KHOA KỸ THUẬT VÀ CÔNG NGHỆ
BỘ MÔN CÔNG NGHỆ THÔNG TIN

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

ĐỀ CƯƠNG CHI TIẾT
ĐỒ ÁN TỐT NGHIỆP NGÀNH CÔNG NGHỆ THÔNG TIN

Họ tên sinh viên: **Nguyễn Đại Hoàng Phúc** MSSV: 110121087

Lớp: Công nghệ Thông tin B Khóa: 2021-2025

Tên đề tài: Nghiên cứu RAG và xây dựng chatbot hỗ trợ môn Cơ sở dữ liệu.

1. Mục tiêu của đề tài:

- Nghiên cứu lý thuyết và kiến trúc Retrieval-Augmented Generation (RAG), các thành phần cốt lõi (embedding, vector database, large language models - LLMs), và các kỹ thuật nâng cao như semantic search và reranking.
- Áp dụng các kỹ thuật Xử lý Ngôn ngữ Tự nhiên (NLP) vào việc tiền xử lý tài liệu và tạo embedding.
- Xây dựng cơ sở tri thức từ tài liệu môn Cơ sở dữ liệu (đa định dạng) dưới dạng vector embeddings.
- Thiết kế và xây dựng hệ thống chatbot RAG hoàn chỉnh, sử dụng semantic search (với Qdrant) và kỹ thuật reranking, có khả năng trả lời câu hỏi chính xác và trích dẫn nguồn.
- Phát triển giao diện người dùng web (dự kiến Next.js) thân thiện, hỗ trợ quản lý tài liệu và tương tác với chatbot.
- Tích hợp giải pháp lưu trữ (dự kiến Supabase) cho lịch sử hội thoại và dữ liệu người dùng.

2. Nội dung thực hiện:

Chương 1: Tổng quan:

- Giới thiệu chung về chatbot, AI, NLP và kỹ thuật RAG là giải pháp cho các hạn chế của LLM truyền thống.
- Nêu bật nhu cầu cần một công cụ hỗ trợ 24/7 cho môn Cơ sở dữ liệu và vai trò của chatbot RAG như một "trợ giảng ảo" đáng tin cậy.

- Trình bày mục tiêu xây dựng hệ thống chatbot, đối tượng và phạm vi nghiên cứu trong môn CSDL, cùng cấu trúc 6 chương của khóa luận.

Chương 2: Cơ sở lý thuyết:

- Trình bày các khái niệm NLP và các bước tiền xử lý văn bản tiếng Việt như làm sạch, tách từ và phân đoạn (chunking).
- Giới thiệu kỹ thuật Embedding, tập trung vào Sentence Transformers để tạo vector ngữ nghĩa cho văn bản.
- Mô tả vai trò của Vector Database, tập trung vào Qdrant để lưu trữ và thực hiện tìm kiếm ngữ nghĩa (semantic search) hiệu quả.
- Giới thiệu kiến trúc Transformer của Large Language Models (LLMs) và mô hình Gemini sẽ được sử dụng.
- Phân tích kiến trúc Retrieval-Augmented Generation (RAG), bao gồm quá trình Semantic Search và bước Reranking để tối ưu hóa kết quả truy xuất.
- Tóm tắt nội dung chính của môn Cơ sở dữ liệu và liệt kê các công nghệ dự kiến sử dụng: Next.js, FastAPI, và Supabase.

Chương 3: Phân tích và thiết kế hệ thống:

- Phân tích yêu cầu:
    - Chức năng: Hỏi đáp CSDL, trích dẫn nguồn, quản lý tài liệu, lịch sử hội thoại, gợi ý câu hỏi, phân tích SQL.
    - Phi chức năng: Hiệu năng, thân thiện, dễ sử dụng.
- Thiết kế kiến trúc tổng thể:
    - Modules chính: Backend (FastAPI - Xử lý tài liệu, Embedding, Tương tác Vector Store, Lõi RAG, API), Frontend (Next.js - Giao diện người dùng, Tương tác API), Database (Supabase - Lưu trữ dữ liệu).
    - Luồng dữ liệu và tương tác giữa các module.
- Thiết kế cơ sở tri thức:
    - Chiến lược tiền xử lý và phân đoạn (Chunking) cho các loại tài liệu, tối ưu hóa truy xuất.
    - Cấu trúc lưu trữ Vector Store (Qdrant) với metadata phù hợp.

