TIỀN XỬ LÝ DỮ LIỆU VÀ XÂY DỰNG CƠ SỞ TRI THỨC CHO HỆ THỐNG HỎI ĐÁP DÙNG LLM: ỨNG DỤNG TRONG PHÂN TÍCH DẤU CHÂN CARBON CÁ NHÂN

[LINK VIDEO MÔ TẢ THỰC HÀNH DỰ ÁN:

https://www.tiktok.com/@8801011999a/video/7581507361462930708?is_from_webapp=1&sender_device=pc&web_id=7572178763338090005]

Nguyễn Quang Minh[1], Đỗ Huyền Châm[2], Vũ Tuấn Anh[3], Nguyễn Tấn Dũng[4], Vũ Văn Thanh[5].

[1]Khoa/Phòng Công Nghệ Thông Tin, Trường/Viện Cao Đẳng FPT Polytechncic

[2]Khoa/Phòng Công Nghệ Thông Tin, Trường/Viện Cao Đẳng FPT Polytechncic

[3]Khoa/Phòng Công Nghệ Thông Tin, Trường/Viện Cao Đẳng FPT Polytechncic

[4]Khoa/Phòng Công Nghệ Thông Tin, Trường/Viện Cao Đẳng FPT Polytechncic

[5]Khoa/Phòng Công Nghệ Thông Tin, Trường/Viện Cao Đẳng FPT Polytechncic

*Email: thanhlaptrinhfpt@gmail.com

TÓM TẮT

Bài viết đề xuất một quy trình tiền xử lý dữ liệu và xây dựng cơ sở tri thức nhằm hỗ trợ hệ thống hỏi đáp dựa trên mô hình ngôn ngữ lớn (LLM), với trường hợp ứng dụng trong phân tích dấu chân carbon cá nhân. Nghiên cứu tập trung xây dựng một pipeline gồm năm giai đoạn: thu thập dữ liệu hành vi liên quan đến phát thải, chuẩn hóa đơn vị và định dạng để đảm bảo tính nhất quán, tổng hợp hệ số CO_2e từ các nguồn khoa học uy tín, mô hình hóa tri thức dưới dạng cấu trúc (schema) dễ truy xuất, và thiết kế lớp ngữ cảnh – prompt giúp LLM suy luận chính xác hơn. Đóng góp chính của đề tài nằm ở việc hình thành một khung phương pháp (framework) có thể tái sử dụng, giúp các hệ thống AI tận dụng dữ liệu bền vững và hạn chế sai lệch trong câu trả lời. Mặc dù không triển khai mô hình AI thực nghiệm, bài viết cung cấp cách tiếp cận mang tính ứng dụng cao, phù hợp cho các nền tảng hỏi đáp dựa trên tri thức như NotebookLM hoặc các hệ thống LLM hiện đại khác.

Từ khóa: Tiền xử lý dữ liệu, Cơ sở tri thức, Mô hình ngôn ngữ lớn – LLM, Dấu chân carbon cá nhân, Thiết kế ngữ cảnh và prompt

GIỚI THIỆU

Trong bối cảnh các hệ thống hỗ trợ ra quyết định ngày càng dựa vào dữ liệu, việc hiểu và đánh giá dấu chân carbon cá nhân đang trở thành một nhu cầu thiết thực đối với cả cá nhân lẫn cộng đồng. Các mô hình ngôn ngữ lớn (LLM) mang đến khả năng phân tích thông tin linh hoạt, diễn giải tự nhiên và hỗ trợ tương tác hỏi đáp theo ngữ cảnh. Tuy nhiên, LLM chỉ thực sự phát huy hiệu quả khi được cung cấp một nguồn tri thức đáng tin cậy và dữ liệu đầu vào đã được chuẩn hóa. Nếu thiếu bước tiền xử lý và xây dựng tri thức, LLM dễ tạo ra thông tin sai lệch, không nhất quán hoặc thiếu căn cứ khoa học.

Đề tài này tập trung nghiên cứu quy trình **tiền xử lý dữ liệu** và **xây dựng cơ sở tri thức** phục vụ cho hệ thống hỏi đáp về dấu chân carbon cá nhân. Thay vì phát triển một ứng dụng hoàn chỉnh, nghiên cứu hướng đến việc đề xuất một quy trình thực tiễn, có khả năng áp dụng ngay vào các hệ thống LLM hiện đại như NotebookLM, ChatGPT hoặc Gemini. Dữ liệu được thu thập từ các nguồn khoa học như GHG Protocol, IPCC và các tài liệu môi trường chính thống, sau đó được chuẩn hóa và mô hình hóa để đảm bảo tính nhất quán khi LLM truy xuất.

