TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI

KHOA CÔNG NGHỆ THÔNG TIN

……

MÔN HỌC: CÔNG NGHỆ PHÁT TRIỂN ỨNG DỤNG DOANH NGHIỆP

ĐỀ TÀI: TÍCH HỢP AL HỎI ĐÁP, HỖ TRỢ BÀI BÁO VÀ DỊCH VĂN BẢN

Giảng viên hướng dẫn: Nguyễn Trọng Phúc

Nhóm: 4 – KS CNTT 2 – K62

Thành viên:

STT	Họ và Tên	Mã Sinh Viên
1	Chu Văn Dũng	211210740
2	Hồ Anh Minh	211211807
3	Nguyễn Trần Công Cường	211214539
4	Phạm Hải Nhi	211203484

Hà Nội, tháng 9 năm 2025

LỜI NÓI ĐẦU

Trong thời đại công nghệ 4.0, lượng thông tin học thuật và khoa học tăng trưởng vượt bậc. Mỗi ngày có hàng nghìn bài báo nghiên cứu, tạp chí, báo cáo kỹ thuật mới được công bố. Đây là kho tàng tri thức vô cùng quý giá, nhưng cũng tạo ra áp lực không nhỏ cho sinh viên, giảng viên và nhà nghiên cứu khi cần cập nhật liên tục.

Một trong những rào cản lớn là thời gian đọc và xử lý. Việc phải đọc hàng chục trang để rút ra vài ý chính khiến hiệu quả nghiên cứu giảm sút. Bên cạnh đó, rào cản ngôn ngữ cũng là một thách thức: phần lớn tài liệu học thuật ở tiếng Anh hoặc các ngôn ngữ khác, khiến người học trong nước khó tiếp cận.

Những công cụ hiện có như Google Translate, ChatGPT hay Perplexity giúp phần nào, nhưng vẫn còn nhiều giới hạn: dịch thuật chưa theo ngữ cảnh, thiếu khả năng hỏi–đáp dựa trên tài liệu cụ thể, phụ thuộc vào cloud dẫn đến rủi ro bảo mật và chi phí cao.

Từ thực tế đó, nhóm chúng tôi xây dựng đề tài **“Tích hợp AI hỏi đáp, hỗ trợ bài báo và dịch văn bản”**, nhằm phát triển một hệ thống chạy local (trên máy cá nhân), vừa đảm bảo quyền riêng tư dữ liệu, vừa hỗ trợ người dùng hỏi đáp, tóm tắt và dịch tài liệu học thuật một cách hiệu quả.

MỤC LỤC

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Trong bối cảnh toàn cầu hóa và sự phát triển nhanh chóng của khoa học công nghệ, tri thức nhân loại đang tăng trưởng với tốc độ chưa từng có. Mỗi ngày, hàng nghìn bài báo khoa học, báo cáo kỹ thuật, và tài liệu học thuật được công bố trên khắp thế giới. Đây là nguồn tài nguyên quý giá cho sinh viên, giảng viên và các nhà nghiên cứu, nhưng đồng thời cũng đặt ra những thách thức không nhỏ:

- **Khối lượng thông tin khổng lồ**: Việc phải tiếp cận và xử lý hàng chục, thậm chí hàng trăm tài liệu trong một nghiên cứu khiến người học dễ rơi vào tình trạng quá tải, khó khăn trong việc chọn lọc và tổng hợp thông tin.
- **Rào cản ngôn ngữ**: Phần lớn tài liệu học thuật chất lượng cao được viết bằng tiếng Anh hoặc các ngôn ngữ khác, trong khi không phải người học nào cũng có đủ khả năng ngôn ngữ để tiếp cận.
- **Hạn chế của công cụ hiện tại**: Các phần mềm dịch hoặc tìm kiếm truyền thống tuy hữu ích, nhưng thường không hỗ trợ hỏi đáp ngữ cảnh, không cung cấp khả năng hội thoại liên tục, và phần lớn phụ thuộc vào nền tảng đám mây, gây ra các vấn đề về bảo mật dữ liệu và chi phí sử dụng.

