TRƯỜNG ĐẠI HỌC NÔNG LÂM TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

TP. Hồ Chí Minh, ngày 19 tháng 12 năm 2024

ĐỀ CƯƠNG CHI TIẾT KHOÁ LUẬN TỐT NGHIỆP

1. Tên đề tài:

- Xây dựng chatbot tư vấn học vụ Nông Lâm kết hợp giữa hệ thống GRAG (Graph Retrieval-Augmented Generation) và RAG (Retrieval-Augmented Generation).

2. Sinh viên thực hiện:

Sinh viên 1:

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

Sinh viên 2:

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

3. Giảng viên hướng dẫn: ThS. Nguyễn Đức Công Song

4. Lý do chọn đề tài:

- Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) đã nổi lên trong việc phân loại văn bản đến các nhiệm vụ phức tạp như tóm tắt, dịch tự động và trả lời câu hỏi. Một lĩnh vực đặc biệt quan trọng trong NLP là Tạo sinh ngôn ngữ tự nhiên (Natural Language Generation – NLG). Mục tiêu chính của NLG là giúp máy tính tạo ra văn bản mạch lạc và phù hợp với ngữ cảnh. Với sự tiến bộ của AI, yêu cầu đối với các mô hình tạo sinh nội dung có ngữ cảnh và căn cứ thực tế ngày càng tăng, dẫn đến những thách thức và cải tiến mới trong NLG. Kiến trúc sequence-to-sequence, đã đạt được những bước tiến lớn trong việc tạo ra văn bản tự nhiên và mạch lạc. Tuy nhiên, các mô hình này phụ thuộc nhiều vào dữ liệu huấn luyện và gặp khó khăn khi phải sản xuất thông tin chính xác hay giàu ngữ cảnh trong những trường hợp yêu cầu kiến thức vượt ra ngoài phạm vi dữ liệu huấn luyện.

- Để giải quyết vấn đề này đã có các nghiên cứu nói về hệ thống RAG(Retrievalaugmented generation) hỗ trợ cập nhập dữ liệu mới mà không cần huấn luyện lại từ đầu bằng phương pháp trích xuất dữ liệu từ PDF sau đó lưu trữ trong vector database từ đó người dùng có thể truy xuất thông tin mà mô hình LLM chưa được huấn luyện. Hệ thống RAG thông thường có nhược điểm sau:
    - Bỏ qua các mối quan hệ trong dữ liệu và không thể trích xuất thông tin một cách tổng quát.
    - Đối với các truy xuất yêu cầu tài liệu bán cấu trúc cần có giải pháp riêng, nhưng RAG chỉ truy xuất được các yêu cầu tài liệu không cấu trúc.
- Đối với GRAG(Graph Retrieval-Augmented Generation) lại chỉ có thể truy xuất được dữ liệu có cấu trúc. Điều này tạo ra hai thách thức lớn sau:
    - Cần phải tạo ra một hệ thống duy nhất vừa truy xuất được nguồn dạng không cấu trúc, cấu trúc và kết hợp.
    - Thậm chí nếu xây dựng được hệ thống, đối với các câu hỏi phức tạp có nhiều phần. Hệ thống có thể trả lời sai ngay từ đầu do đó cần phải có module để phản hồi và nhận xét câu trả lời.
- Các nghiên cứu gần đây đã chỉ ra rằng việc kết hợp giữa RAG và GRAG tạo ra Agentic RAG có thể giải quyết hai vấn đề trên. Vì vậy nhóm đã chọn đề tài "Xây dựng chatbot tư vấn học vụ Nông Lâm kết hợp giữa hệ thống GRAG (Graph Retrieval-Augmented Generation) và RAG (Retrieval-Augmented Generation)".

  Điểm mới của đề tài:
    - Kết hợp hai kiến trúc truy xuất không cấu trúc và có cấu trúc.
    - Thêm critic module để phản ánh câu trả lời có tốt hay chưa từ đó điều chỉnh câu trả lời đúng ngữ cảnh hơn.

5. Mục tiêu của đề tài:

- Sử dụng LLM Gemini-1.5-flash để làm base model.
- Tạo ra hệ thống kết hợp giữa RAG và GRAG xây dựng hệ thống chatbot tư vấn học vụ Nông Lâm.
- Đánh giá hệ thống trên các tập dataset StaRK-MAG, StaRK-Prime, CRAG và ZaloAI sử dụng các metric Hit@1, Hit@5, Recall@20m, MRR, Person Cosine và Spearman Cosine để so sánh với mô hình của đề tài trước. Chứng minh rằng kiến trúc này có khả năng trả lời câu hỏi tốt hơn kiến trúc của mô hình cũ.