Ngoài ra, nghiên cứu còn đề xuất phương pháp thiết kế ngữ cảnh (context) và prompt để tối ưu hóa khả năng suy luận của mô hình, giúp LLM trả lời chính xác hơn dựa trên tri thức đã nạp thay vì suy luận cảm tính. Đây là bước quan trọng nhằm đảm bảo tính tin cậy và giảm thiểu hiện tượng "hallucination" thường gặp trong các hệ thống AI.

Kết quả của nghiên cứu không phải là một hệ thống hoàn chỉnh mà là một **khung phương pháp (framework)** giúp sinh viên, nhà phát triển hoặc người quan tâm có thể triển khai một hệ thống hỏi đáp về dấu chân carbon dựa trên LLM một cách đúng đắn, nhất quán và có căn cứ khoa học.

MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU

Mục tiêu của đề tài

Đề tài hướng đến việc xây dựng một quy trình ứng dụng có tính khả thi, giúp chuyển đổi dữ liệu thô liên quan đến dấu chân carbon cá nhân thành một cơ sở tri thức rõ ràng, nhất quán và có thể sử dụng trực tiếp trong các hệ thống hỏi đáp dựa trên mô hình ngôn ngữ lớn (LLM). Cụ thể, đề tài đặt ra các mục tiêu sau:

1. **Đề xuất quy trình thu thập dữ liệu về dấu chân carbon cá nhân** từ các nguồn khoa học và các công trình nghiên cứu môi trường có độ tin cậy cao.
2. **Xây dựng bộ quy tắc tiền xử lý dữ liệu**, bao gồm làm sạch, chuẩn hóa đơn vị đo lường và chuyển đổi dữ liệu sang cấu trúc phù hợp cho mô hình tri thức.
3. **Phát triển một cơ sở tri thức nền (Knowledge Base)** bao gồm hệ số phát thải CO_2e, quy tắc tính toán và các mô tả hoạt động thường gặp trong đời sống cá nhân.
4. **Thiết kế phương pháp xây dựng ngữ cảnh và kỹ thuật prompt engineering** nhằm tối ưu hóa khả năng truy xuất và suy luận của LLM dựa trên tri thức đã cung cấp.
5. **Mô phỏng quy trình hỏi đáp bằng NotebookLM hoặc các LLM tương đương** nhằm kiểm tra khả năng diễn giải thông tin sau khi tri thức được chuẩn hóa.
6. **Đề xuất hướng ứng dụng thực tế**, giúp người dùng hiểu rõ lượng phát thải carbon của bản thân và đưa ra các quyết định giảm thiểu tác động môi trường.

Mục tiêu của nghiên cứu nằm ở việc **chuẩn hóa và tối ưu dữ liệu cho AI**, thay vì xây dựng một hệ thống phần mềm hoàn chỉnh.

Đối tượng hướng đến

Đề tài được xây dựng hướng đến các nhóm đối tượng sau:

- **Sinh viên và người học ngành CNTT, IoT hoặc môi trường** muốn tiếp cận cách ứng dụng LLM trong xử lý dữ liệu thực tế mà không cần triển khai mô hình phức tạp.

- **Nhà phát triển quan tâm đến chatbot tri thức (knowledge-grounded chatbot)**, cần một quy trình chuẩn để xây dựng và nạp dữ liệu vào LLM.
- **Các cá nhân hoặc tổ chức đang nghiên cứu về dấu chân carbon** và mong muốn có công cụ phân tích thân thiện, dễ truy cập.
- **Người dùng phổ thông** muốn hiểu rõ tác động môi trường từ hoạt động hằng ngày thông qua hệ thống hỏi đáp tự nhiên.

Nhóm đối tượng hướng đến trải rộng từ học thuật đến ứng dụng thực tiễn, giúp đề tài có giá trị cả về giáo dục lẫn triển khai thực tế.

