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

MÔ HÌNH 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

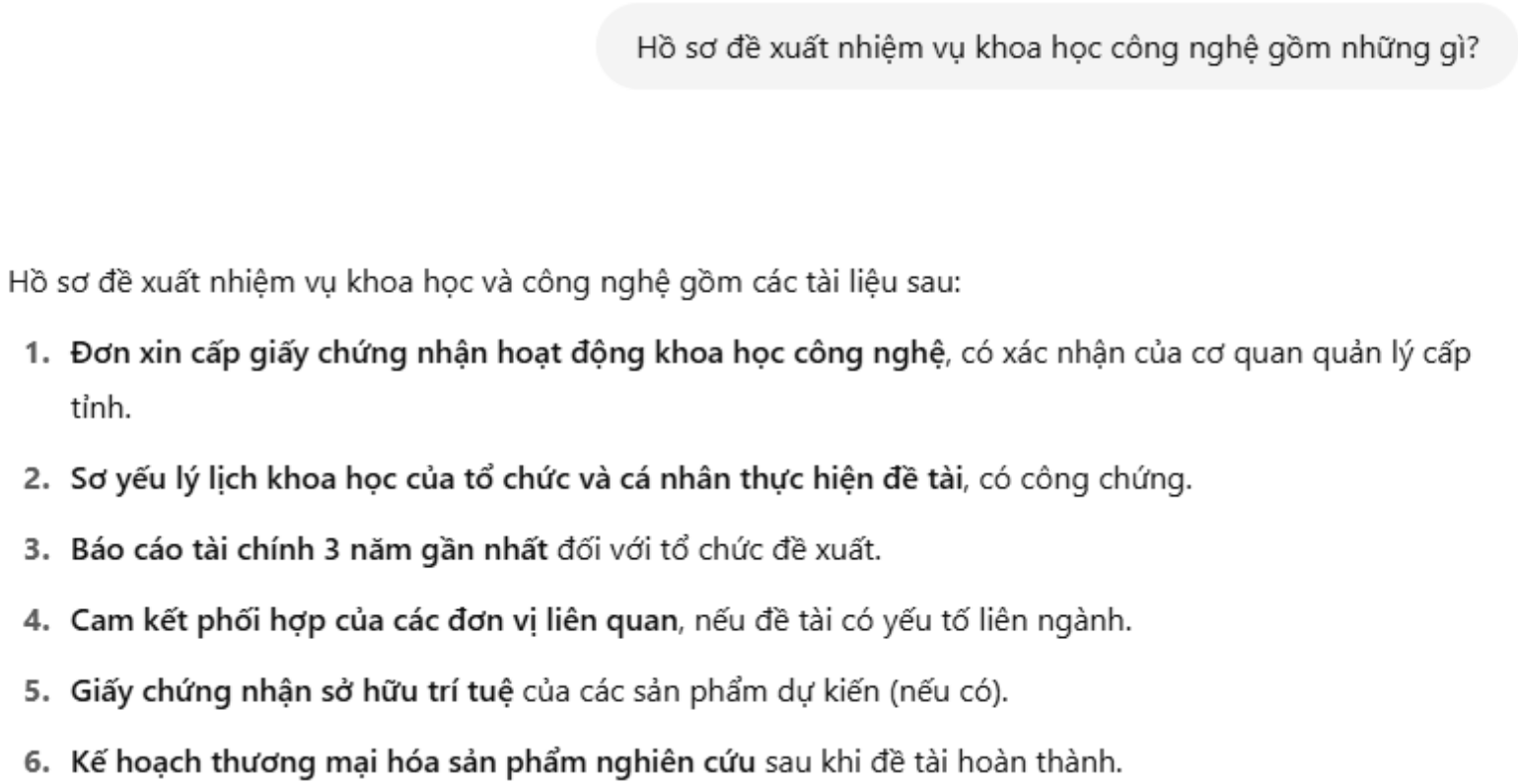
Giáo viên hướng dẫn:
TS. Nguyễn Tiến Huy
ThS. Nguyễn Trần Duy Minh

Nội dung

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

1.1 Mô tả vấn đề

- 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.