Những khó khăn trên cho thấy nhu cầu cấp thiết về một hệ thống thông minh, có khả năng hỏi đáp trực tiếp dựa trên tài liệu, tóm tắt thông tin nhanh chóng, dịch thuật theo ngữ cảnh chuyên ngành, đồng thời có thể vận hành trên môi trường local để đảm bảo quyền riêng tư và tiết kiệm chi phí. Đây chính là động lực để nhóm chúng tôi lựa chọn và triển khai đề tài **"Tích hợp AI hỏi đáp, hỗ trợ bài báo và dịch văn bản"**.

1.2 Ý nghĩa thực tiễn

1.2.1 Đối với sinh viên và nghiên cứu sinh

Việc tiếp cận và xử lý khối lượng lớn tài liệu học thuật thường gây áp lực rất lớn cho sinh viên, đặc biệt trong các giai đoạn học tập và làm nghiên cứu. Hệ thống mà nhóm đề xuất mang lại những lợi ích thiết thực như:

- **Rút ngắn thời gian đọc và ghi chú**: thay vì phải dành hàng giờ đọc toàn bộ tài liệu, sinh viên có thể nhận ngay bản tóm tắt cô đọng và những luận điểm chính.
- **Tăng khả năng tương tác với tài liệu**: thông qua chức năng hỏi đáp trực tiếp, người học có thể nhanh chóng tìm ra câu trả lời cho những thắc mắc cụ thể, thay vì phải tự tìm kiếm thủ công.

- **Hỗ trợ dịch thuật theo ngữ cảnh chuyên ngành**: nhờ vào mô hình dịch tích hợp, hệ thống không chỉ dịch nghĩa bề mặt mà còn bảo toàn được thuật ngữ và ý nghĩa chuyên sâu trong từng lĩnh vực. Điều này đặc biệt hữu ích với các ngành có ngôn ngữ chuyên ngành phức tạp như y học, công nghệ thông tin hay luật học.

1.2.2 Đối với giảng viên và nhà khoa học

Trong công tác giảng dạy và nghiên cứu, giảng viên và các nhà khoa học thường xuyên phải rà soát, tổng hợp hàng loạt tài liệu. Công cụ này giúp:

- **Rà soát tài liệu nhanh chóng và chính xác**: hỗ trợ quá trình chuẩn bị giáo trình, bài giảng hoặc báo cáo khoa học.
- **Tiết kiệm chi phí dịch thuật**: không cần thuê ngoài hoặc mua gói dịch vụ đắt tiền, đặc biệt khi xử lý khối lượng lớn tài liệu nước ngoài.
- **Nâng cao hiệu quả nghiên cứu**: nhờ việc rút ngắn thời gian tìm hiểu và dịch tài liệu, các nhà nghiên cứu có thể tập trung nhiều hơn vào phân tích, sáng tạo và phát triển ý tưởng mới, từ đó rút ngắn chu kỳ nghiên cứu.

1.2.3 Đối với tổ chức và đơn vị đào tạo

Các tổ chức, viện nghiên cứu hay trường đại học có thể ứng dụng hệ thống này để nâng cao năng lực quản lý tri thức nội bộ:

- **Xây dựng công cụ khai thác tri thức dùng chung**: hệ thống có thể tích hợp vào thư viện số hoặc cơ sở dữ liệu học liệu, tạo ra môi trường học tập thông minh cho toàn bộ giảng viên và sinh viên.
- **Thúc đẩy quá trình chuyển đổi số trong giáo dục và nghiên cứu**: góp phần tạo ra môi trường học tập hiện đại, ứng dụng AI để nâng cao chất lượng đào tạo, từ đó tăng khả năng cạnh tranh của các cơ sở giáo dục trong bối cảnh toàn cầu hóa.

1.3 Các giải pháp hiện có và giới hạn

Trong những năm gần đây, nhiều công cụ đã được phát triển để hỗ trợ quá trình đọc hiểu và dịch thuật tài liệu học thuật. Tuy nhiên, các giải pháp hiện có vẫn còn nhiều hạn chế khi đặt trong bối cảnh nghiên cứu thực tiễn:

- **Công cụ dịch thuật** (Google Translate, DeepL, v.v.): mặc dù dịch nhanh và tiện lợi, nhưng chất lượng dịch thường dừng lại ở mức cơ bản, thiếu ngữ cảnh chuyên ngành, đôi khi gây sai lệch về ý nghĩa, đặc biệt trong các lĩnh vực có thuật ngữ phức tạp.

