

PHIẾU ĐĂNG KÝ

PHƯƠNG THỨC ĐÀO TẠO VÀ ĐỀ TÀI LUẬN VĂN THẠC SĨ

1. Tên đề tài (ghi IN HOA):

- Tên tiếng Việt:
XÂY DỰNG HỆ THỐNG ĐỌC HIỂU, TRẢ LỜI VÀ GIẢI THÍCH CÂU TRẢ LỜI DỰA TRÊN OPENAI - ỨNG DỤNG TRONG LĨNH VỰC ĐÀO TẠO.
- Tên tiếng Anh:
AN OPENAI-BASED EXPLAINABLE QUESTION ANSWERING SYSTEM FOR EDUCATION.
- Hướng đề tài luận văn: Hướng ứng dụng (12 TC)

2. Ngành học và Mã ngành: 8480101 (Khoa học máy tính)

3. Cán bộ hướng dẫn:

- Họ tên: TS. Dương Việt Hằng
- Email: hangdv@gm.uit.edu.vn
- Điện thoại: (+84) 919 196 708
- Đơn vị công tác: Trường Đại học Công Nghệ Thông Tin – ĐHQG TP.HCM

4. Thời gian thực hiện: 6 tháng. Từ tháng 11/2025

5. Học viên thực hiện:

- Họ tên: Trần Văn Tịnh
 - Mã số: 220101039
 - Email: tinhtv.17@grad.uit.edu.vn
- Khóa: 17 Đợt: 2
Điện thoại: (+84) 358 444 463

TP. HCM, ngày 13 tháng 01 năm 2026

Xác nhận của CBHD
(Ký tên và ghi rõ họ tên)

TS. Dương Việt Hằng

Học viên
(Ký tên và ghi rõ họ tên)

Trần Văn Tịnh

ĐỀ CƯƠNG ĐỀ TÀI LUẬN VĂN THẠC SĨ

1. Giới thiệu

Đề tài “XÂY DỰNG HỆ THỐNG ĐỌC HIỀU, TRẢ LỜI VÀ GIẢI THÍCH CÂU TRẢ LỜI DỰA TRÊN OPENAI - ÚNG DỤNG TRONG LĨNH VỰC ĐÀO TẠO” nhằm ứng dụng các dịch vụ của OpenAI để xây dựng một hệ thống trả lời câu hỏi cho lĩnh vực giáo dục giúp giải đáp thắc mắc của học sinh một cách nhanh chóng, kịp thời. Hệ thống không chỉ đưa ra câu trả lời mà còn cung cấp giải thích chi tiết từng bước suy luận ra câu trả lời giúp nâng cao độ tin cậy của người dùng vào sự chính xác của hệ thống. **Bên cạnh đó, đề tài cũng hướng đến việc nghiên cứu suy luận kết hợp với các kỹ thuật Symbolic Engine nhằm tăng cường khả năng giải thích hệ thống dựa trên biểu diễn tri thức và suy luận thay vì chỉ dựa đơn thuần trên các mẫu kiến thức tổng quát đã được học của mô hình ngôn ngữ.**

Theo đó, input và output của hệ thống được mô tả như sau:

- Input: gồm hai thành phần:
 - Module ID: giúp xác định phạm vi cụ thể của câu hỏi. Ví dụ: các câu hỏi về quy chế đào tạo, hoặc các câu hỏi chuyên về một môn học cụ thể (như Toán, Lý, Hóa).
 - Một câu hỏi trong phạm vi đó.
- Output: gồm hai thành phần:
 - Câu trả lời trực tiếp cho câu hỏi đã cung cấp.
 - Lời giải thích chi tiết từng bước cho câu trả lời.