Công cụ AI sử dụng

CÔNG CỤ AI	VAI TRÒ	MỤC ĐÍCH SỬ DỤNG
NotebookLM	Mô phỏng hệ thống hỏi đáp dựa trên tri thức.	Nạp tài liệu, kiểm tra khả năng truy xuất tri thức, thử nghiệm hỏi đáp và đánh giá chất lượng ngữ cảnh/prompt.
GPT/Gemini	Công cụ hỗ trợ xử lý ngôn ngữ và tối ưu hóa tương tác.	Chuẩn hóa dữ liệu mô tả, thiết kế prompt, tổng hợp nội dung và kiểm tra tính nhất quán của tri thức.
Google sheet/ Microsoft Excel	Công cụ xử lý, làm sạch và chuẩn hóa dữ liệu.	Tạo bảng dữ liệu, chuyển đổi đơn vị, xây dựng bảng hệ số phát thải và chuẩn hóa định dạng tri thức.
Công cụ phân tích văn bản của AI	Hỗ trợ thu thập và xác minh dữ liệu từ tài liệu khoa học.	Trích xuất hệ số phát thải, tóm tắt tài liệu, đối chiếu nguồn và kiểm tra độ tin cậy của dữ liệu.

Kịch bản sản phẩm và kế hoạch triển khai

Kịch bản sản phẩm

Hệ thống được mô phỏng dưới dạng một công cụ hỏi đáp về dấu chân carbon cá nhân sử dụng mô hình ngôn ngữ lớn (LLM). Người dùng nhập vào các thông tin như quãng đường di chuyển, mức tiêu thụ điện, khẩu phần ăn hay thói quen mua sắm. Hệ thống dựa trên cơ sở tri thức đã được chuẩn hóa để:

1. Xác định loại hoạt động và đơn vị đo phù hợp.
2. Tra cứu hệ số phát thải CO_2e tương ứng.
3. Tính toán lượng phát thải cho từng hoạt động.
4. LLM diễn giải kết quả bằng ngôn ngữ tự nhiên, cung cấp khuyến nghị giảm thiểu phù hợp.

Sản phẩm cuối cùng mang tính mô phỏng, giúp thể hiện quy trình xử lý dữ liệu và cách LLM hoạt động dựa trên tri thức thay vì tạo ra một ứng dụng hoàn chỉnh.

Kế hoạch triển khai

Quy trình được thiết kế theo 5 bước chính, nhằm bảo đảm tính rõ ràng, dễ áp dụng và phù hợp với mục tiêu xây dựng tri thức cho LLM.

Bước 1 – Thu thập dữ liệu

- Tìm nguồn hệ số phát thải từ IPCC, GHG Protocol, tài liệu khoa học và số liệu từ ngành môi trường.
- Thu thập các mẫu dữ liệu hoạt động cá nhân (km di chuyển, kWh điện năng, đồ ăn, sản phẩm tiêu dùng).
- Xây dựng bộ dữ liệu mô phỏng phục vụ cho quá trình thử nghiệm.

Bước 2 – Tiền xử lý và chuẩn hóa

- Làm sạch dữ liệu: loại bỏ lỗi nhập liệu, giá trị thiếu, dữ liệu không hợp lệ.
- Chuẩn hóa đơn vị đo: chuyển đổi về dạng CO_2e thống nhất.
- Mã hóa dữ liệu theo danh mục (phương tiện, loại điện, nhóm thực phẩm…).
- Thiết kế cấu trúc bảng dữ liệu chuẩn cho Knowledge Base.

Bước 3 – Xây dựng cơ sở tri thức

- Tạo bảng hệ số phát thải với các trường: Hoạt động – Đơn vị – Hệ số CO_2e – Nguồn trích dẫn.
- Mô tả quy tắc tính toán cho từng nhóm hoạt động.
- Bổ sung ví dụ minh họa để hỗ trợ khả năng suy luận của LLM.
- Đóng gói tri thức theo định dạng phù hợp để nạp vào NotebookLM hoặc GPT.

Bước 4 – Thiết lập ngữ cảnh và prompt engineering

- Tạo ngữ cảnh (context) bắt buộc LLM chỉ trả lời dựa trên tri thức đã nạp.
- Thiết kế bộ prompt chuẩn để hướng dẫn LLM quy trình tính toán và trình bày kết quả.
- Kiểm tra độ ổn định của câu trả lời thông qua nhiều tình huống mô phỏng.