- **Các mô hình ngôn ngữ lớn trên nền tảng đám mây** (ChatGPT, Gemini, Perplexity, v.v.): có khả năng sinh văn bản và tóm tắt mạnh mẽ, song phụ thuộc hoàn toàn vào dịch vụ cloud. Điều này đặt ra vấn đề về bảo mật dữ liệu nghiên cứu cũng như chi phí sử dụng cho cá nhân và tổ chức.
- **Hệ thống tìm kiếm học thuật** (Google Scholar, ResearchGate): cung cấp kho dữ liệu khổng lồ nhưng chỉ dừng ở mức tìm kiếm tài liệu, không có khả năng hỏi đáp, dịch thuật hay tóm tắt nội dung.

Từ phân tích trên, có thể thấy khoảng trống công nghệ hiện nay là chưa có một hệ thống tích hợp đầy đủ chức năng hỏi đáp, dịch và tóm tắt, có khả năng lưu lịch sử hội thoại, và đặc biệt có thể triển khai tại chỗ (local) để đảm bảo bảo mật và giảm chi phí. Đây chính là điểm mới và là hướng giải quyết mà đề tài của chúng tôi hướng tới.

CHƯƠNG 2: MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU

2.1 Mục tiêu nghiên cứu

2.1.1 Mục tiêu tổng quát

Đề tài hướng tới việc xây dựng một hệ thống hỗ trợ học tập và nghiên cứu toàn diện, có khả năng:

- **Hỏi đáp trực tiếp dựa trên nội dung tài liệu**: cho phép người dùng đặt câu hỏi tự nhiên và nhận được câu trả lời chính xác, trích xuất từ chính bài báo hoặc tài liệu đã cung cấp.
- **Tóm tắt thông minh**: rút gọn nội dung của các tài liệu dài thành những ý chính, giúp người dùng tiết kiệm thời gian đọc và tập trung vào những luận điểm quan trọng.
- **Dịch thuật theo ngữ cảnh chuyên ngành**: đảm bảo tính chính xác về thuật ngữ học thuật và ngữ nghĩa, hỗ trợ người dùng dễ dàng tiếp cận các tài liệu quốc tế.
- **Tương tác liên tục với hệ thống**: lưu giữ lịch sử hội thoại (history) để hỗ trợ các câu hỏi liên quan theo chuỗi, giúp quá trình tìm hiểu tài liệu trở nên mạch lạc và thuận tiện hơn.
- **Vận hành trên môi trường local**: triển khai hệ thống trên máy cá nhân (sử dụng Python virtual environment), đảm bảo tính riêng tư, bảo mật dữ liệu và tiết kiệm chi phí sử dụng dịch vụ cloud.

2.1.2 Mục tiêu cụ thể

Các mục tiêu cụ thể được xây dựng theo tiêu chí **SMART (Specific – Measurable – Achievable – Relevant – Time-bound)**:

- **Tính chính xác**: đạt được độ chính xác trung bình (precision@k) ≥ 85% trong việc truy xuất thông tin từ cơ sở dữ liệu văn bản.
- **Tốc độ xử lý**: thời gian phản hồi cho một truy vấn đơn giản không quá 5 giây trên máy tính cấu hình phổ biến (CPU hoặc GPU tầm trung).
- **Chất lượng dịch và tóm tắt**: bản dịch và bản tóm tắt đạt mức đánh giá ≥ 4/5 theo khảo sát ý kiến người dùng thử nghiệm (sinh viên, giảng viên).
- **Khả năng mở rộng**: hệ thống có thể tích hợp thêm giao diện web/app trong tương lai mà không cần thay đổi nhiều kiến trúc cốt lõi.