2. Bối cảnh, lý do chọn đề tài và khả năng ứng dụng.

Trong bối cảnh chuyển đổi số giáo dục mạnh mẽ của nước ta trong thời kì 4.0, các hệ thống hỗ trợ thông minh đóng một vai trò quan trọng. Phần lớn các hệ thống hiện nay vẫn dựa nhiều vào nguồn lực con người gây bất cập cho học sinh. Học sinh phải chờ đợi lâu để được hỗ trợ, giải đáp thắc mắc bởi giáo viên và những người có chuyên

môn. Việc xây dựng thành công hệ thống không chỉ giúp tối ưu chi phí, nguồn nhân lực cần thiết mà còn thuận tiện cho học sinh khi nhận được giải đáp tức thời.

OpenAI là tổ chức nghiên cứu trí tuệ nhân tạo hàng đầu thế giới, đặc biệt tiên phong trong lĩnh vực xử lý ngôn ngữ tự nhiên. Việc ứng dụng OpenAI để phát triển ứng dụng trả lời câu hỏi có các ưu điểm sau: (1) Các mô hình của họ mạnh mẽ về khả năng hiểu ngôn ngữ và có tiềm năng thực hiện lý luận (reasoning) tốt (như GPT-5, O-series). (2) Các mô hình của họ đa ngôn ngữ phù hợp để phát triển các ứng dụng trên ngôn ngữ ít tài nguyên như tiếng Việt. (3) Nền tảng OpenAI cung cấp cả API và giao diện người dùng tiện lợi cho quá trình phát triển và bảo trì. (4) OpenAI hỗ trợ dễ dàng tích hợp các phần mềm bên thứ 3 như WebSearch, Code Interpreter, File Search & Retrieval, và Công cụ tùy chỉnh. Những ưu điểm đó rất hữu ích cho việc phát triển ứng dụng và điều chỉnh chuyên biệt để tối ưu cho miền kiến thức cụ thể và cơ sở dữ liệu riêng.

Các hệ thống trả lời câu hỏi dựa đơn thuần vào LLM thường gây ra lo ngại về độ tin cậy khi ứng dụng vào giáo dục vì các mô hình này chủ yếu thực hiện suy luận thống kê dựa trên mẫu dữ liệu đã học, chưa phản ánh đầy đủ bản chất và logic của tri thức. Suy luận ký hiệu (Symbolic Reasoning) là một hướng phát triển hứa hẹn dựa trên kiến thức dựa trên biểu diễn tri thức và luật suy diễn, từ đó tạo ra lời giải thích rõ ràng. Tuy nhiên, việc tích hợp Symbolic Engine cũng đặt ra thách thức trong việc đảm bảo sự nhất quán khi ánh xạ từ ngôn ngữ tự nhiên sang ngôn ngữ ký hiệu, cũng như khả năng kiểm soát lỗi và mở rộng suy luận của hệ thống. Với lý do đó, đề tài này cũng tập trung vào nghiên cứu các giải pháp để giải quyết thách thức này và ứng dụng để xây dựng hệ thống suy luận kết hợp LLM và Symbolic Engine.

Nếu phát triển thành công, hệ thống sẽ cho phép tích hợp vào các nền tảng hỗ trợ học trực tuyến (ứng dụng theo môn học cụ thể). Ngoài ra, hệ thống cũng có thể là cơ sở để triển khai các ứng dụng ở các cơ sở đào tạo trong việc giải đáp quy chế tự động giúp tiết kiệm chi phí và nhân lực.

3. Mục tiêu của đề tài.

- Tìm hiểu về các dịch vụ của OpenAI để phát triển khung cơ sở để triển khai các ứng dụng trả lời câu hỏi có kèm theo giải thích chi tiết trong lĩnh vực giáo dục và đào tạo.
- Đề xuất và thử nghiệm các phương pháp, chiến lược cải tiến nâng cao **dựa trên Symbolic Engine** để nâng cao hiệu quả giải thích của hệ thống dựa trên vào bản chất kiến thức.