- Thiết kế luồng xử lý RAG: Tiền xử lý câu hỏi -> Embedding -> Semantic Search trên Qdrant -> Áp dụng Reranking -> Chuẩn bị Prompt (với ngữ cảnh) -> Gọi LLM -> Hậu xử lý và trích dẫn.
- Thiết kế giao diện người dùng (UI/UX): Xác định các API chính và phác thảo giao diện người dùng trực quan, dễ sử dụng.

Chương 4: Triển khai hệ thống:

- Lựa chọn công nghệ: Python, FastAPI, Next.js, TypeScript, Sentence Transformers, Qdrant, Gemini API, Supabase, thư viện NLP và xử lý tài liệu, model Reranker.
- Xây dựng Backend: Triển khai các module xử lý tài liệu, embedding, truy xuất ngữ cảnh (thực hiện cơ chế semantic search và reranking trên Qdrant), sinh câu trả lời, hậu xử lý, API và tích hợp Supabase.
- Xây dựng Frontend: Triển khai giao diện người dùng và logic tương tác với backend.
- Tích hợp hệ thống và Đóng gói/Triển khai: Kết nối các thành phần, nghiên cứu sử dụng Docker/Docker Compose.

Chương 5: Đánh giá và kết quả:

- Xây dựng bộ câu hỏi kiểm thử: Chuẩn bị bộ câu hỏi (khoảng 20-30 câu) đa dạng về các chủ đề trong môn Cơ sở dữ liệu.
- Phương pháp đánh giá:
    - Độ liên quan của câu trả lời: Với từng câu hỏi trong bộ kiểm thử, chạy chatbot và đánh giá câu trả lời theo thang điểm (ví dụ: 0-Không liên quan, 1-Liên quan một phần, 2-Hoàn toàn liên quan).
    - Kiểm tra tính đúng đắn dựa trên nguồn (cơ bản): Với các câu trả lời được đánh giá cao về độ liên quan, kiểm tra nhanh xem nội dung có khớp với nguồn trích dẫn không (Đánh dấu: Có/Không).
    - Tốc độ phản hồi: Ghi nhận thời gian chatbot trả lời cho mỗi câu hỏi.
- Phân tích kết quả: Tổng hợp số liệu từ đánh giá (tỷ lệ câu trả lời liên quan, tỷ lệ khớp nguồn, tốc độ trung bình).
- Kết luận.

Chương 6: Kết luận và hướng phát triển:

- Tóm tắt kết quả, đóng góp của đề tài.
- Hạn chế của hệ thống và nghiên cứu.
- Đề xuất hướng phát triển: Cải thiện độ chính xác, chunking, mở rộng tính năng, tối ưu hóa, đánh giá toàn diện hơn.

3. Phương pháp thực hiện:

- Nghiên cứu lý thuyết về RAG, NLP, LLMs và các công nghệ liên quan (Qdrant, Sentence Transformers, semantic search, reranking).
- Phân tích, thiết kế hệ thống theo phương pháp phát triển phần mềm.
- Lập trình và phát triển sử dụng các công nghệ đã chọn.
- Thử nghiệm, kiểm thử các thành phần cốt lõi.
- Đánh giá thực nghiệm thông qua các chỉ số tổng hợp.

4. Bố cục đề tài:

Đề tài dự kiến được chia thành 6 chương như đã trình bày chi tiết trong Mục 2 (Nội dung thực hiện).

