TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
Khoa Công Nghệ Thông Tin
Bộ môn: Khoa Học Máy Tính

BẢO VỆ KHÓA LUẬN TỐT NGHIỆP

XÂY DỰNG BỘ DỮ LIỆU PHÁT HIỆN ẢO GIÁC CỦA MÔ HÌNH NGÔN NGỮ LỚN TRONG NGỮ CẢNH DỊCH VỤ CÔNG

Sinh viên thực hiện:
21120108 – Nguyễn Tiến Nhật
21120201 – Bùi Đình Bảo

Giáo viên hướng dẫn:
TS. Nguyễn Tiến Huy
ThS. Nguyễn Trần Duy Minh

Nội dung

1. Giới thiệu
2. Các công trình liên quan
3. Phương pháp đề xuất
4. Kết quả thí nghiệm
5. Kết luận

1. Giới thiệu

1.1 Lý do chọn đề tài

- LLMs tiềm năng hỗ trợ dịch vụ công (tra cứu, hướng dẫn thủ tục).
- Hallucination gây sai lệch thông tin, dễ hiểu lầm pháp lý trong khi dịch vụ công đòi hỏi tính chính xác và minh bạch.
- Chưa có bộ dữ liệu chuyên biệt cho bài toán phát hiện ảo giác trong dịch vụ công tại Việt Nam.

Mục tiêu

Phát triển phương pháp đánh giá độ ảo giác của LLM trong ngữ cảnh dịch vụ công bằng cách xây dựng bộ dữ liệu tiếng Việt.

Câu hỏi thường gặp: Người tố cáo có được rút đơn tố cáo không?
Câu trả lời đúng: Người tố cáo có quyền rút toàn bộ nội dung tố cáo hoặc một phần nội dung tố cáo trước khi người giải quyết tố cáo ra kết luận nội dung tố cáo. Việc rút tố cáo phải được thực hiện bằng văn bản (Khoản 1 Điều 33 Luật tố cáo 2018).
Câu trả lời ảo giác: Người tố cáo không có quyền rút đơn tố cáo một khi đã nộp đơn, bất kể là toàn bộ hay một phần nội dung tố cáo. Việc này phải được thực hiện bằng hình thức gọi điện thoại và không cần văn bản xác nhận (Khoản 1 Điều 29 Luật tố cáo 2018).

Bảng 1.1: Một ví dụ về câu hỏi thường gặp, câu trả lời đúng, câu trả lời ảo giác của mô hình ngôn ngữ lớn trong bộ dữ liệu

Cho một câu hỏi liên quan đến dịch vụ công và một câu trả lời từ mô hình LLM, hãy xác định xem câu trả lời này có chứa thông tin ảo giác hay không.

Để làm được điều này cần có:

- Bộ câu hỏi – câu trả lời đúng.
- Bộ câu trả lời chứa ảo giác (được sinh ra có kiểm soát).
- Hệ thống đánh giá mô hình dựa trên đầu vào và đầu ra.

2. Các công trình liên quan

2.1 HaluEval:

- Là tập dữ liệu quy mô lớn đánh giá khả năng **phát hiện hallucination** trong phản hồi của LLMs
- **Quy trình gồm 2 bước**: Lấy mẫu phản hồi từ nhiều chỉ dẫn và lọc ra những phản hồi có khó nhận diện hallucination nhất**.**
- Kết hợp **chú thích thủ công** để xác định thông tin sai lệch hoặc không kiểm chứng.
- Một số chiến lược như **bổ sung tri thức ngoài** hoặc **suy luận theo bước (reasoning)** có thể giúp mô hình phát hiện ảo giác hiệu quả hơn.

2.2 SelfCheckGPT-WikiBio:

- Là đề xuất đầu tiên phát hiện hallucination của LLM mà không cần dữ liệu tham chiếu.
- Đánh giá ảo giác dựa trên so sánh nhiều phản hồi được sinh ra từ cùng một đầu vào.
- Hiệu quả cao với cả mô hình hộp đen và hộp xám.
- Công bố thêm bộ dữ liệu gán nhãn thủ công từ GPT-3 để phục vụ nghiên cứu.

