21DTC
* Khoa: Công nghệ thông tin
* Khóa: 21
* Số điện thoại: 0839060487

Sinh viên 2:

* MSSV: 21130467
* Tên: Nguyễn Việt Pha
* Lớp: DH21DTC
* Khoa: Công nghệ thông tin
* Khóa: 21
* Số điện thoại: 0982352578

1. **Giảng viên hướng dẫn:** ThS. Nguyễn Đức Công Song.
2. **Lý do chọn đề tài**

* Trong thời đại số hóa hiện nay, lượng dữ liệu văn bản ngày càng gia tăng đột biến, khiến việc trích xuất thông tin và xử lý tài liệu thủ công trở nên khó khăn và kém hiệu quả. Các mô hình ngôn ngữ lớn (Large Language Models - LLMs) đã chứng minh được hiệu quả vượt trội trong việc xử lý ngôn ngữ tự nhiên và trả lời câu hỏi từ các dữ liệu văn bản lớn. Nhu cầu phát triển hệ thống chatbot tự động để hỗ trợ truy vấn dữ liệu dựa trên nội dung văn bản đã trở nên cấp thiết, giúp cải thiện hiệu suất công việc và tối ưu hóa quá trình tìm kiếm thông tin trong lĩnh vực giáo dục đặc biệt là thói quen tìm hiểu sách. Việc ứng dụng các mô hình này không chỉ nâng cao độ chính xác trong trả lời câu hỏi mà còn giúp tiết kiệm thời gian và tài nguyên.
* Nhằm giải quyết các vấn đề này, nhiều mô hình ngôn ngữ lớn đã cung cấp mã nguồn mở. Các dự án nổi tiếng như vibert-base-cased, phobert-base-v2, PhoGPT-4B, Gemini 1.5 Pro,… Đã đóng góp mạnh mẽ, giúp giải quyết vấn đề tạo văn bản, đặc biệt là trong ngữ cảnh tiếng Việt.
* Vì vậy nhóm đã chọn đề tài “Nghiên cứu mô hình ngôn ngữ lớn (Large Language Models) và xây dựng hệ thống chatbot truy vấn dữ liệu dành cho tiếng việt”.

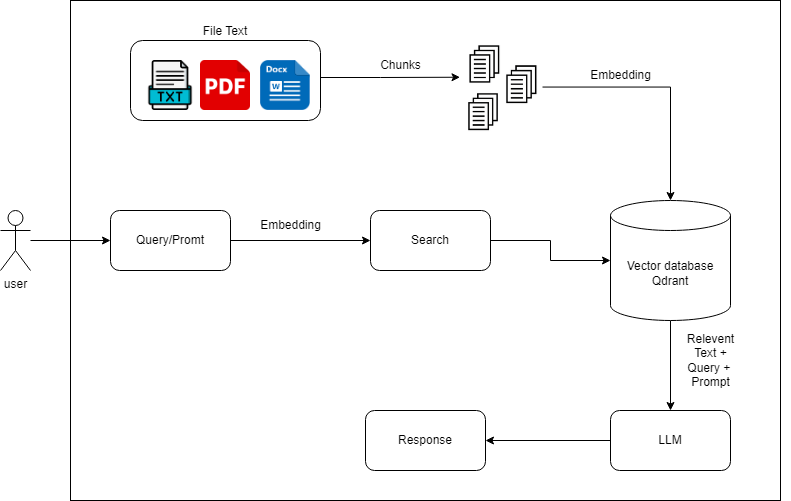
1. **Mục tiêu đề tài**

* Nghiên cứu và đánh giá kết quả của mô hình LLMs Gemini 1.5 Flash trên tập dữ liệu là các cuốn sách.
* Hiểu được quy trình của RAG(Retrieval-Augmented Generation)
* Hiểu được quy trình xử lý dữ liệu của LangChain
* Hiểu được các tương tác và lưu trữ của vector database Qdrant
* Ứng dụng Gemini 1.5 Flash xây dụng ứng dụng chatbot tiếng việt hỗ trợ truy vấn văn bản tiếng việt thông qua .pdf, .docx, .txt

1. **Nội dung và phạm vi nghiên cứu**

* **Nội dung nghiên cứu**
  + Nghiên cứu mô hình LLMs Gemini 1.5 Flash
  + Nghiên cứu framework LangChain
  + Nghiên cứu RAG system
  + Nghiên cứu framework backend FastAPI.
  + Nghiên cứu framework frontend ReactJS.
  + Nghiên cứu vector database Qdrant.
* **Phạm vi nghiên cứu**
  + Nghiên cứu mô hình LLMs Gemini 1.5 Flash để huấn luyện bằng ngôn ngữ tiếng việt.
  + Nghiên cứu RAG giúp tạo ra câu truy vấn hiệu quả và tiết kiệm chi phí
  + Nghiên cứu framework LangChain có thể giúp xây dựng model hiệu quả
  + Nghiên cứu framework backend FastAPI để xây dựng server.
  + Nghiên cứu framework fontend ReactJS để xây dựng giao diện trang chatbot.
  + Nghiên cứu vector database Qdrant để lưu trữ các vector và các thông tin về văn bản.