Bước 5 – Mô phỏng hoạt động hệ thống

- Nạp dữ liệu vào NotebookLM.
- Đặt các câu hỏi thực tế như: “Tôi đi xe máy 20 km mỗi ngày thì thải ra bao nhiêu CO_2e?”
- Quan sát cách hệ thống truy xuất tri thức và diễn giải kết quả.
- Đánh giá độ tin cậy, nhất quán và khả năng ứng dụng thực tiễn.

Giá trị của kế hoạch triển khai

Kế hoạch triển khai giúp định hình một quy trình hoàn chỉnh, phù hợp để:

- Huấn luyện sinh viên hiểu cách chuẩn hóa dữ liệu trước khi đưa vào AI.
- Tạo nền tảng cho chatbot tri thức trong tương lai.
- Hỗ trợ người dùng đánh giá dấu chân carbon cá nhân một cách minh bạch và khoa học.
- Hạn chế sai lệch thông tin của LLM nhờ tri thức chuẩn hóa.

DỰ KIẾN KẾT QUẢ ĐẦU RA

Đề tài tập trung vào việc xây dựng tri thức và mô phỏng khả năng hỏi đáp của LLM, do đó kết quả đầu ra chủ yếu nằm ở mức dữ liệu, mô hình hóa và quy trình triển khai. Các kết quả dự kiến bao gồm:

Bộ dữ liệu đã được chuẩn hóa

- Bộ dữ liệu mô phỏng về hoạt động cá nhân (di chuyển, tiêu thụ điện, ăn uống, mua sắm).
- Dữ liệu đã được xử lý lỗi, thống nhất đơn vị đo và sắp xếp theo cấu trúc dễ truy xuất.
- Bảng chuyển đổi đơn vị → CO_2e ở định dạng có thể nạp vào LLM (CSV, Sheets, bảng tri thức).

Bảng hệ số phát thải CO_2e (Emission Factors)

- Bảng hệ số phát thải được trích xuất từ các nguồn khoa học (GHG Protocol, IPCC…).
- Bao gồm các trường: *Hoạt động – Đơn vị – Hệ số phát thải – Ghi chú – Nguồn trích dẫn*.
- Đây là nền tảng trung tâm để LLM tính toán và trả lời chính xác.

Cơ sở tri thức (Knowledge Base) cho LLM

- Tập hợp dữ liệu đã chuẩn hóa + mô tả + quy tắc tính toán.
- Có thể đưa trực tiếp vào NotebookLM hoặc các LLM tương đương.
- Tri thức này giúp mô hình trả lời dựa trên dữ liệu có thật, không suy diễn cảm tính.

Bộ hệ thống prompt và ngữ cảnh

- Tập hợp các prompt tiêu chuẩn dùng để truy vấn dữ liệu, yêu cầu giải thích, hoặc yêu cầu mô hình tính toán lượng CO_2e.
- Tài liệu mô tả cấu trúc ngữ cảnh (context template) để kiểm soát hành vi của LLM và hạn chế việc tạo thông tin sai.
- Các prompt được thử nghiệm để đảm bảo tính ổn định và nhất quán.

Mô phỏng tương tác hỏi đáp

- Các trường hợp hỏi đáp mẫu như:
    - "Tôi đi xe máy 30 km mỗi ngày, một tuần thải ra bao nhiêu CO_2e?"
    - "Giữa gà và thịt bò, loại nào gây ra phát thải carbon cao hơn?"
    - "Làm sao giảm phát thải khi sử dụng điều hòa hằng ngày?"
- Kết quả mô phỏng thể hiện cách LLM sử dụng tri thức đã nạp để trả lời rõ ràng và có căn cứ khoa học.

Bộ quy trình triển khai hoàn chỉnh

- Pipeline gồm 5 bước: thu thập → tiền xử lý → xây dựng tri thức → thiết kế prompt → mô phỏng hỏi đáp.
- Quy trình được mô tả rõ ràng, có thể áp dụng cho các đề tài tương tự như giáo dục, y tế, phân tích hành vi hoặc IoT.

Ý nghĩa ứng dụng

Kết quả đầu ra giúp sinh viên và người phát triển có nền tảng rõ ràng để:

- hiểu cách xử lý dữ liệu trước khi tích hợp vào AI,
- triển khai chatbot tri thức trong tương lai,
- xây dựng hệ thống phân tích carbon minh bạch và dễ tiếp cận cho người dùng.