ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập - Tự Do - Hạnh Phúc

ĐỀ CƯƠNG CHI TIẾT

TÊN ĐỀ TÀI TIẾNG VIỆT:

XÂY DỰNG HỆ THỐNG HỎI ĐÁP TÀI LIỆU VÀ CƠ SỞ DỮ LIỆU CHO DỰ ÁN CÔNG NGHỆ THÔNG TIN

TÊN ĐỀ TÀI TIẾNG ANH:

BUILDING A DOCUMENT AND DATABASE QUESTION ANSWER SYSTEM FOR INFORMATION TECHNOLOGY PROJECTS

Cán bộ hướng dẫn: ThS. Hà Lê Hoài Trung

Thời gian thực hiện: Từ ngày 3/3/2025 đến ngày 8/6/2025

Sinh viên thực hiện:

Nguyễn Sỹ Lê Hoàng – 21520870

Trần Ngọc Tố Như – 21520385

Nội dung đề tài

1. Giới thiệu đề tài

Sự phát triển mạnh mẽ của các mô hình ngôn ngữ lớn (Large Language Models - LLMs) đã mở ra nhiều cơ hội mới trong việc ứng dụng AI vào cuộc sống và công việc hàng ngày. Các doanh nghiệp hiện nay đang chạy đua trong việc khai thác sức mạnh của AI để tăng năng suất, tối ưu quy trình làm việc và cải thiện chất lượng dịch vụ.

Trong bối cảnh đó, việc xây dựng một hệ thống hỏi đáp thông minh dựa trên tài liệu dự án công nghệ thông tin, bao gồm về tài liệu thiết kế hệ thống, yêu cầu dự án, thiết kế dữ liệu và cơ sở dữ liệu (MySQL, PostgreSQL,...) trở thành một giải pháp tiềm năng. Hệ thống này không chỉ giúp lập trình viên hay người quản lý dự án dễ dàng tra cứu thông

tin, giảm thiểu thời gian tìm kiếm dữ liệu mà còn đảm bảo tính chính xác và nhất quán trong quá trình thực hiện dự án.

Đề tài **"**Xây dựng hệ thống hỏi đáp tài liệu và cơ sở dữ liệu cho dự án công nghệ thông tin**"** có ý nghĩa quan trọng trong việc ứng dụng AI để tối ưu hóa hoạt động nội bộ. Hệ thống này có thể được tích hợp với các công cụ hiện có, hỗ trợ tìm kiếm thông tin nhanh chóng, hiểu ngữ cảnh tốt hơn và cung cấp câu trả lời chính xác dựa trên nguồn dữ liệu đáng tin cậy.

Đồng thời, đề tài này còn giúp nhóm nghiên cứu nắm bắt xu hướng và khám phá tiềm năng của các mô hình ngôn ngữ lớn, mở ra những cơ hội ứng dụng AI tiên tiến trong bối cảnh các tập đoàn công nghệ hàng đầu đang cạnh tranh mạnh mẽ để dẫn đầu trong cuộc đua trí tuệ nhân tạo.

2. Mục tiêu của đề tài

Để thực hiện đề tài XÂY DỰNG HỆ THỐNG HỎI ĐÁP TÀI LIỆU VÀ CƠ SỞ DỮ LIỆU CHO DỰ ÁN CÔNG NGHỆ THÔNG TIN, nhóm đã đặt ra những mục tiêu sau:

- Xây dựng một hệ thống hỏi đáp tài liệu, bao gồm về tài liệu thiết kế hệ thống, yêu cầu dự án, thiết kế dữ liệu,... và cơ sở dữ liệu của dự án công nghệ thông tin. Hệ thống có khả năng hiểu ngữ cảnh câu hỏi, truy xuất thông tin nhanh chóng, cung cấp câu trả lời, tóm tắt hoặc đưa ra truy vấn phù hợp dựa trên nguồn dữ liệu nội bộ của dự án, từ đó đảm bảo tính chính xác và nhất quán trong quá trình thực hiện dự án.
- Tìm hiểu và thử nghiệm các mô hình ngôn ngữ lớn (LLM) phù hợp cho hỏi đáp, đồng thời cho bài toán chuyển đổi ngôn ngữ tự nhiên sang câu lệnh truy vấn cơ sở dữ liệu.
- Tìm hiểu quy trình và tài liệu khi quản lý và phát triển một dự án công nghệ thông tin, cũng như các khó khăn khi quản lý, tìm kiếm thông tin, phân quyền, phối hợp nhóm, cũng như cách tận dụng công cụ hỗ trợ để tối ưu hiệu suất và giảm thiểu rủi ro.
- Xây dựng mô hình RAG [1] với vector database để tối ưu hóa việc lưu trữ và tìm kiếm dữ liệu, tăng độ chính xác cho hệ thống hỏi đáp.
- Tìm hiểu và phát triển cơ chế hỗ trợ tối ưu hóa prompt, giúp người dùng đặt câu hỏi rõ ràng, đầy đủ ngữ cảnh, nâng cao chất lượng câu trả lời từ hệ thống AI.
- Đánh giá hiệu quả của mô hình trong việc xử lý các câu hỏi phức tạp và tối ưu hóa quá trình tìm kiếm thông tin trong dự án.