4. Nội dung nghiên cứu của đề tài.

- Tìm hiểu OpenAI platform, các models hỗ trợ và khả năng ứng dụng của chúng.
- Ứng dụng OpenAI để xây dựng khung cơ sở cho hệ thống trả lời và giải thích câu trả lời cho câu hỏi trong lĩnh vực giáo dục và đào tạo.
- **Đề xuất và thử nghiệm các phương pháp diễn giải câu trả lời kết hợp suy luận ký hiệu (Symbolic Engine) trên 2 bộ dữ liệu XAI Challenge TRNS-AI@IJCNN25 và bộ dữ liệu tự thu thập câu hỏi Hóa Học trung học phổ thông.**

5. Phương pháp thực hiện.

- **Xây dựng bộ dữ liệu:**

Bộ dữ liệu quy chế đào tạo từ XAI Challenge TRNS-AI@IJCNN25

- *Thu thập dữ liệu:* Thu thập từ trang web công khai chính thức của thử thách **XAI Challenge TRNS-AI@IJCNN25**
- *Xử lý và gán nhãn dữ liệu:* tạo danh sách câu hỏi từ bộ dữ liệu thu thập và tiến hành gán nhãn loại câu hỏi phù hợp. Mỗi mẫu dữ liệu gồm danh sách các tiền đề cho trước và 1 nhãn được gán dựa theo logic lập luận như sau:
 - *Yes/No/Uncertain Question:* dựa vào các tiền đề đã cho, xác định xem mệnh đề trong câu hỏi là đúng, sai, hoặc không thể kết luận.
 - *Multiple Choices Question:* dựa vào các tiền đề đã cho, xác định lựa chọn nào là phù hợp với câu hỏi.
 - *Numerical Question:* các câu hỏi liên quan đến số liệu và tính toán.

Bộ dữ liệu câu hỏi Hóa Học trung học phổ thông

- *Thu thập dữ liệu*: Thu thập danh sách câu hỏi từ các nguồn sách bài tập hóa học 10, 11, 12 gồm 3 bộ sách (Kết Nối Tri Thức Với Cuộc Sống, Chân Trời Sáng Tạo, và Cánh Diều). Sau đó sử dụng công cụ Regular Expression, Mathpix,... để trích xuất câu hỏi, câu trả lời, công thức hóa học từ PDF để tạo dữ liệu thô.
- *Xử lý dữ liệu*: Tạo danh sách câu hỏi từ dữ liệu thô và gán nhãn phù hợp. Mỗi câu hỏi gồm 2 nhãn:
 - Theo loại câu hỏi: bao gồm 2 loại trắc nghiệm và tự luận.
 - Theo chủ đề câu hỏi: được phân theo 3 nhóm câu hỏi:
 - + *Lý thuyết*: các liên quan đến các khái niệm hóa học như chất, cấu tạo nguyên tử, và các khái niệm cơ bản khác.
 - + *Chất – Hợp chất*: các câu hỏi liên quan đến các đặc điểm và ứng dụng của một chất – hợp chất. Gồm các chủ đề nhỏ: phân tích, nhận biết, vận dụng.
 - + *Phản ứng – chuỗi phản ứng hóa học*: các câu hỏi xoay quanh phản ứng hóa học – chuỗi phản ứng hóa học. Gồm các chủ đề nhỏ hơn: phân tích, vận dụng, tính toán.

Bộ dataset sẽ được dùng để thử nghiệm và đánh giá ở các bước sau đó.

- **Xây dựng khung cơ sở cho hệ thống trả lời và giải thích câu trả lời cho câu hỏi giáo dục:**
 - *Xây dựng khung OpenAI cơ sở*: khung cơ sở được thiết kế gồm 3 giai đoạn (dựa trên kết quả nền tảng của học viên từ thử thách **XAI Challenge TRNS-AI@IJCNN25**): Classification Agent → Reasoning and Solver Agent → Synthesizes and Generation Agent.

Trong đó:

- ***Classification Agent***: chịu trách nhiệm xác định câu hỏi. Bao gồm xác định loại câu hỏi và chủ đề câu hỏi như trong bộ dữ liệu.
- ***Reasoning and Solver Agent***: chịu trách nhiệm phân tích câu hỏi, truy xuất kiến thức, thực hiện chi tiết suy luận (bao gồm LLM và Symbolic engine) và giải quyết câu hỏi.