Hình 1: Một phản hồi ảo giác của GPT 4.0 về dịch vụ công

1.2 Mô tả bài toán

Cho một câu hỏi liên quan đến dịch vụ công và một câu trả lời từ 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.

Bảng 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

Câu hỏi: 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).

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.

HaluEval:

Ưu điểm:

- Là tập dữ liệu dùng để đánh giá khả năng phát hiện hallucination trong phản hồi của LLMs.
- Cho thấy kết quả đánh giá có cải thiện nếu cung cấp tri thức nền cho LLMs.

Nhược điểm:

- Công trình hiện tại đánh giá trên dữ liệu tiếng Anh, không hỗ trợ tiếng Việt.

Dichvucong.me:

Ưu điểm:

- Là chatbot hỗ trợ tra cứu thủ tục hành chính tại Việt Nam, ứng dụng LLM để trả lời câu hỏi người dân về CCCD, hộ chiếu, đăng ký khai sinh,...
- Đây là ví dụ điển hình cho việc LLM được áp dụng vào lĩnh vực dịch vụ công tiếng Việt.

Nhược điểm:

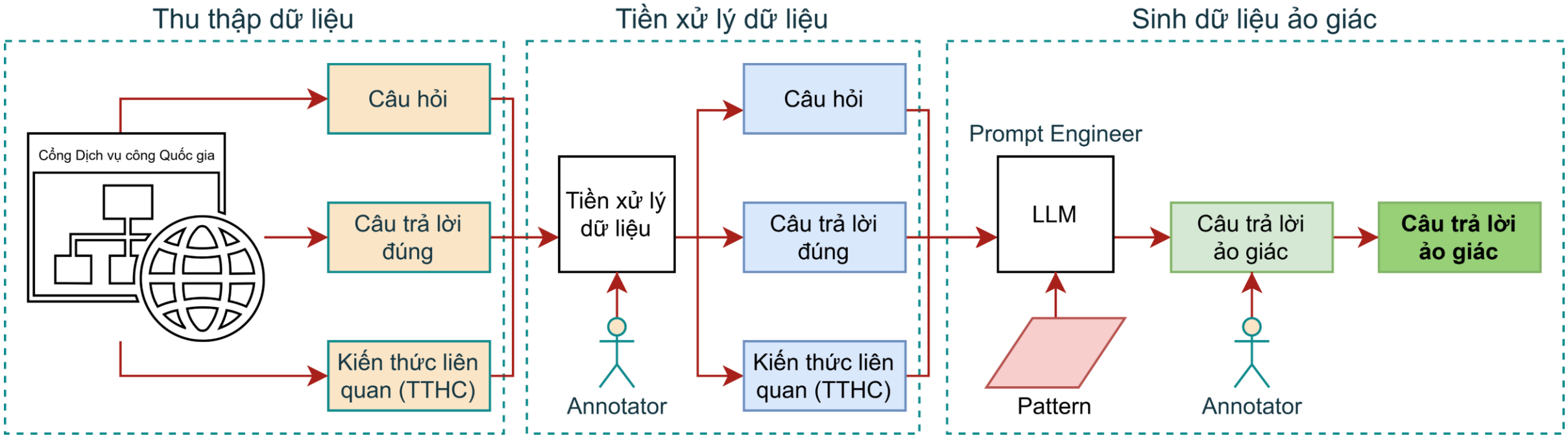
- Chưa có công trình nào đánh giá mức độ hallucination của hệ thống này, nhất là trong bối cảnh pháp lý và hậu quả nếu trả lời sai.

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.