Sơ đồ xử lý của hệ thống:

****

1. **Cơ sở khoa học và thực tiễn**
   1. **Cơ sở khoa học**

* Các mô hình LLMs là các mô hình được huấn luyện dựa trên dữ liệu ngôn ngữ khổng lồ để thực hiện các nhiệm vụ NLP như hệ thống hỏi đáp, truy vấn dữ liệu, tóm tắt văn bản, phân loại văn bản,…
* Gemini là thành quả của sự hợp tác quy mô lớn từ các đội ngũ tại Google, bao gồm cả nhóm Google Research. Được thiết kế với khả năng đa phương thức ngay từ ban đầu, Gemini có thể xử lý và kết hợp nhiều loại dữ liệu khác nhau như văn bản, mã nguồn, âm thanh, hình ảnh và video một cách mượt mà. Đây là mô hình linh hoạt nhất mà Google từng phát triển, có khả năng hoạt động hiệu quả trên nhiều nền tảng, từ các trung tâm dữ liệu lớn đến các thiết bị di động. Với những tính năng tiên tiến, Gemini hứa hẹn sẽ nâng cao đáng kể cách các nhà phát triển và doanh nghiệp xây dựng và triển khai AI.
* FastAPI là một khung web hiện đại, hiệu suất cao để xây dựng các API bằng Python, tận dụng các gợi ý kiểu của Python tiêu chuẩn nhằm tăng cường tính dễ sử dụng và hiệu quả.
* LangChain là một framework được thiết kế để đơn giản hóa việc phát triển các ứng dụng dựa trên các mô hình ngôn ngữ lớn (LLMs). Nó hỗ trợ tối ưu hóa mọi giai đoạn trong vòng đời ứng dụng LLM, từ phát triển, triển khai đến mở rộng quy mô.
* ReactJS, thường được gọi là React, là một thư viện JavaScript phổ biến để xây dựng giao diện người dùng, đặc biệt dành cho việc tạo ra các ứng dụng web động và tương tác. Được phát triển bởi Facebook, React được sử dụng rộng rãi để xây dựng các ứng dụng front-end.
* Qdrant là một hệ thống tìm kiếm tương đồng vector, sẵn sàng cho triển khai sản xuất, với API dễ sử dụng, cho phép lưu trữ, tìm kiếm và quản lý các vector cùng với dữ liệu bổ sung. Dữ liệu bổ sung này có thể được xem như thông tin thêm giúp cải thiện quá trình tìm kiếm và cung cấp những thông tin giá trị cho người dùng.
  1. **Cơ sở thực tiễn**
* Hệ thống chatbot truy vấn dữ liệu nhằm phục vụ cho lĩnh vực giáo dục, tạo điều kiện thuận lợi cho việc khuyến khích và hiểu được nội dung của cuốn sách nhanh chóng và tiết kiệm thời gian.
* Đề tài nhằm nghiên cứu mô hình LLMs xử lý tiếng việt để áp dụng truy vấn dữ liệu. Đây được coi là đề tài phục vụ cho việc tham khảo và nghiên cứu thêm về mô hình ngôn ngữ lớn.

1. **Thời gian thực hiện:** 8 tháng (bắt đầu từ HKII năm học 2024-2025)
2. **Sản phẩm của đề tài**

* Đưa ra được bộ tài liệu nghiên cứu, phân tích và so sánh giữa ba mô hình LLMs vibert-base-cased, phobert-base-v2 và PhoGPT-4B được huấn luyện dựa trên tiếng việt.
* Đưa ra được tài liệu áp dụng được model LLMs Gemini 1.5 Flash kết hợp với RAG và lưu trữ trên vector database để xây dựng được chatbot truy vấn dữ liệu
* Đưa ra được life cycle từ lúc huấn luyện cho tới khi hoàn thành trang web chatbot.

1. **Tài liệu tham khảo**

**[1]: Introducing Gemini – Sundar Pichai, Demis Hassabis**

<https://blog.google/technology/ai/google-gemini-ai/#introducing-gemini>

**[2]: FastAPI - Sebastián Ramírez, Marcelo, Alejandra, Sofie, Patrick, Esteban Maya, et al.**

https://fastapi.tiangolo.com/#create-it

**[3]: LangChain - Harrison Chase, et al.**

<https://python.langchain.com/docs/introduction/>

**[4]: React Introduction - Jordan Walke, et al.**

<https://www.geeksforgeeks.org/reactjs-introduction/>

**[5]: What is Qdrant - André Zayarni**

<https://qdrant.tech/documentation/overview/>

**Giảng viên hướng dẫn Sinh viên thực hiện**

(Ký và ghi rõ họ tên) (Ký và ghi rõ họ tên)