- ***Synthesizes and Generation Agent:*** thực hiện tổng hợp và tạo câu trả lời theo cấu trúc thống nhất.
- *Tối ưu hệ thống*: nghiên cứu các mẫu, chiến lược prompting, truy xuất hiện đại để tối ưu kết quả. Thực hiện đánh giá cơ bản các mô hình OpenAI khác nhau (gpt-5, gpt-4, O-series) để lựa chọn models phù hợp và tối ưu.
- **Nghiên cứu, thử nghiệm và tích hợp diễn giải câu trả lời dựa trên suy luận ký hiệu (Symbolic Engine).**

Bộ dữ liệu quy chế đào tạo từ XAI Challenge TRNS-AI@IJCNN25

- Nghiên cứu và thử nghiệm các phương pháp dựa trên truy xuất tăng cường để đảm bảo mô hình có được sự nhất quán về khái niệm khi ánh xạ từ ngôn ngữ tự nhiên sang ngôn ngữ ký hiệu.
- Nghiên cứu tích hợp công cụ suy luận ký hiệu (như Z3 Solver) để tăng hiệu quả giải thích theo 2 hướng hoặc dùng làm công cụ suy diễn chính hoặc dùng làm công cụ xác thực các bước suy luận riêng lẻ của LLM.
- Đánh giá và so sánh kết quả với phương pháp dựa đơn thuần trên LLM.

Bộ dữ liệu câu hỏi Hóa Học trung học phổ thông

- Tìm hiểu các cơ sở dữ liệu hóa học PubChem, ChemSpider và cách truy xuất tri thức hóa học từ các cơ sở dữ liệu đó.
- Thử nghiệm các phương pháp phân tích câu hỏi (nhằm chọn lọc các khái niệm, chất hóa học cần thiết) và truy xuất cơ sở tri thức để có kiến thức cơ sở cho việc suy luận.
- Đánh giá các phương pháp ánh xạ từ ngôn ngữ tự nhiên sang ngôn ngữ ký hiệu và tích hợp công cụ suy luận ký hiệu tương tự như cách trên bộ dữ liệu quy chế đào tạo. Phân tích và đề xuất điều chỉnh phù hợp trên bộ dữ liệu câu hỏi Hóa Học trung học phổ thông.
- Đánh giá và so sánh kết quả với phương pháp dựa đơn thuần trên LLM.

6. Kết quả, sản phẩm dự kiến

- Bộ dataset trả lời câu hỏi môn Hóa học trung học phổ thông ngôn ngữ tiếng Việt.

- Một khung OpenAI (framework) cơ sở để triển khai các ứng dụng trả lời câu hỏi cho giáo dục và đào tạo.
- Ứng dụng trả lời câu hỏi Hóa học trung học phổ thông, một triển khai cụ thể dựa trên khung cơ sở để trả lời câu hỏi Hóa học trung học phổ thông ngôn ngữ tiếng Việt.

7. Tài liệu tham khảo.

[1] OpenAI, “OpenAI Platform,” *Openai.com*, 2025.

<https://platform.openai.com/docs/overview>

[2] “OpenAI Cookbook,” *Openai.com*, 2025. <https://cookbook.openai.com/>.

[3] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. v, & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

[4] “Introduction - PubChemPy 1.0.5 documentation,” *Pubchempy.org*, 2025.

<https://docs.pubchempy.org/en/latest/guide/introduction.html>.

[5] “ChemSpiPy — ChemSpiPy 2.0.0 documentation,” *Readthedocs.io*, 2018.

<https://chemspipy.readthedocs.io/en/latest/index.html>.