- **Mức độ hài lòng của người dùng**: ≥ 80% người tham gia thử nghiệm đánh giá hài lòng về tính hữu ích và tính dễ sử dụng của hệ thống.

2.2 Phạm vi nghiên cứu

2.2.1 Phạm vi chức năng

Hệ thống được xây dựng tập trung vào các chức năng chính:

- Nhập dữ liệu từ các nguồn PDF, file văn bản (.txt, .docx), bài báo trực tuyến (HTML).
- Tiền xử lý văn bản, tách đoạn (chunking) và tạo vector embedding từ dữ liệu.
- Lưu trữ và quản lý embedding trên ChromaDB (vector database).
- Hỗ trợ người dùng thực hiện tìm kiếm, hỏi đáp, dịch thuật, tóm tắt thông qua giao diện tương tác.
- Cung cấp kết quả theo dạng streaming output để tăng trải nghiệm, đồng thời lưu giữ lịch sử hội thoại.

2.2.2 Phạm vi ứng dụng

- Đối tượng chính: sinh viên, nghiên cứu sinh, giảng viên, và các nhà nghiên cứu có nhu cầu xử lý tài liệu học thuật.
- Phạm vi áp dụng: hỗ trợ học tập, nghiên cứu khoa học, giảng dạy, biên soạn giáo trình.
- Ngôn ngữ hỗ trợ: ưu tiên tiếng Anh và tiếng Việt; có thể mở rộng sang các ngôn ngữ khác khi tích hợp thêm mô hình dịch.

2.2.3 Giới hạn nghiên cứu

Trong khuôn khổ đề tài, hệ thống còn một số giới hạn:

- Chưa hỗ trợ xử lý **dữ liệu phi văn bản** như hình ảnh, âm thanh, hoặc video.
- Chưa tích hợp các công cụ **OCR nâng cao** để xử lý tài liệu PDF dạng ảnh (scan).
- Chưa tối ưu hoàn toàn cho môi trường đa người dùng, hệ thống mới tập trung cho quy mô cá nhân hoặc nhóm nhỏ.
- Kết quả dịch và tóm tắt phụ thuộc vào mô hình ngôn ngữ được lựa chọn, có thể cần hiệu chỉnh thêm để phù hợp với từng lĩnh vực chuyên sâu.

CHƯƠNG 3: HƯỚNG GIẢI QUYẾT

3.1. Phương án tiếp cận bài toán

Để giải quyết bài toán hỏi đáp, dịch và tóm tắt dựa trên tài liệu học thuật, nhóm đã cân nhắc ba phương án phổ biến:

- **Search Engine truyền thống**: sử dụng từ khóa để tìm kiếm tài liệu. Ưu điểm là đơn giản, dễ triển khai nhưng độ chính xác thấp, không hỗ trợ hội thoại và không cho phép tóm tắt hay dịch trực tiếp.
- **Hỏi đáp dựa trên Embedding (QA Embedding)**: chuyển đổi tài liệu thành vector và so khớp với truy vấn của người dùng. Phương án này cải thiện độ chính xác nhưng vẫn còn hạn chế trong việc tạo câu trả lời tự nhiên, khó tích hợp thêm chức năng dịch và tóm tắt.
- **Retrieval-Augmented Generation (RAG)**: kết hợp giữa khả năng tìm kiếm ngữ nghĩa (retrieval) và mô hình sinh ngôn ngữ (generation). Đây là phương án cân bằng, vừa đảm bảo độ chính xác, vừa tạo ra câu trả lời tự nhiên, đồng thời dễ dàng mở rộng thêm tính năng dịch và tóm tắt.