3. Phạm vi

- Quản lý người dùng & phân quyền: Hỗ trợ đăng nhập/đăng ký, quản lý người dùng theo dự án, phân quyền truy cập theo dự án hoặc tài liệu.
- Hệ thống Bot Hỏi Đáp & Kho Tài Liệu: Bot hỗ trợ hỏi đáp dựa trên tài liệu/quy trình/dữ liệu dự án, có thể mở rộng ra tích hợp các công cụ quản lý dự án (Jira, Slack,...). Kho tài liệu cho phép upload, tóm tắt, lưu trữ, xem trước và sử dụng dữ liệu để bot học.
- Tìm kiếm thông minh: Hỗ trợ tìm kiếm tài liệu, câu hỏi theo chủ đề, từ khóa, thẻ, ngày đăng với khả năng phân loại theo nhiều tiêu chí.
- Ứng dụng mô hình ngôn ngữ lớn (LLM): Hiểu ngữ cảnh để chuyển đổi câu hỏi thành SQL phù hợp và tóm tắt văn bản chính xác bằng các mô hình
- Tích hợp kỹ thuật RAG cùng LLM, kết hợp với vector database để tối ưu hóa khả năng tìm kiếm và lưu trữ dữ liệu.
- Phát triển cơ chế tối ưu hóa prompt, giúp cải thiện chất lượng câu hỏi và tăng độ chính xác của câu trả lời.

4. Phương pháp thực hiện

Nội dung 1: Khảo sát và lựa chọn các mô hình ngôn ngữ lớn

- **Mô tả**: Nghiên cứu và thử nghiệm với các mô hình LLM tiên tiến mã nguồn mở như Llama3 [2], Qwen [3],... các mô hình thương mại như gpt-4o [4], gemini-1.5 [5],... đánh giá hiệu suất và khả năng của chúng trong việc xử lý ngữ nghĩa câu hỏi, tìm kiếm thông tin trong tài liệu và tác vụ text2SQL.
- **Phương pháp**: Sử dụng các mô hình mã nguồn mở thông qua HuggingFace, Ollama, LMStudio hoặc thông các API của OpenAI, Google Gemini.
- **Kết quả dự kiến**: Lựa chọn mô hình LLM phù hợp nhất với mục tiêu của đề tài.

Nội dung 2: Nghiên cứu kỹ thuật RAG [5] và xây dựng cơ sở dữ liệu vector

- **Mô tả**: Nghiên cứu kỹ thuật RAG và các biến thể của nó để chọn mô hình phù hợp cho bài toán hỏi đáp tài liệu. Đồng thời, thiết kế và xây dựng cơ sở dữ liệu vector để lưu trữ và truy xuất thông tin nhanh chóng, hỗ trợ RAG trong việc cung cấp dữ liệu khi mô hình cần trả lời câu hỏi từ tài liệu.
- **Phương pháp**: Một số cơ sở dữ liệu vector nhóm dự kiến sẽ tìm hiểu là Qdrant, Weaviate, Atlas vector database. Thử nghiệm các kỹ thuật RAG và kết hợp với cơ

sở dữ liệu vector để lưu trữ và tối ưu hóa truy vấn dữ liệu, nâng cao hiệu quả trả lời câu hỏi.

- **Kết quả dự kiến**: Xây dựng và lựa chọn mô hình RAG phù hợp, xác định cách thức tích hợp với LLM để cải thiện hiệu suất tìm kiếm và trả lời câu hỏi dựa trên tài liệu. Kết hợp với cơ sở dữ liệu vector có cấu trúc hợp lý, dễ sử dụng và có khả năng mở rộng theo thời gian.