8. Kế hoạch

Công việc	Thời gian Dự kiến (Ngày)	Ghi Chú
1. Xây dựng bộ dữ liệu.	23	<ul style="list-style-type: none"> - Thu thập từ trang web công khai của thử thách XAI Challenge TRNS-AI@IJCNN25 và gán nhãn dữ liệu. (3 ngày) - Thu thập dữ liệu câu hỏi hóa học trung học phổ thông (Hóa 10-12) từ 3 bộ sách (Kết Nối Tri Thức Với Cuộc Sống, Chân Trời Sáng Tạo, Cánh Diều) theo chương trình giáo dục mới. (20 ngày)

2. Tìm hiểu và Xây dựng khung cơ sở OpenAI cho hệ thống trả lời và giải thích câu trả lời cho câu hỏi giáo dục.	20	<ul style="list-style-type: none"> - Xây dựng khung cơ sở OpenAI 3 giai đoạn. (10 ngày) - Thực hiện các tối ưu lời nhắc và lựa chọn mô hình (10 ngày).
3. Nghiên cứu, thử nghiệm và tích hợp diễn giải câu trả lời dựa trên suy luận ký hiệu (Symbolic Engine).	60	<ul style="list-style-type: none"> - Phương pháp suy luận kết hợp symbolic engine trên bộ dữ liệu XAI Challenge TRNS-AI@IJCNN25. (25 ngày) - Phương pháp suy luận kết hợp Symbolic Engine trên bộ dữ liệu câu hỏi Hóa Học trung học phổ thông. (35 ngày)
4. Phân tích kết quả và viết luận văn.	27	<ul style="list-style-type: none"> - Thu thập tài liệu cần thiết cho việc viết luận văn (5 ngày). - Viết bản thảo luận văn (15 ngày) - Hoàn thiện và sửa đổi luận văn dựa trên phản hồi của giảng viên hướng dẫn (5 ngày) - Hoàn thiện và nộp bản thảo luận văn (2 ngày)
5. Làm Slides báo cáo	9	<ul style="list-style-type: none"> - Làm slides báo cáo và nhận phản hồi từ giảng viên hướng dẫn. (5 ngày) - Hoàn thiện slides báo cáo dựa trên phản hồi từ giảng viên hướng dẫn (4 ngày)
Tổng cộng	139	

Biểu đồ Grant

GIẢI TRÌNH CHỈNH SỬA

Nhận xét của ĐVCM	Vị trí chỉnh sửa	Nội dung chỉnh sửa (được in đậm)
Học viên cần làm rõ sự khác biệt của giải pháp dự kiến với cách phương thức, hệ thống hỏi-đáp hiện có như ChatGPT (OpenAI), Notebooklm... trong đó, đã cho phép người dùng hỏi-đáp và được giải thích trên kiến thức thức tổng quát, sử dụng dữ liệu thu thập từ web, hoặc kho tài liệu do người dùng cung cấp. Bài toán ứng dụng mà học viên dự kiến có sự khác biệt gì ?	Bổ sung, Mục 1 – Giới thiệu đề tài.	Đề tài “ XÂY DỰNG HỆ THỐNG ĐỌC HIẾU, TRẢ LỜI VÀ GIẢI THÍCH CÂU TRẢ LỜI DỰA TRÊN OPENAI - ÚNG DỤNG TRONG LĨNH VỰC ĐÀO TẠO ” nhằm ứng dụng các dịch vụ của OpenAI để xây dựng một hệ thống trả lời câu hỏi cho lĩnh vực giáo dục giúp giải đáp thắc mắc của học sinh một cách nhanh chóng, kịp thời. Hệ thống không chỉ đưa ra câu trả lời mà còn cung cấp giải thích chi tiết từng bước suy luận ra câu trả lời giúp nâng cao độ tin cậy của người dùng vào sự chính xác của hệ thống. Bên cạnh đó, đề tài cũng hướng đến việc nghiên cứu suy luận kết hợp với các kỹ thuật Symbolic Engine nhằm tăng cường khả năng giải thích hệ thống dựa trên biểu diễn tri thức và suy luận thay vì chỉ dựa đơn thuần trên các mẫu kiến thức tổng quát đã được học của mô hình ngôn ngữ.
Bài toán ứng dụng mà học viên dự kiến có sự khác biệt gì ? có thách thức gì cần phải nghiên cứu giải quyết, mà không thể chỉ đơn thuần sử dụng các giải pháp hiện tại, đặc biệt trong domain quy chế đào tạo, hay Hóa học. Trên cơ sở đó, cho thấy ý nghĩa, đóng góp về khoa	Bổ sung, Mục 2 – lý do chọn đề tài.	OpenAI là tổ chức nghiên cứu trí tuệ nhân tạo hàng đầu thế giới, đặc biệt tiên phong trong lĩnh vực xử lý ngôn ngữ tự nhiên. Việc ứng dụng OpenAI để phát triển ứng dụng trả lời câu hỏi có các ưu điểm sau: (1) Các mô hình của họ mạnh mẽ về khả năng hiểu ngôn ngữ và có tiềm năng thực hiện lý luận (reasoning) tốt (như GPT-5, O-series). (2) Các mô hình của họ đa ngôn ngữ phù hợp để phát triển các ứng dụng trên ngôn ngữ ít tài nguyên như tiếng Việt. (3) Nền tảng OpenAI cung cấp cả API và giao diện người dùng tiện lợi