3.1 Tổng quan quy trình

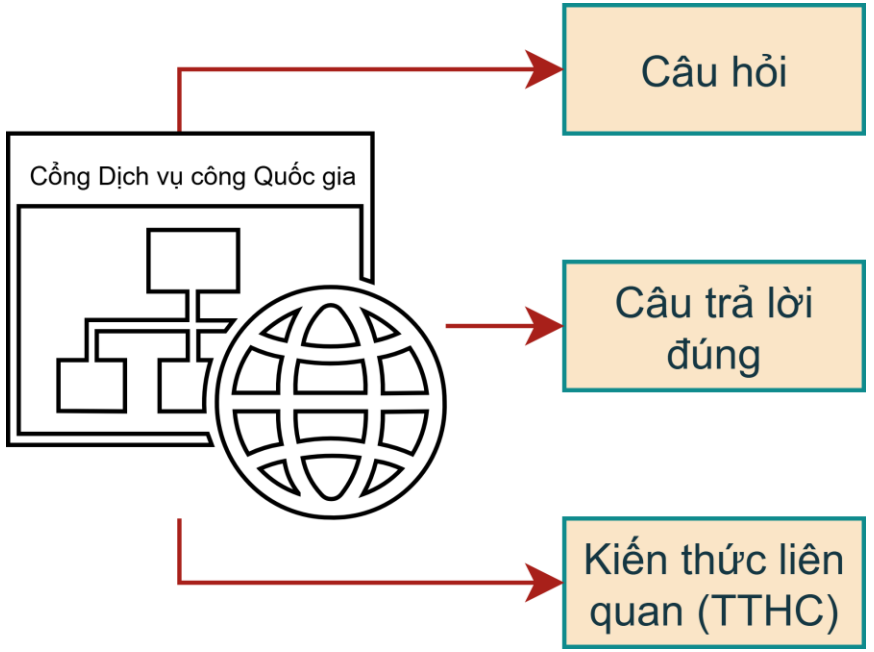
- **Bước 1:** Tìm kiếm và thu thập dữ liệu
- **Bước 2:** Tiền xử lý dữ liệu
- **Bước 3:** Sinh dữ liệu ảo giác



Hình 2: Quy trình xây dựng bộ dữ liệu phát hiện ảo giác trong ngữ cảnh dịch vụ công

Bộ dữ liệu thu được gồm:

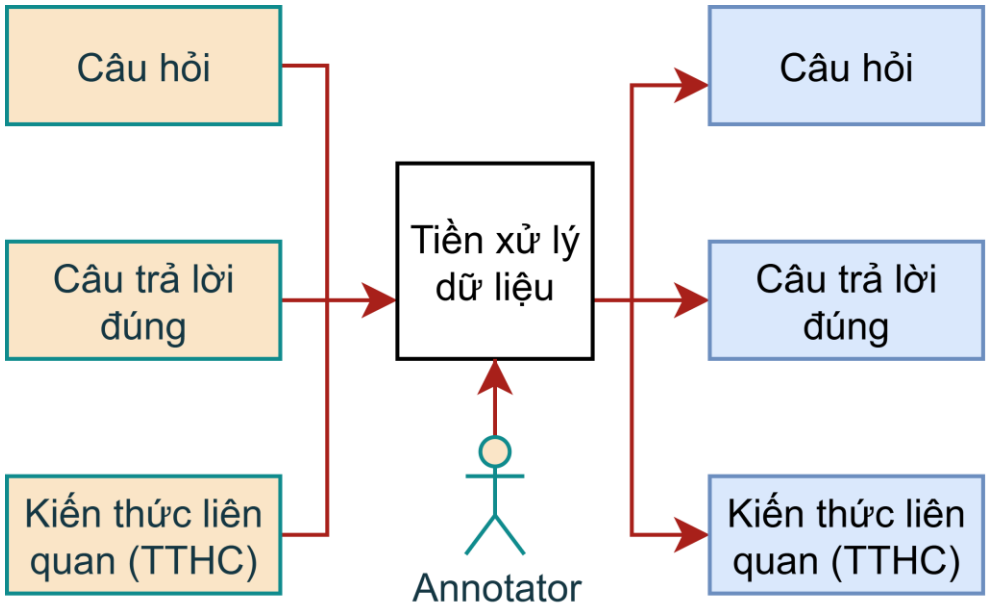
- **9.452** cặp câu hỏi và câu trả lời đúng trong ngữ cảnh dịch vụ công.
- **2.695** thủ tục hành chính chứa kiến thức liên quan của câu hỏi.