Nội dung 3: Xây dựng mô hình hỏi đáp

- **Mô tả**: Áp dụng mô hình ngôn ngữ lớn để xử lý các câu hỏi của người dùng và trích xuất thông tin từ tài liệu hoặc cơ sở dữ liệu.
- **Phương pháp**: Tiến hành huấn luyện và tinh chỉnh các mô hình ngôn ngữ đã chọn, sử dụng dữ liệu đã thu thập để mô phỏng các câu hỏi và câu trả lời thực tế.
- **Kết quả dự kiến**: Mô hình có khả năng trả lời chính xác và nhanh chóng các câu hỏi từ người dùng.

Nội dung 4: Kiểm thử, đánh giá và cải thiện hệ thống hỏi đáp sử dụng LLM

Mô tả: Nghiên cứu và thử nghiệm các phương pháp tối ưu hóa prompt để cải thiện độ chính xác và hiệu suất của hệ thống hỏi đáp LLM, giúp mô hình hiểu rõ hơn ý định người dùng và tối ưu hóa truy vấn cơ sở dữ liệu.

Phương pháp: Khảo sát và thử nghiệm các kỹ thuật tối ưu prompt, kiểm thử hệ thống với các tập dữ liệu đa dạng, đánh giá độ chính xác, độ phủ và thời gian phản hồi để tối ưu hóa hiệu suất và chất lượng.

Kết quả dự kiến: Phát triển chiến lược tối ưu prompt hiệu quả, giúp hệ thống trả lời chính xác, nhanh chóng và nâng cao độ chính xác trong truy xuất tài liệu và thực hiện truy vấn cơ sở dữ liệu.

Nội dung 5: Xây dựng hệ thống hỏi đáp tài liệu hoàn chỉnh

- **Mô tả:** Tìm hiểu về lập trình web và phương pháp tích hợp khoa học dữ liệu và LLM vào ứng dụng.
- **Phương pháp nghiên cứu**:
    - Tìm hiểu một số công nghệ lập trình web, các thư viện và framework phù hợp với đề tài.

- Tìm hiểu các phương pháp, các framework phục vụ phát triển ứng dụng cùng với khoa học dữ liệu và mô hình ngôn ngữ lớn như LangChain, LangGraph, Streamlit,....

- **Kết quả dự kiến**: Phát triển hệ thống hỏi đáp tài liệu toàn diện, đáp ứng đầy đủ chức năng theo mục tiêu đề ra.

Kế hoạch thực hiện:

Nội dung nghiên cứu	**Thời gian**	**Phân công**
1. Nghiên cứu và thử nghiệm các công nghệ và kỹ thuật để xây dựng hệ thống hỏi đáp và dựa trên LLM.	1 tuần	Cả hai
2. Nghiên cứu và thử nghiệm kỹ thuật RAG cho hỏi đáp tài liệu.	1.5 tuần	Cả hai
3. Nghiên cứu và thử nghiệm LLM cho hỏi đáp cơ sở dữ liệu.	1 tuần	Cả hai
4. Nghiên cứu và thử nghiệm các kỹ thuật tối ưu hóa/hỗ trợ cải thiện truy vấn người dùng.	1 tuần	Cả hai
5. Nghiên cứu các thư viện, framework hỗ trợ quá trình kết hợp khoa học dữ liệu hay các mô hình ngôn ngữ lớn vào ứng dụng.	1.5 tuần	Cả hai
6. Phát triển và hoàn thiện hệ thống.	4 tuần	Cả hai
7. Kiểm thử, hoàn thiện báo cáo và chuẩn bị bảo vệ.	2 tuần	Cả hai
Tổng thời gian: 12 tuần		

Tài liệu tham khảo:

[1] Patrick Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2021. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401

[2] Grattafiori et al. The Llama 3 Herd of Models. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783

[3] Bai et al. Qwen Technical Report. 2023. arXiv: 2309.16609 [cs.CL]. URL: https://arxiv.org/abs/2309.16609

[4] OpenAI et al. GPT-4o System Card. 2024. arXiv: 2410.21276 [cs.CL]. URL: https://arxiv.org/abs/2407.21783

[5] Gemini Team et al. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. 2024. arXiv: 2403.05530 [cs.CL]. URL: https://arxiv.org/abs/2403.05530

Xác nhận của CBHD (Ký tên và ghi rõ họ tên)	**TP. HCM, ngày 21 tháng 2 năm 2025** **Sinh viên 1** (Ký tên và ghi rõ họ tên)	**TP. HCM, ngày 21 tháng 2 năm 2025** **Sinh viên 2** (Ký tên và ghi rõ họ tên)