- Chương 1: Tổng quan
- Chương 2: Cơ sở lý thuyết
- Chương 3: Phân tích và thiết kế hệ thống
- Chương 4: Triển khai hệ thống
- Chương 5: Đánh giá và kết quả
- Chương 6: Kết luận và hướng phát triển

5. Tài liệu tham khảo:

[1] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *arXiv preprint* arXiv:2005.11401, May 2020. [Online]. Available: https://arxiv.org/abs/2005.11401. Accessed: May 5, 2025

[2] W. Yeo, K. Kim, S. Jeong, J. Baek, and S. J. Hwang, “UniversalRAG: Retrieval-augmented generation over corpora of diverse modalities and granularities,” *arXiv preprint* arXiv:2504.20734, Apr.2025. [Online]. Available: https://arxiv.org/abs/2504.20734. Accessed: May 5, 2025

[3] Q. D. Nguyen, H. S. Le, D. N. Nguyen, D. N. N. Nguyen, T. H. Le, and V. S. Dinh, “Towards comprehensive Vietnamese retrieval-augmented generation and large language models,” *arXiv preprint* arXiv:2403.01616, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.01616. Accessed: May 5, 2025

[4] Y. Gao et al., “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint* arXiv:2312.10997, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2312.10997. Accessed: May 5, 2025

[5] Z. Jiang et al., “Active retrieval-augmented generation,” *arXiv preprint* arXiv:2305.06983, May 2023. [Online]. Available: https://arxiv.org/abs/2305.06983. Accessed: May 5, 2025

[6] D. Lin, “Revolutionizing retrieval-augmented generation with enhanced PDF structure recognition,” *arXiv preprint* arXiv:2401.12599, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.12599. Accessed: May 5, 2025

[7] A. A. Khan, M. T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, “Developing retrieval-augmented generation (RAG)-based LLM systems from PDFs: An experience report,” *arXiv preprint* arXiv:2410.15944, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.15944. Accessed: May 5, 2025

[8] C. Alario-Hoyos, R. Kemcha, C. D. Kloos, P. Callejo, I. Estévez-Ayres, D. Santín-Cristóbal, F. Cruz-Argudo, and J. L. López-Sánchez, “Tailoring your code companion: Leveraging LLMs and RAG to develop a chatbot to support students in a programming course,” in *Proc. IEEE Int. Conf. Teaching, Assessment and Learning for Engineering (TALE)*, Bengaluru, India, Dec. 9–12, 2024, pp. 1–8. [Online]. Available: https://dblp.org/rec/conf/tale/Alario-HoyosKKC24. Accessed: May 5, 2025

[9] G. Lang and T. Gürpinar, “AI-powered learning support: A study of retrieval-augmented generation (RAG) chatbot effectiveness in an online course,” *Information Systems Education Journal*, vol. 23, no. 2, pp. 4–13, Mar. 2025. [Online]. Available: https://files.eric.ed.gov/fulltext/EJ1467995.pdf. Accessed: May 5, 2025

[10] H. A. Modran, I. C. Bogdan, D. Ursuțiu, C. Samoilă, and P. L. Modran, "LLM intelligent agent tutoring in higher education courses using a RAG approach," in *Proc. Int. Conf. Interactive Collaborative Learning (ICL)*, Tallinn, Estonia, Sep. 24–27, 2024, pp. 589– 599. [Online]. Available: https://www.preprints.org/manuscript/202407.0519/v1. Accessed: May 5, 2025

6. Kế hoạch thực hiện đề tài:

Tuần	Từ ngày - đến ngày	Công việc thực hiện	Ghi chú
1	Từ ngày 07/4/2025 đến ngày 13/4/2025	- Nghiên cứu sâu lại lý thuyết RAG, Embedding (Sentence Transformers), Vector DB (Qdrant), LLM API (Gemini). - Hoàn thiện kế hoạch chi tiết, xác định rõ phạm vi cuối cùng. - Tìm hiểu các tiêu chuẩn trình bày khóa luận. - Thiết lập môi trường phát triển cơ bản.	Nguyễn Đại Hoàng Phúc
2	Từ ngày 14/4/2025 đến ngày 20/4/2025	- Thu thập tài liệu môn Cơ sở dữ liệu (giáo trình, slide, bài tập,...). - Bắt đầu tiền xử lý dữ liệu: Trích xuất text từ PDF, DOCX, TXT, SQL,... - Nghiên cứu và lựa chọn chiến lược phân đoạn. - Đăng ký và cấu hình API keys (Qdrant, Google Gemini).	Nguyễn Đại Hoàng Phúc
3	Từ ngày 21/4/2025 đến ngày 27/4/2025	- Hoàn thiện việc tiền xử lý và làm sạch dữ liệu.	Nguyễn Đại Hoàng Phúc