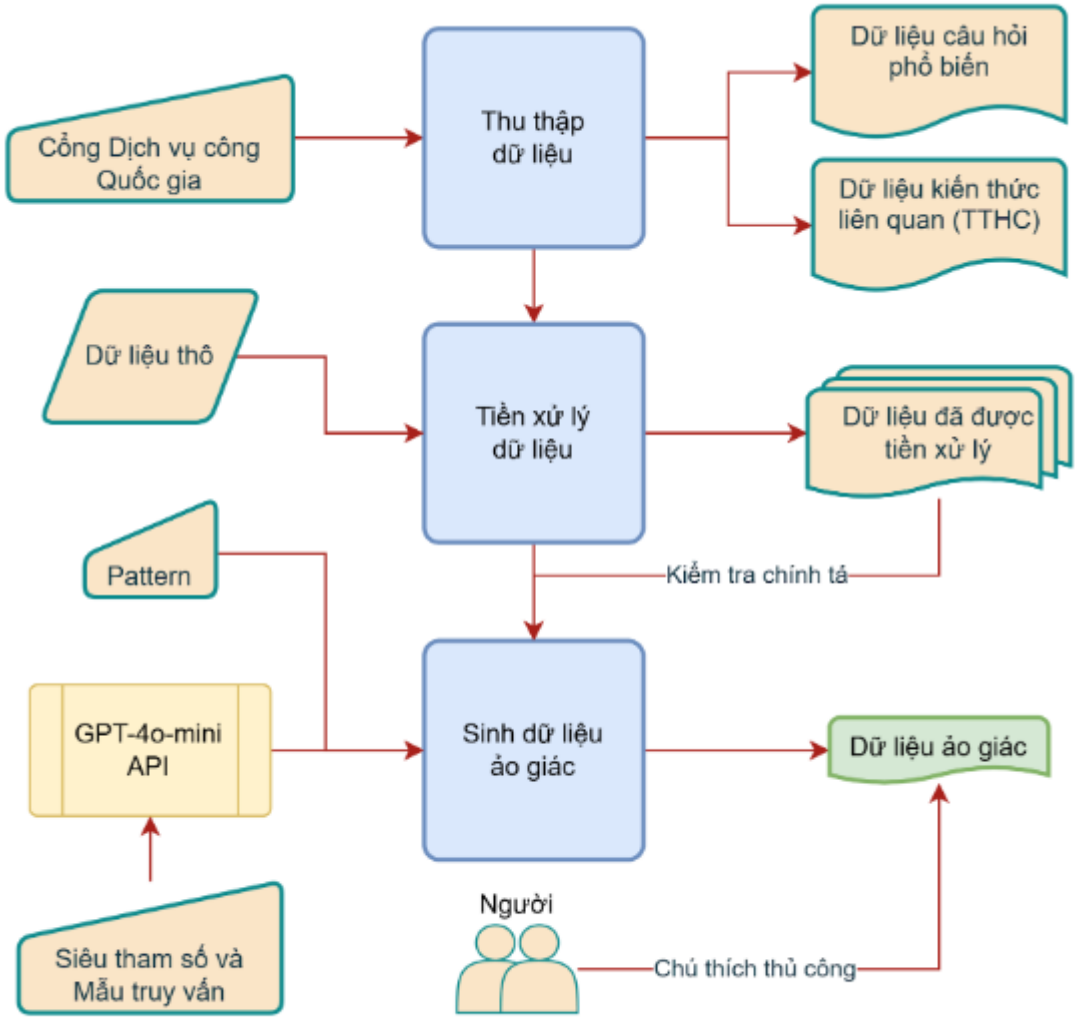
3. Phương pháp đề xuất

3.1 Tổng quan quy trình

Bước 1: Tìm kiếm và thu thập dữ liệu

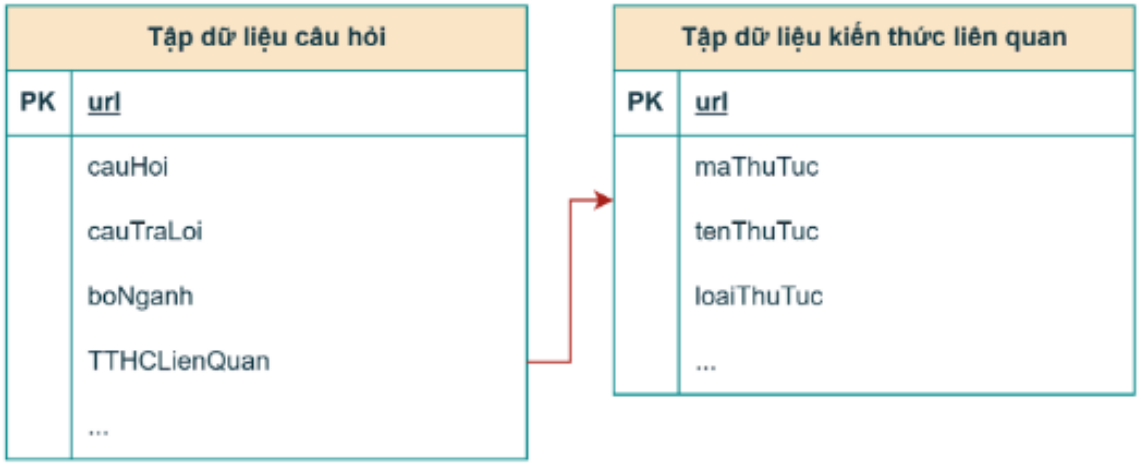
Bước 2: Tiền xử lý dữ liệu

Bước 3: Sinh dữ liệu ảo giác



Hình 3.1: Quy trình xây dựng bộ dữ liệu ảo giác trong ngữ cảnh dịch vụ công

- **Tập dữ liệu câu hỏi:** Gồm 9.452 câu hỏi thường gặp, đi kèm câu trả lời chính thức và danh sách các TTHC có liên quan.
- **Tập dữ liệu kiến thức liên quan (TTHC):** Gồm 2.695 thủ tục hành chính, mỗi thủ tục chứa đầy đủ thông tin như điều kiện, hồ sơ, cơ sở pháp lý,...



Hình 3.2: Lược đồ quan hệ các tập dữ liệu có trong quy trình

3.3 Tiền xử lý dữ liệu

- Lọc trùng lặp.
- Ưu tiên giữ lại các câu có đầy đủ thông tin về bộ/ ngành và nhiều thủ tục hành chính liên quan.
- Loại bỏ bản ghi thiếu thông tin.

- **Kiểm tra dữ liệu thủ công:**
  - Kiểm tra lỗi chính tả.
  - Phát hiện lỗi logic hoặc lỗi ngữ nghĩa.
  - Kiểm tra định dạng của các mẫu trong bộ dữ liệu trước khi tiến hành sinh ảo giác.

- **Kết quả thu được:**
  - 3.717 câu hỏi – câu trả lời hợp lệ.
  - 1.820 thủ tục hành chính liên quan được giữ lại.
  - Dữ liệu được gán nhãn theo 14 bộ/ngành khác nhau, phục vụ cho phân tích theo lĩnh vực.

Lựa chọn mô hình và siêu tham số:

- temperature=1: tăng độ ngẫu nhiên trong việc chọn từ tiếp theo.
- top_p=1: mở rộng không gian lựa chọn từ tiếp theo dựa trên xác suất tích lũy.