6. Nội dung và phạm vi nghiên cứu:

a. Nội dung nghiên cứu

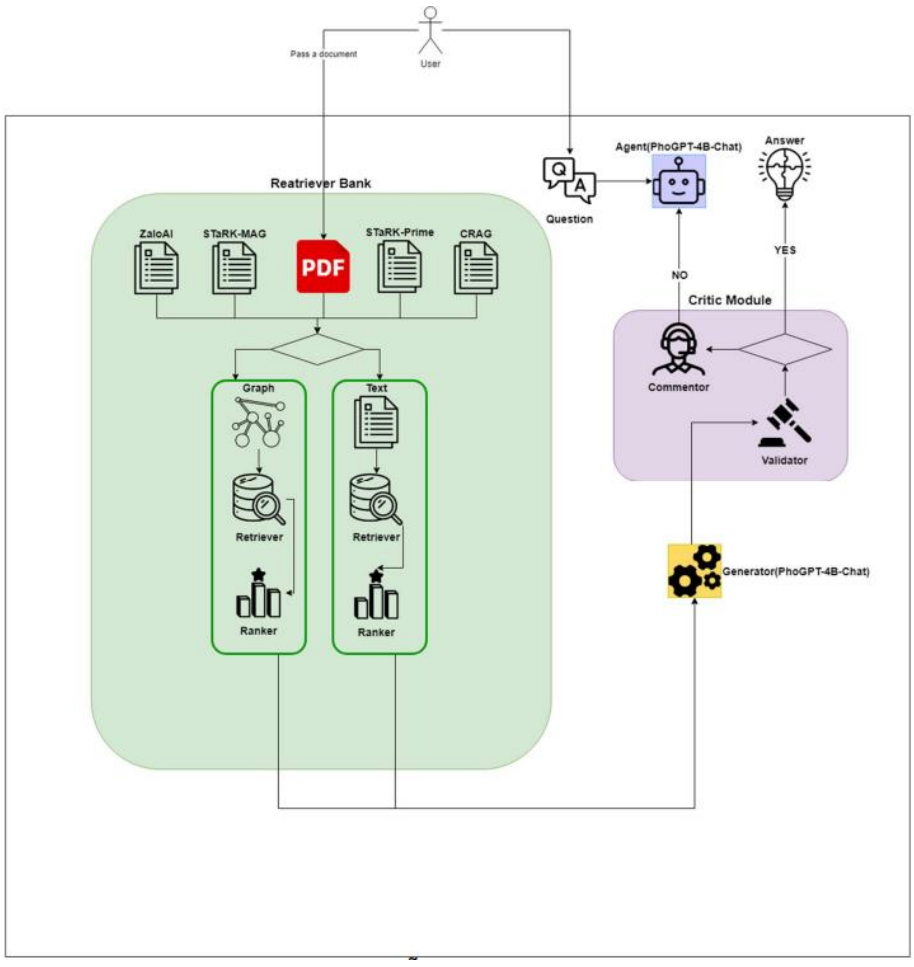
- Nghiên cứu mô hình LLM Gemini-1.5-flash.
- Nghiên cữu RAG và GRAG.

- Nghiên cứu ba dataset StaRK-MAG, StaRK-Prime, CRAG và ZaloAI để đánh giá hệ thống sử dụng Hit@1, Hit@5, Recall@20m, MRR, Person Cosine và Spearman Cosine.
- Nghiên cứu framework LangChain.
- Nghiên cứu thư viện pyvis.
- Nghiên cứu vector database Qdrant.

b. Phạm vi nghiên cứu

- Dùng LLM Gemini-1.5-flash để làm base model.
- Áp dụng dataset StaRK-MAG, StaRK-Prime, CRAG và ZaloAI để đánh giá hiệu suất của hệ thống sử dụng các metric Hit@1, Hit@5, Recall@20m, MRR, Person Cosine và Spearman Cosine.
- Áp dụng dataset học vụ Nông Lâm trên Hugggingface và bổ sung thêm dữ liệu nếu cần thiết.
- Áp dụng kết hợp hệ thống GRAG và RAG để xây dựng chatbot tư vấn học vụ Nông Lâm.
- Áp dụng framework Langchain giúp xây dựng model LLM dễ dàng.
- Áp dụng thư viện pyvis để vẽ Knowledge Graph.
- Áp dụng vector database Qdrant để lưu trữ các embedding xây dựng reatriever bank.

Sơ đồ xử lý của hệ thống:



7. Cơ sở khoa học và thực tiễn:

a. Cơ sở khoa học

- Gemini-1.5-flash là một mô hình trí tuệ nhân tạo tiên tiến thuộc dòng Gemini của Google, được phát triển dựa trên nền tảng Transformer với khả năng học sâu (Deep Learning) và học tăng cường từ phản hồi con người (Reinforcement Learning from Human Feedback - RLHF). Mô hình này cho phép cải thiện độ chính xác, sự hiểu ngữ cảnh và tính sáng tạo. Được huấn luyện trên lượng dữ liệu lớn từ nhiều nguồn. Gemini-1.5-flash có khả năng phân tích, tổng hợp và tạo nội dung phù hợp trong các lĩnh vực đa dạng, từ xử lý ngôn ngữ tự nhiên (NLP) đến các ứng dụng AI chuyên sâu, đồng thời tối ưu hóa hiệu suất tính toán để đáp ứng nhu cầu thực tiễn.
- RAG(Retrieval Augmented Generation) là một hệ thống cho phép mô hình có thể truy xuất vào tài liệu nội bộ ở dạng không cấu trúc, cho phép mô hình cải thiện hiệu suất, chất lượng của truy vấn theo từng hoàn cảnh khác nhau, mà dữ liệu đó chưa được huấn luyện trước.
- GRAG(Graph Retrieval Augmented Generation) là một phương pháp mới của Microsoft Research nghiên cứu nhằm giải quyết giới hạn của RAG trong nhiệm vụ QFS bằng cách truy xuất vào tài liệu ở dạng có cấu trúc, sử dụng LLM để tạo ra một đồ thị tri thức dựa trên tài liệu nội bộ thông qua hai giai đoạn trích xuất một đồ thị kiến thức thực thể từ các tài liệu nguồn, sau đó tạo trước các bản tóm tắt cộng đồng cho tất cả các nhóm thực thể có liên quan chặt chẽ. Điều này đã chứng minh hiệu suất cải thiện đáng kể so với phương pháp RAG cơ bản về cả tính toàn diện và tính đa dạng của các câu trả lời được tạo ra.
- Kết hợp giữa RAG và GRAG tạo nên hệ thống Agentic RAG, có thể tự cải thiện hành động của Agent thông qua cơ chế tự phản hồi, đồng thời có thể giải quyết những câu hỏi yêu cầu nguồn kiến thức kết hợp trong một hệ thống duy nhất. Điều này giúp giảm lỗi biện minh và giảm thiểu ảo tưởng cho LLM. GRAG đã chứng minh vượt trội hơn tất cả các hệ thống RAG cơ sở thông thường trên tập dataset StaRK-MAG, StaRK-Prime và CRAG trên các metric Hit@1, Hit@5, Recall@20m, MRR.

b. Cơ sở thực tiễn

- Đề tài phát triển nhằm mục đích nghiên cứu LLM, đánh giá và kết hợp RAG và GRAG trên tập dữ liệu là học vụ Nông Lâm.
- Xây dựng ứng dụng chatbot tiếng việt hỗ trợ sinh viên Nông Lâm.
- Đây là đề tài kế thừa của đề tài "Nghiên cứu mô hình ngôn ngữ lớn (Large Language Models) và ứng dụng Chatbot tiếng việt hỗ trợ giải đáp sinh viên Nông Lâm" trong HK2/2023-2024.

Tồn đọng của đề tài cũ:

- Đối với các câu hỏi yêu cầu truy xuất thông tin có cấu trúc hoặc kết hợp có cấu trúc và không cấu trúc sẽ trả lời không đúng vì truy xuất sai nguồn dữ liệu.
- Có thể bị trả lời sai ngay lần đầu tiên.

Hướng giải quyết:

- Cần xây dựng hai cơ sở dữ liệu để chữa dữ liệu không cấu trúc và có cấu trúc.

- Sử dụng thêm critic module làm trung gian để có thể phản ảnh câu trả lời có được tốt hay không, từ đó model sẽ ra quyết định chỉnh sửa hoặc truy xuất vào nguồn dữ liệu phù hợp.

Tính mới của đề tài:

- Kết hợp hai kiến trúc truy xuất không cấu trúc và có cấu trúc
- Critic module để phản ánh câu trả lời có tốt hay chưa từ đó điều chỉnh câu trả lời đúng ngữ cảnh hơn.

8. **Thời gian thực hiện:** 6 tháng (bắt đầu từ HKII năm học 2024-2025).
9. **Sản phẩm của đề tài:**

- Hệ thống chatbot giúp hỗ trợ sinh viên thực hiện tư vấn học vụ Nông Lâm.
- Đưa ra được tài liệu chi tiết áp dụng được model LLM Gemini-1.5-flash làm base model trong hệ thống kết hợp RAG và GRAG và lưu trữ trên vector database để xây dựng được chatbot truy vấn dữ liệu.

10. **Tài liệu tham khảo:**

[1]: A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions - Shailja Gupta, Rajesh Ranjan, Surya Narayan Singh

https://arxiv.org/abs/2410.12837

[2]: Agent-G: An Agentic Framework for Graph Retrieval Augmented Generation: Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, Christos Faloutsos

https://openreview.net/forum?id=g2C947jjjQ

[3]: STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases: Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N. Ioannidis, Karthik Subbian, James Zou, Jure Leskovec, Department of Computer Science, Stanford University Amazon

https://arxiv.org/abs/2404.13207

[4]: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks: Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela

https://arxiv.org/abs/2005.11401

[5]: From Local to Global: A Graph RAG Approach to Query-Focused Summarization: Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Jonathan Larson

https://arxiv.org/abs/2404.16130

[6]: Introducing Gemini: our largest and most capable AI model: Sundar Pichai, Demis Hassabis

https://blog.google/technology/ai/google-gemini-ai/

Giảng viên hướng dẫn	**Sinh viên thực hiện**
(Ký và ghi rõ họ tên)	*(Ký và ghi rõ họ tên)*