Hình 3: Quy trình thu thập dữ liệu

Bộ dữ liệu thu được gồm:

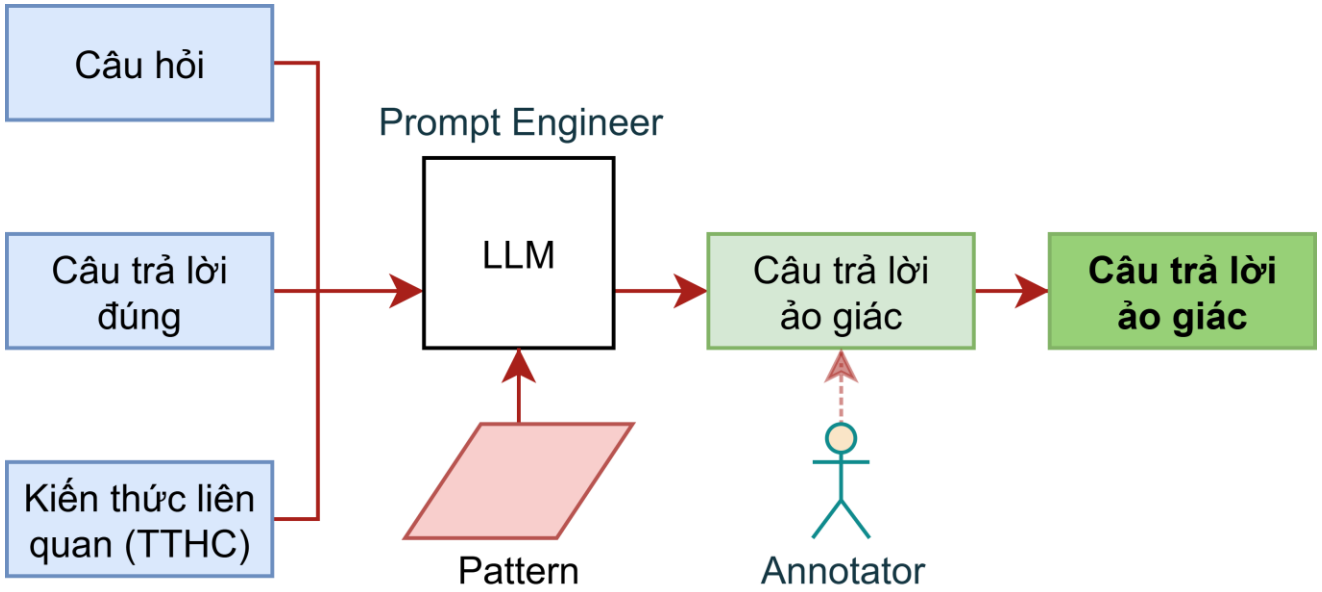
- **3.717** cặp câu hỏi và câu trả lời đúng trong ngữ cảnh dịch vụ công.
- **1.820** thủ tục hành chính chứa kiến thức liên quan của câu hỏi.