- max_output_tokens=512: đảm bảo độ dài phản hồi đủ lớn để diễn đạt được ảo giác.

Thiết kế truy vấn gồm:

- System prompt (cung cấp ngữ cảnh).
- User prompt (cung cấp tri thức, câu hỏi, câu trả lời đúng).

Phân loại phản hồi ảo giác thành 4 dạng (pattern):

- Hiểu sai ngữ cảnh câu hỏi.
- Mâu thuẫn với tri thức đầu vào.
- Quá chung chung hoặc quá chi tiết.
- Suy luận sai từ tri thức đúng.

Truy vấn hệ thống:
Bạn sẽ đóng vai trò là một trình tạo câu trả lời ảo giác (hallucination answer generator). Với một câu hỏi, câu trả lời đúng, và kiến thức liên quan, mục tiêu của bạn là viết một câu trả lời ảo giác mà nghe có vẻ đúng nhưng thực tế lại sai. {pattern} Bạn nên cố gắng hết sức để làm cho câu trả lời trở nên ảo giác. #Câu trả lời ảo giác# chỉ có thể nhiều hơn #Câu trả lời đúng# khoảng 5 từ.
Truy vấn người dùng (học ngữ cảnh không có ví dụ):
#Kiến thức liên quan#: {knowledge} #Câu hỏi#: {question} #Câu trả lời đúng#: {right_answer} #Câu trả lời ảo giác#:

Bảng 3.1: Mẫu truy vấn cho quá trình sinh ảo giác

➢ Kiểm tra dữ liệu câu trả lời ảo giác bằng thủ công:

- Kiểm tra lỗi chính tả.
- Phát hiện lỗi logic hoặc lỗi ngữ nghĩa.
- Kiểm tra định dạng của các mẫu trong bộ dữ liệu.

➢ Kết quả thu được:

- Tất cả phản hồi đều được xác nhận là **ảo giác hợp lệ**.
- Đảm bảo dùng cho đánh giá mô hình trong ngữ cảnh pháp lý.

→ Thu được **3717** câu trả lời ảo giác được tạo từ mô hình GPT 4o-mini.