		- Xây dựng module Data Ingestion với chức năng đọc, xử lý và phân đoạn tài liệu theo chiến lược đã chọn. - Kiểm thử module Ingestion với các loại tài liệu khác nhau. - Tạo cấu trúc thư mục dự án chuẩn.	
4	Từ ngày 28/4/2025 đến ngày 04/5/2025	- Nghiên cứu và lựa chọn mô hình Sentence Transformer phù hợp. - Xây dựng module Embedding: Tạo vector embeddings cho các đoạn tài liệu. - Tích hợp Qdrant: Cấu hình collection, tải embeddings và metadata lên Qdrant. - Kiểm thử quá trình tạo và lưu trữ embeddings.	Nguyễn Đại Hoàng Phúc
5	Từ ngày 05/5/2025 đến ngày 11/5/2025	- Xây dựng module Vector Retrieval: Tạo embedding cho câu hỏi người dùng, thực hiện tìm kiếm trên Qdrant. - Lấy ra các đoạn ngữ cảnh (context) liên quan nhất. - Kiểm thử hiệu quả của việc truy xuất ngữ cảnh.	Nguyễn Đại Hoàng Phúc
6	Từ ngày 12/5/2025 đến ngày 18/5/2025	- Xây dựng module Generation: Tích hợp Google Gemini API. - Thiết kế và thử nghiệm các prompt template để kết hợp câu hỏi và ngữ cảnh truy xuất được. - Sinh câu trả lời từ LLM. - Xây dựng module Postprocessing cơ bản.	Nguyễn Đại Hoàng Phúc

		- Kiểm thử luồng RAG hoàn chỉnh đầu tiên.	
7	Từ ngày 19/5/2025 đến ngày 25/5/2025	- Thiết kế kiến trúc API. - Xây dựng Backend API sử dụng FastAPI để đóng gói luồng RAG. - Triển khai các endpoint chính. - Viết các unit test cơ bản cho API.	Nguyễn Đại Hoàng Phúc
8	Từ ngày 26/5/2025 đến ngày 01/6/2025	- Thiết kế giao diện người dùng (UI) web đơn giản. - Xây dựng Frontend sử dụng Next.js. - Tích hợp Frontend với Backend API để gửi câu hỏi và hiển thị câu trả lời. - Kiểm thử luồng tương tác người dùng.	Nguyễn Đại Hoàng Phúc
9	Từ ngày 02/6/2025 đến ngày 08/6/2025	- Tích hợp toàn bộ hệ thống. - Triển khai CICD lên host sử dùng Github Action và chạy với docker-compose. - Đánh giá hệ thống. - Phân tích kết quả, xác định ưu/nhược điểm. - Thực hiện các tinh chỉnh nhỏ và sửa lỗi dựa trên kết quả đánh giá.	Nguyễn Đại Hoàng Phúc
10	Từ ngày 09/6/2025 đến ngày 15/6/2025	- Hoàn thiện bản thảo cuối cùng của khóa luận tốt nghiệp. - Định dạng tài liệu theo đúng quy định. - Chuẩn bị slide thuyết trình và demo sản phẩm. - Duyệt lại toàn bộ khóa luận, kiểm tra lỗi chính tả, định dạng. - Hoàn tất nộp khóa luận tốt nghiệp.	Nguyễn Đại Hoàng Phúc

Trà Vinh, ngày tháng 4 năm 2025

GIẢNG VIÊN HƯỚNG DẪN	SINH VIÊN THỰC HIỆN
TS. Nguyễn Bảo Ân	**Nguyễn Đại Hoàng Phúc**