Hình 4: Quy trình tiền xử lý dữ liệu

Pattern (phân loại ảo giác):

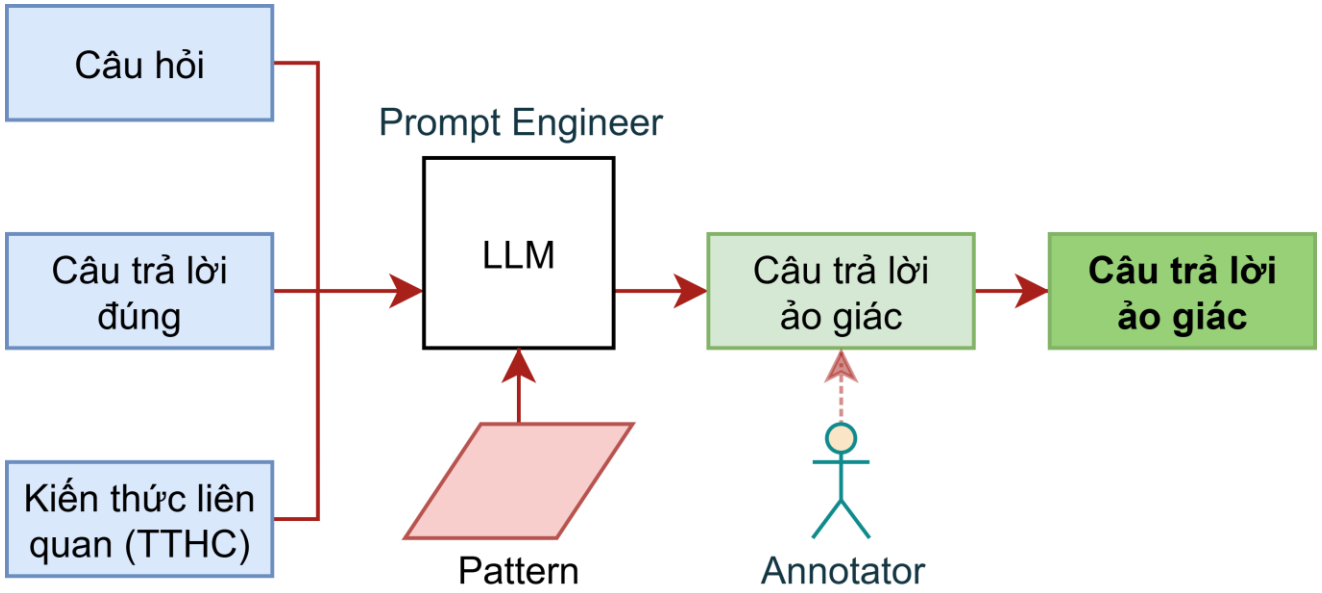
- Hiểu sai ngữ cảnh và mục đích (P-I).
- Mâu thuẫn giữa câu trả lời và tri thức (P-II).
- Quá chung chung hoặc quá chi tiết (P-III).
- Suy luận sai từ tri thức (P-IV).



Hình 5: Quy trình sinh dữ liệu ảo giác

3.4 Sinh dữ liệu ảo giác

- LLM được sử dụng để sinh ảo giác: **GPT-4o-mini**.
- Bộ dữ liệu ảo giác thu được gồm: **3.717** câu trả lời ảo giác.