học hoặc ứng dụng của đề tài này.		<p>cho quá trình phát triển và bảo trì. (4) OpenAI hỗ trợ dễ dàng tích hợp các phần mềm bên thứ 3 như WebSearch, Code Interpreter, File Search & Retrieval, và Công cụ tùy chỉnh. Những ưu điểm đó rất hữu ích cho việc phát triển ứng dụng và điều chỉnh chuyên biệt để tối ưu cho miền kiến thức cụ thể và cơ sở dữ liệu riêng.</p> <p>Các hệ thống trả lời câu hỏi dựa đơn thuần vào LLM thường gây ra lo ngại về độ tin cậy khi ứng dụng vào giáo dục vì các mô hình này chủ yếu thực hiện suy luận thống kê dựa trên mẫu dữ liệu đã học, chưa phản ánh đầy đủ bản chất và logic của tri thức. Suy luận ký hiệu (Symbolic Reasoning) là một hướng phát triển hứa hẹn dựa trên kiến thức dựa trên biểu diễn tri thức và luật suy diễn, từ đó tạo ra lời giải thích rõ ràng. Tuy nhiên, việc tích hợp Symbolic Engine cũng đặt ra thách thức trong việc đảm bảo sự nhất quán khi ánh xạ từ ngôn ngữ tự nhiên sang ngôn ngữ ký hiệu, cũng như khả năng kiểm soát lỗi và mở rộng suy luận của hệ thống. Với lý do đó, đề tài này cũng tập trung vào nghiên cứu các giải pháp để giải quyết thách thức này và ứng dụng để xây dựng hệ thống suy luận kết hợp LLM và Symbolic Engine.</p>
Hiệu chỉnh mô tả "Kết quả, sản phẩm dự kiến". Cụ thể, "hệ thống OpenAI nền tảng" là hệ thống hoàn chỉnh, hay chỉ là	Chỉnh sửa, Mục 5– Kết quả, sản phẩm dự kiến.	<ul style="list-style-type: none"> Bộ dataset trả lời câu hỏi môn Hóa học trung học phổ thông ngôn ngữ tiếng Việt. Một khung OpenAI (framework) cơ sở để triển khai các ứng dụng trả lời câu hỏi cho giáo dục và đào tạo.

<p>framework/pipeline/work flow dựa trên dịch vụ do OpenAI cung cấp qua API ? Module trả lời câu hỏi giáo dục môn Hóa học là một module mới ? module này là một phần của hệ thống nào ?</p>		<ul style="list-style-type: none"> Ứng dụng trả lời câu hỏi Hóa học trung học phổ thông, một triển khai cụ thể dựa trên khung cơ sở để trả lời câu hỏi Hóa học trung học phổ thông ngôn ngữ tiếng Việt.
---	--	--