Bảng 3.1: So sánh các hướng tiếp cận

Tiêu chí	**Search Engine**	**QA Embedding**	**RAG (đề xuất)**
Cơ chế hoạt động	Tìm kiếm từ khóa, trả về văn bản gốc	So khớp vector giữa truy vấn và tài liệu	Kết hợp tìm kiếm ngữ nghĩa + sinh văn bản
Độ chính xác	Thấp, phụ thuộc từ khóa	Trung bình đến khá	Cao, tận dụng ngữ cảnh và LLM
Khả năng hội thoại	Không	Hạn chế	Có, lưu history và hỗ trợ follow-up query
Khả năng tóm tắt	Không	Hạn chế (phải xây dựng thêm)	Có, dễ tích hợp trực tiếp vào pipeline

Tiêu chí	Search Engine	QA Embedding	RAG (đề xuất)
Khả năng dịch thuật	Không	Chưa tích hợp	Có thể kết hợp cùng LLM để dịch theo ngữ cảnh
Tính mở rộng	Thấp, chỉ dùng cho tìm kiếm cơ bản	Trung bình, khó mở rộng thêm chức năng	Cao, dễ tích hợp dịch, tóm tắt, giao diện web
Ứng dụng phù hợp	Tra cứu nhanh, quy mô nhỏ	Trả lời câu hỏi đơn giản từ dữ liệu có sẵn	Nghiên cứu, học tập, phân tích chuyên sâu

3.2 Kiến trúc tổng thể của hệ thống

Quy trình xử lý dữ liệu của hệ thống được thiết kế theo các bước sau:

```
Tài liệu đầu vào (PDF, web, text)
      ↓
Tiền xử lý (làm sạch, tách đoạn)
      ↓
Embedding (HuggingFace)
      ↓
Lưu trữ vector (ChromaDB)
      ↓
Truy vấn người dùng (query)
      ↓
Tìm kiếm ngữ cảnh liên quan (retrieval)
      ↓
Mô hình ngôn ngữ (LLM) sinh kết quả (Q&A, tóm tắt, dịch)
      ↓
Kết quả trả về (streaming + lưu lịch sử hội thoại)
```

3.3 Các thành phần chính

- **Tiền xử lý dữ liệu (Ingestion & Preprocessing)**
  - Thu thập dữ liệu từ file PDF, web hoặc văn bản.
  - Chuyển đổi và làm sạch dữ liệu (loại bỏ ký tự thừa, chia nhỏ đoạn văn bản).
  - Chia nhỏ tài liệu thành các chunk (200–500 từ) để phù hợp với mô hình ngôn ngữ.
- **Tạo vector ngữ nghĩa (Embedding)**
  - Sử dụng mô hình embedding từ HuggingFace để chuyển đổi các đoạn văn bản thành vector số.
  - Các vector này giúp hệ thống so sánh mức độ tương đồng ngữ nghĩa giữa tài liệu và câu hỏi.
- **Cơ sở dữ liệu vector (ChromaDB)**
  - Lưu trữ toàn bộ embedding cùng thông tin metadata (tên tài liệu, vị trí đoạn).
  - Hỗ trợ tìm kiếm nhanh dựa trên cosine similarity, giúp hệ thống truy xuất những đoạn liên quan nhất đến truy vấn.
- **Truy xuất thông tin (Retrieval)**
  - Khi người dùng đặt câu hỏi, hệ thống chuyển câu hỏi thành vector.
  - So khớp vector truy vấn với vector trong ChromaDB để tìm top-k đoạn văn bản liên quan.
- **Sinh câu trả lời (LLM – Large Language Model)**
  - Các đoạn văn bản được đưa vào mô hình ngôn ngữ (LLM).
  - LLM kết hợp ngữ cảnh để trả lời câu hỏi, dịch văn bản hoặc tạo bản tóm tắt.
- **Tương tác người dùng (Streaming + History)**
  - Kết quả được trả về theo kiểu streaming output để hiển thị nhanh và mượt hơn.
  - Hệ thống lưu lại lịch sử hội thoại để người dùng có thể đặt câu hỏi tiếp theo dựa trên ngữ cảnh cũ

3.4 Định hướng mở rộng

Trong tương lai, hệ thống có thể được phát triển thêm:

- **Hỗ trợ đa ngôn ngữ**: mở rộng sang các ngôn ngữ khác ngoài Anh – Việt.
- **Tích hợp giao diện web/app**: giúp người dùng dễ dàng truy cập và sử dụng.
- **Hỗ trợ dữ liệu đa phương tiện**: ngoài văn bản, có thể bổ sung khả năng phân tích hình ảnh, biểu đồ hoặc file âm thanh.