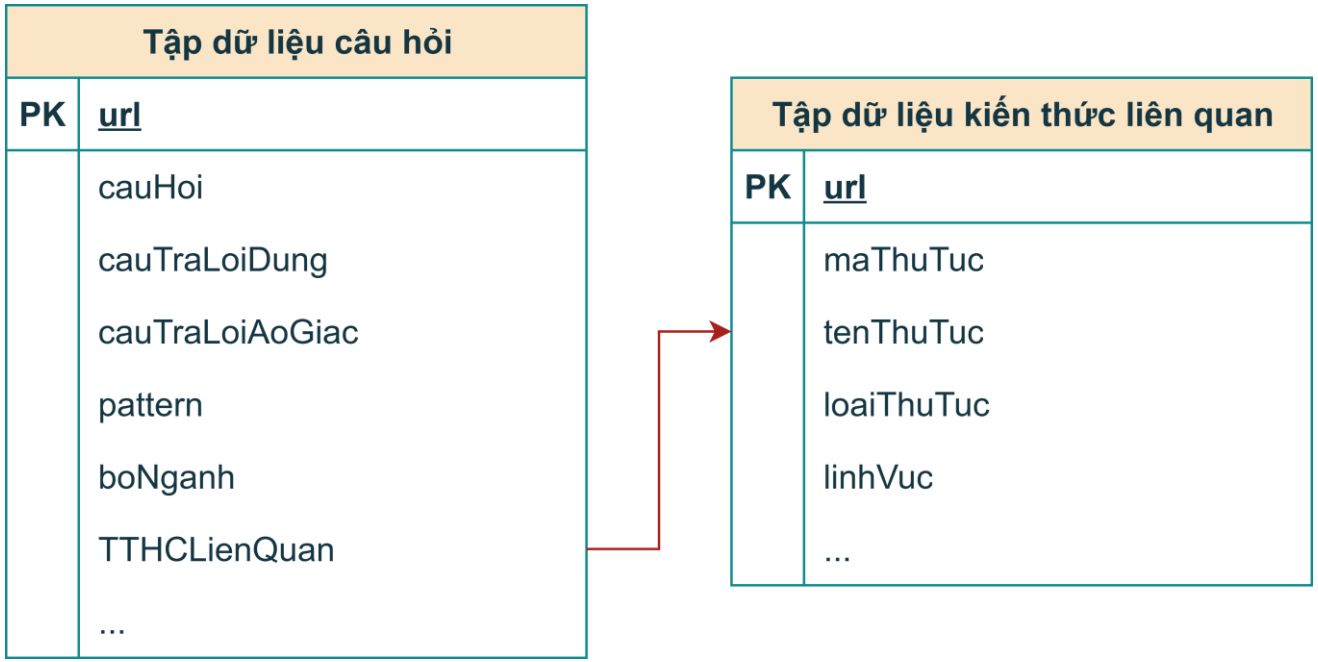


Hình 6: Quy trình sinh dữ liệu ảo giác

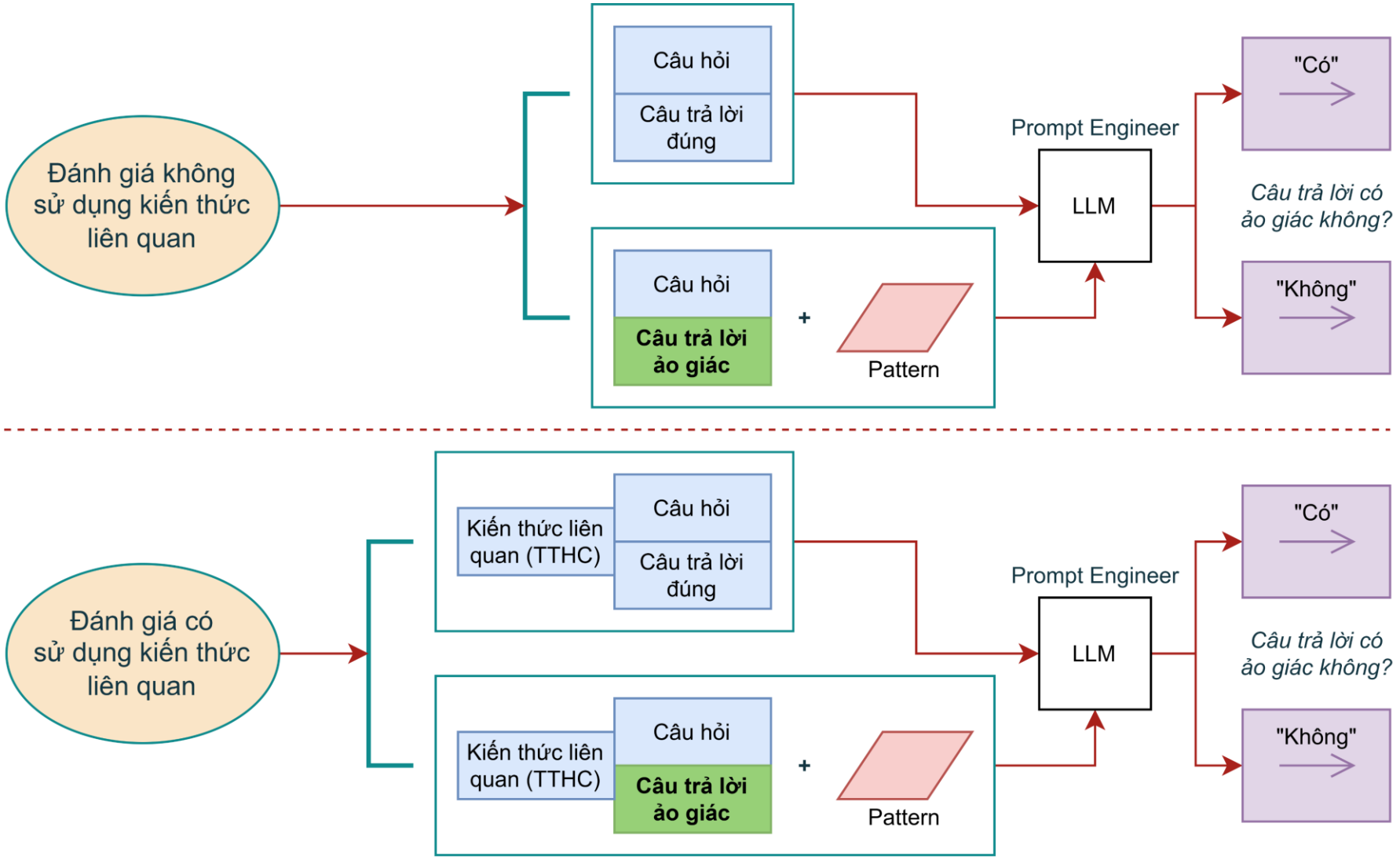
3.5 Mô tả về bộ dữ liệu

Bộ dữ liệu thu được gồm:

- **3.717** cặp câu hỏi và câu trả lời đúng cùng với câu trả lời ảo giác và pattern tương ứng.
- **1.820** thủ tục hành chính chứa kiến thức liên quan của câu hỏi.



Hình 7: Lược đồ quan hệ các tập dữ liệu



Hình 8: Quy trình đánh giá ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công

4.1 Phương pháp đánh giá

Dữ liệu đầu vào:

Không truyền kiến thức: **7.434** cặp câu hỏi và câu trả lời trong ngữ cảnh dịch vụ công.

- **3.717** cặp câu hỏi và câu trả lời đúng (mẫu âm).
- **3.717** cặp câu hỏi và câu trả lời ảo giác cùng với pattern tương ứng (mẫu dương).

Có truyền kiến thức: **2.000** cặp câu hỏi và câu trả lời trong ngữ cảnh dịch vụ công.

*(Được lấy sample từ **7.434** mẫu bên trên)*

- **1.000** cặp câu hỏi và câu trả lời đúng (mẫu âm).
- **1.000** cặp câu hỏi và câu trả lời ảo giác cùng với pattern tương ứng (mẫu dương).

Dữ liệu thu được từ LLM:

- **7.434** nhãn nhị phân khi không truyền kiến thức liên quan.
- **2.000** nhãn nhị phân khi có truyền kiến thức liên quan.