4. Kết quả thí nghiệm

Sử dụng các mô hình mã nguồn đóng và mở để đánh giá khả năng nhận diện hallucination của các mô hình trên bộ dữ liệu đã sinh.

Độ đo: Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$

Truy vấn hệ thống:
Bạn sẽ đóng vai trò là một người đánh giá câu trả lời (answer judge). Với một câu hỏi và câu trả lời, mục tiêu của bạn là xác định xem câu trả lời được cung cấp có chứa thông tin không đúng sự thật hoặc thông tin ảo giác (hallucinated information) hay không. {pattern} Bạn nên cố gắng hết sức để xác định xem câu trả lời có chứa thông tin không đúng sự thật hoặc thông tin ảo giác hay không. Câu trả lời bạn đưa ra bắt buộc CHỈ là "Có" hoặc "Không", và không giải thích gì thêm. Trả lời "Có" nếu câu trả lời chứa thông tin ảo giác, trả lời "Không" nếu câu trả lời không chứa thông tin ảo giác.
Truy vấn người dùng (học ngữ cảnh không có ví dụ):
#Câu hỏi#: {question} #Câu trả lời#: {answer} #Đánh giá của bạn#:

Bảng 4.1: Mẫu truy vấn cho quá trình đánh giá các mô hình ngôn ngữ lớn

4.2 Kết quả và nhận xét

Mô hình	Accuracy	Mô hình	Accuracy	Mô hình	Accuracy
GPT-4o -mini	51.72	Llama-3	53.07	Vistral	50.91
Gemini-2.0 -flash	51.29	Mistral-v0.3	48.24	Qwen-Viet	53.81
DeepSeek-V3 -0324	50.26	Qwen-2.5	51.91		
Claude-3.5 -Haiku	43.58	Vicuna-v1.5	50.59		
		WizardLM-2	**57.61**		

Bảng 4.2: Bảng thống kê các mô hình khi không truyền kiến thức liên quan (đơn vị: %)

Bộ/ngành	Accuracy	Bộ/ngành	Accuracy
Thanh tra Chính phủ	50.38	Bộ Tư pháp	50.42
Bộ Khoa học và Công nghệ	51.22	Bộ Giao thông vận tải	51.66
Bộ Y tế	50.71	Bộ Ngoại giao	51.57
Bộ Quốc phòng	52.07	Bộ Công an	50.70
Bộ Nông nghiệp và Môi trường	51.56	Bộ Tài chính	50.09
Bộ Tài nguyên và Môi trường	50.81	Bộ Nội vụ	51.22
Bộ Công thương	51.08	Bộ Lao động - Thương binh và Xã hội	**52.12**

Bảng 4.3: Accuracy theo từng bộ/ngành (Đơn vị: %)

Pattern	Accuracy
P-I	51.59
P-II	**52.73**
P-III	51.97
P-IV	48.48

Bảng 4.4: Accuracy theo từng phân loại ảo giác (Đơn vị: %)

4.2 Kết quả và nhận xét

Mô hình	Không truyền tri thức		Có truyền tri thức	
	Số lượng mẫu dương (âm)	Accuracy	Số lượng mẫu dương (âm)	Accuracy
GPT-4o-mini	3717	51.72	1000	50.00
Gemini-2.0-flash	3717	51.29	1000	49.85
DeepSeek-V3-0324	3717	50.26	1000	49.90
Claude-3.5-Haiku	3717	43.58	1000	45.55
WizardLM-2	3717	**57.61**	1000	47.50
Qwen-Viet	3717	53.81	1000	**50.15**

Bảng 4.5: So sánh hiệu suất khi không truyền và khi truyền kiến thức liên quan trên những mô hình tiêu biểu (Đơn vị: %)

5. Kết luận

5 Kết luận

- Xây dựng bộ dữ liệu chuyên biệt gồm 3717 mẫu câu hỏi câu trả lời đúng, câu trả lời ảo giác để **phát hiện ảo giác** trong ngữ cảnh **dịch vụ công**.
- Sau khi tiến hành thực nghiệm đánh giá ảo giác của bộ dữ liệu trên nhiều mô hình ngôn ngữ lớn, ta rút ra kết luận: mô hình ChatGPT 4o-mini có kết quả tốt nhất đối với các mô hình mã nguồn đóng, trong khi đó, WizardLM là mô hình mã nguồn mở cho ra kết quả tốt nhất.
- Góp phần nâng cao độ chính xác & độ tin cậy khi ứng dụng LLM vào hành chính công.

Chân thành cảm ơn Quý Thầy Cô đã lắng nghe và theo dõi!