Độ đo sử dụng:

- $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$

Bảng 2: Định nghĩa Confusion Matrix trong quá trình đánh giá

		Thực tế là:	
		Câu trả lời ảo giác	Câu trả lời đúng
Output của LLM là:	Câu trả lời ảo giác	TP (True Positive)	FP (False Positive)
	Câu trả lời đúng	FN (False Negative)	TN (True Negative)

Bảng 3: Accuracy (%) của các mô hình khi không truyền kiến thức liên quan

Mã nguồn đóng/ truy cập qua API		Mã nguồn mở chưa được tinh chỉnh trên tiếng Việt		Mã nguồn mở đã được tinh chỉnh trên tiếng Việt	
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: Accuracy (%) của mô hình WizardLM-2 khi không truyền kiến thức liên quan trên các bộ/ngành khác nhau

Bộ/Ngành	Accuracy	Số mẫu	Bộ/Ngành	Accuracy	Số mẫu
Bộ Nông nghiệp và Môi trường	59.28	1174	Bộ Y tế	58.56	526
Bộ Giao thông vận tải	55.44	900	Bộ Nội vụ	57.45	416
Bộ Khoa học và Công nghệ	57.43	808	Bộ Tài chính	56.16	292
Bộ Tư pháp	58.94	738	Thanh tra Chính phủ	51.92	260
Bộ Công an	54.41	680	Bộ Tài nguyên và Môi trường	56.45	248
Bộ Quốc phòng	62.58	620	Bộ Công Thương	54.76	126
Bộ Ngoại giao	57.19	556	Bộ LĐ-TB và XH	**64.44**	90
			Tổng số mẫu:		**7434**

Bảng 5: Accuracy (%) của mô hình WizardLM-2 khi không truyền kiến thức liên quan trên các phân loại ảo giác (pattern)

Pattern	Mô tả	Accuracy	Số mẫu
P-I	Hiểu sai ngữ cảnh và mục đích	55.80	922
P-II	Mâu thuẫn giữa câu trả lời và tri thức	54.48	915
P-III	Quá chung chung hoặc quá chi tiết	**65.89**	941
P-IV	Suy luận sai từ tri thức	54.15	939
Tổng số mẫu:			**3717**

Bảng 6: Accuracy (%) khi không truyền và khi truyền kiến thức liên quan trên những mô hình tiêu biểu

Mô hình	Không truyền kiến thức		Có truyền kiến thức	
	Số mẫu dương (âm)*	Accuracy	Số mẫu dương (âm)*	Accuracy
GPT-4o-mini	3717	51.72	1000	50.00
WizardLM-2	3717	**57.61**	1000	47.50
Qwen-Viet	3717	53.81	1000	**50.15**

**: Số lượng mẫu dương = Số lượng mẫu âm*

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**.
- 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.

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

Phụ lục trả lời câu hỏi

Bảng: Annotation Guidelines cho giai đoạn tiền xử lý dữ liệu

Annotation Guidelines
➢ Kiểm tra chính tả
➢ Kiểm tra định dạng

Prompt Engineer:

Bảng: Mẫu truy vấn cho quá trình sinh ảo giác

System prompt:
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.
User prompt (zero-shot learning):
#Kiến thức liên quan#: {knowledge} #Câu hỏi#: {question} #Câu trả lời đúng#: {right_answer} #Câu trả lời ảo giác#:

Bảng: Annotation Guidelines cho giai đoạn sinh dữ liệu ảo giác

Annotation Guidelines
➢ Kiểm tra xem câu trả lời do LLM sinh ra có thực sự là ảo giác hay không

Prompt Engineer:

Bảng: Mẫu truy vấn cho quá trình đánh giá không sử dụng kiến thức liên quan

System prompt:
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.
User prompt (zero-shot learning):
#Câu hỏi#: {question} #Câu trả lời#: {answer} #Đánh giá của bạn#:

Prompt Engineer:

Bảng: Mẫu truy vấn cho quá trình đánh giá có sử dụng kiến thức liên quan

System prompt:
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.
User prompt (zero-shot learning):
#Kiến thức liên quan#: {knowledge} #Câu hỏi#: {question} #Câu trả lời#: {answer} #Đánh giá của bạn#: