



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

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

- LLMs tiềm năng hỗ trợ dịch vụ công (tra cứu, hướng dẫn thủ tục).
- LLMs gây sai lệch thông tin, dễ hiểu lầm pháp lý trong khi dịch vụ công đòi hỏi tính chính xác và minh bạch.

Hồ sơ đề xuất nhiệm vụ khoa học công nghệ gồm những gì?

Hồ sơ đề xuất nhiệm vụ khoa học và công nghệ gồm các tài liệu sau:

1. Đơn xin cấp giấy chứng nhận hoạt động khoa học công nghệ, có xác nhận của cơ quan quản lý cấp tỉnh.
2. Sơ yếu lý lịch khoa học của tổ chức và cá nhân thực hiện đề tài, có công chứng.
3. Báo cáo tài chính 3 năm gần nhất đối với tổ chức đề xuất.
4. Cam kết phối hợp của các đơn vị liên quan, nếu đề tài có yếu tố liên ngành.
5. Giấy chứng nhận sở hữu trí tuệ của các sản phẩm dự kiến (nếu có).
6. Kế hoạch thương mại hóa sản phẩm nghiên cứu sau khi đề tài hoàn thành.

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

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

Xây dựng bộ dữ liệu ảo giác tiếng Việt nhằm đánh giá mức độ ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công

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:

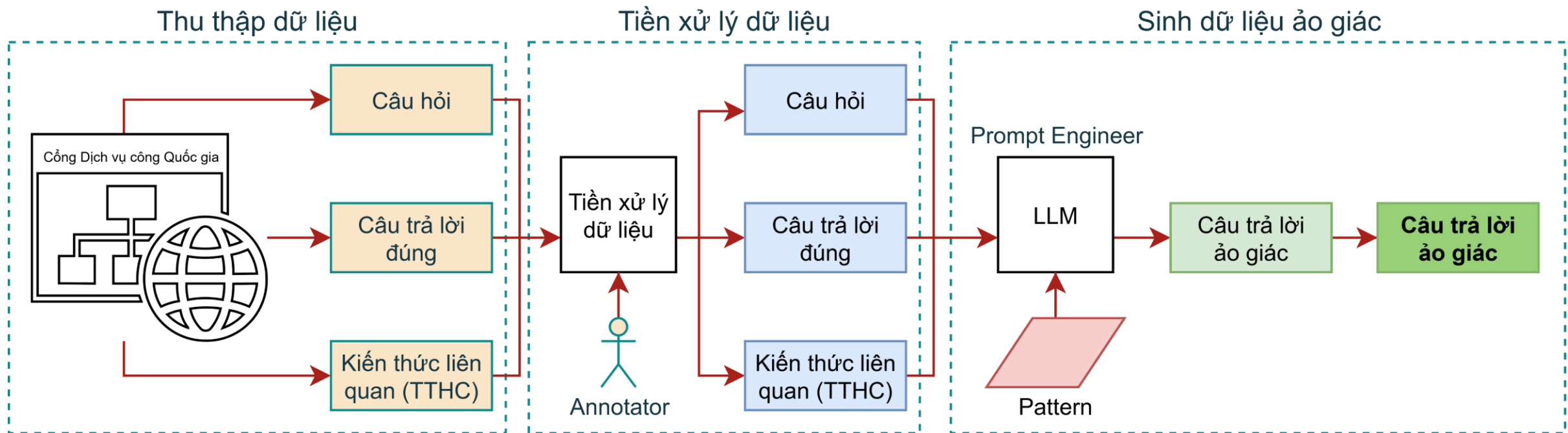
Ưu điểm:

- 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.

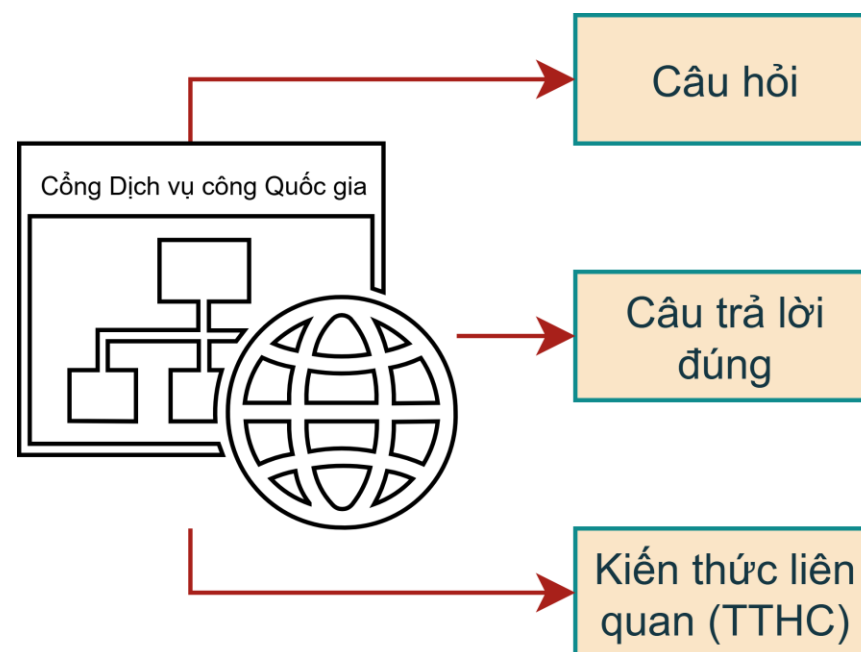
- **Bước 1:** Tìm kiếm và thu thập dữ liệu
- **Bước 2:** Tiền xử lý dữ liệu
- **Bước 3:** Sinh dữ liệu ảo giác



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:

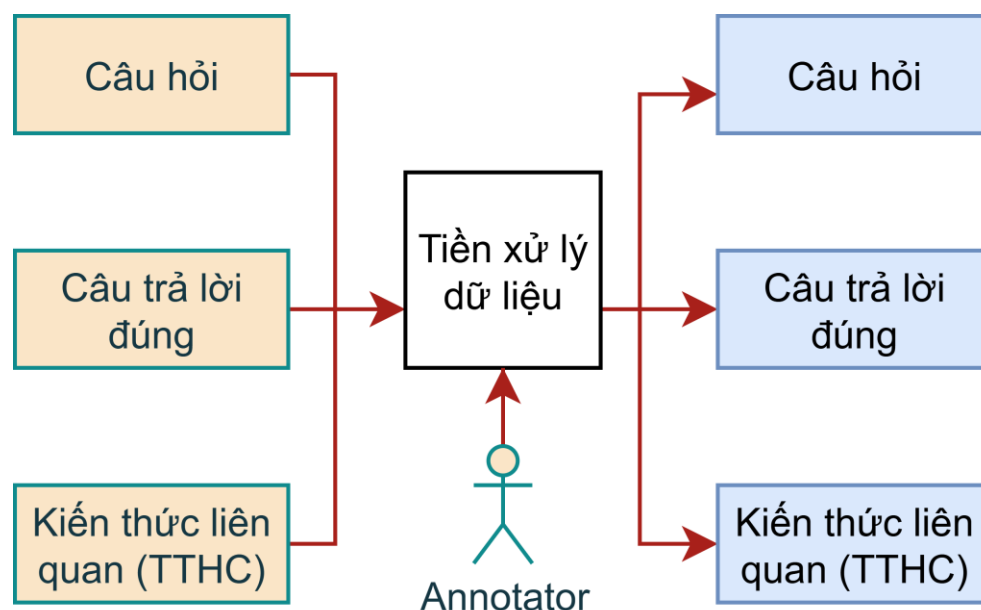
- **9.452** cặp câu hỏi và câu trả lời đúng trong ngữ cảnh dịch vụ công.
- **2.695** thủ tục hành chính chứa kiến thức liên quan của câu hỏi.



Hình 3: Quy trình thu thập dữ liệu

Bộ dữ liệu thu được gồm:

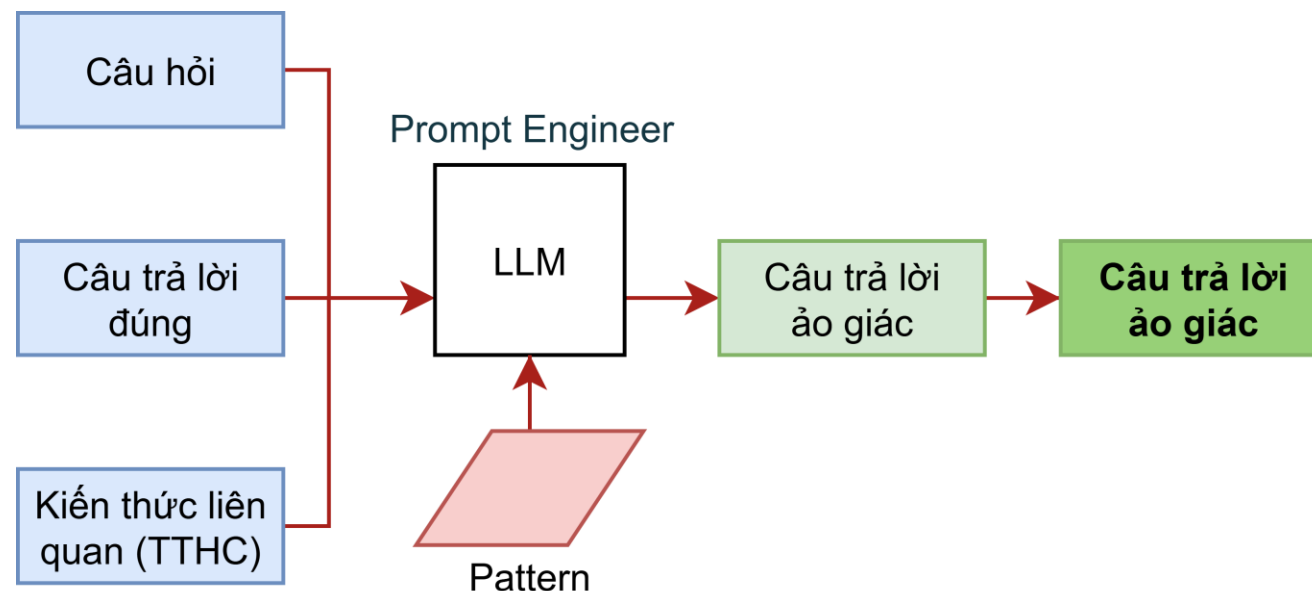
- **3.717** cặp câu hỏi và câu trả lời đúng trong ngữ cảnh dịch vụ công.
- **1.820** thủ tục hành chính chứa kiến thức liên quan của câu hỏi.



Hình 4: Quy trình tiền xử lý dữ liệu

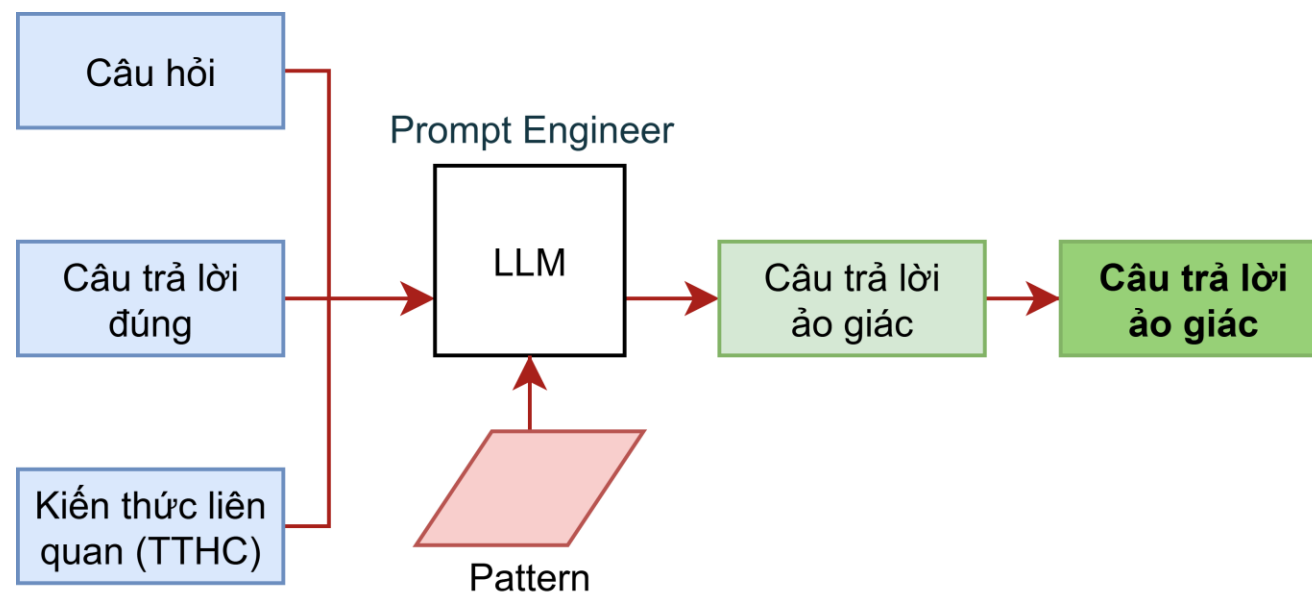
Pattern (phân loại ảo giác):

- 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

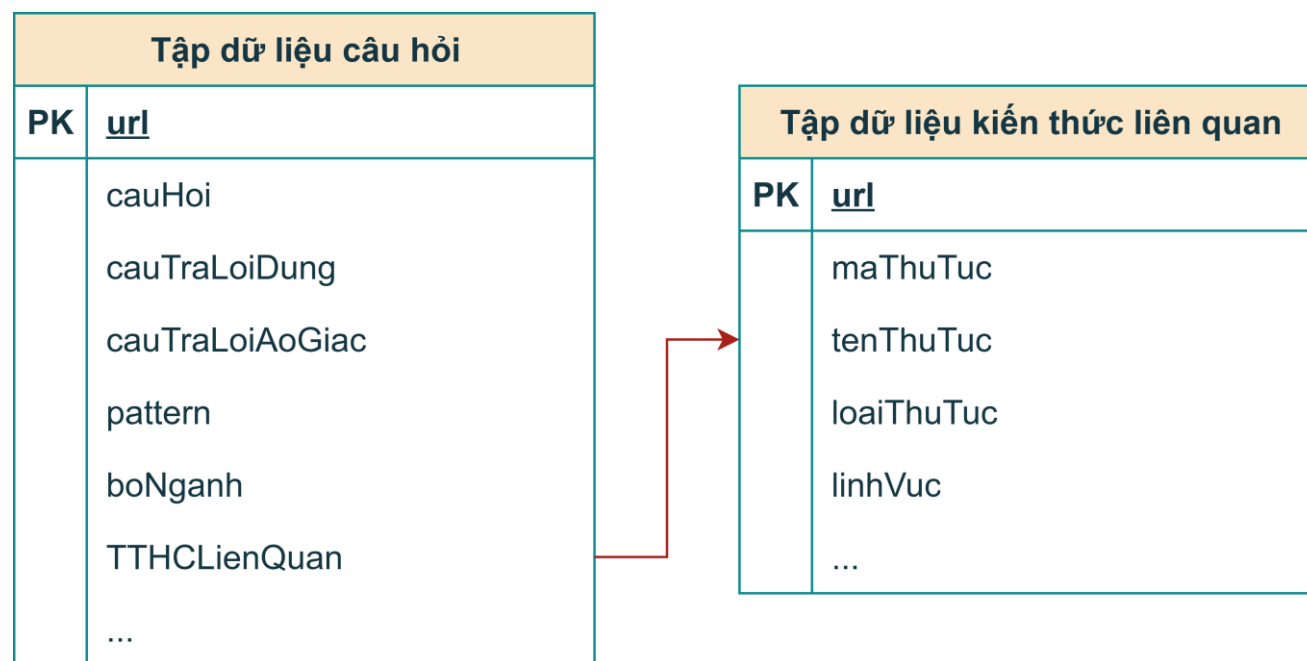
- LLM được sử dụng để sinh ảo giác: **GPT-4o-mini**.
- Bộ dữ liệu ảo giác thu được gồm: **3.717** câu trả lời ảo giác.



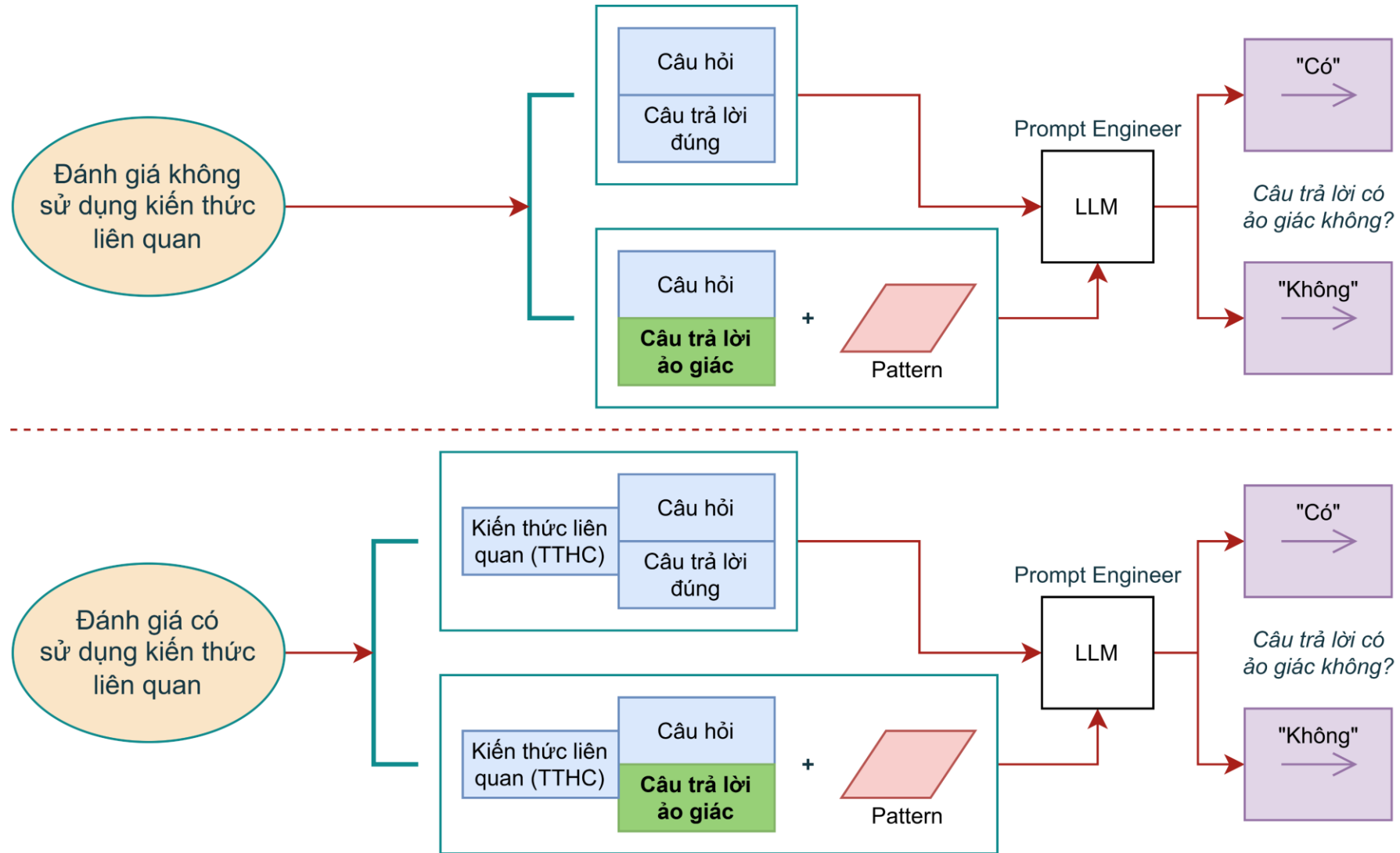
Hình 6: Quy trình sinh dữ liệu ảo giác

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

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).

Kết quả 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:

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

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 | <u>57.61</u> | | |

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 | <u>64.44</u> | 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 | <u>65.89</u> | 941 |
| P-IV | Suy luận sai từ tri thức | 54.15 | 939 |
| <i>Tổng số mẫu: 3717</i> | | | |

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 | 1000 | 51.80 | 1000 | 50.00 |
| WizardLM-2 | 1000 | <u>58.60</u> | 1000 | 47.50 |
| Qwen-Viet | 1000 | 53.80 | 1000 | <u>50.15</u> |

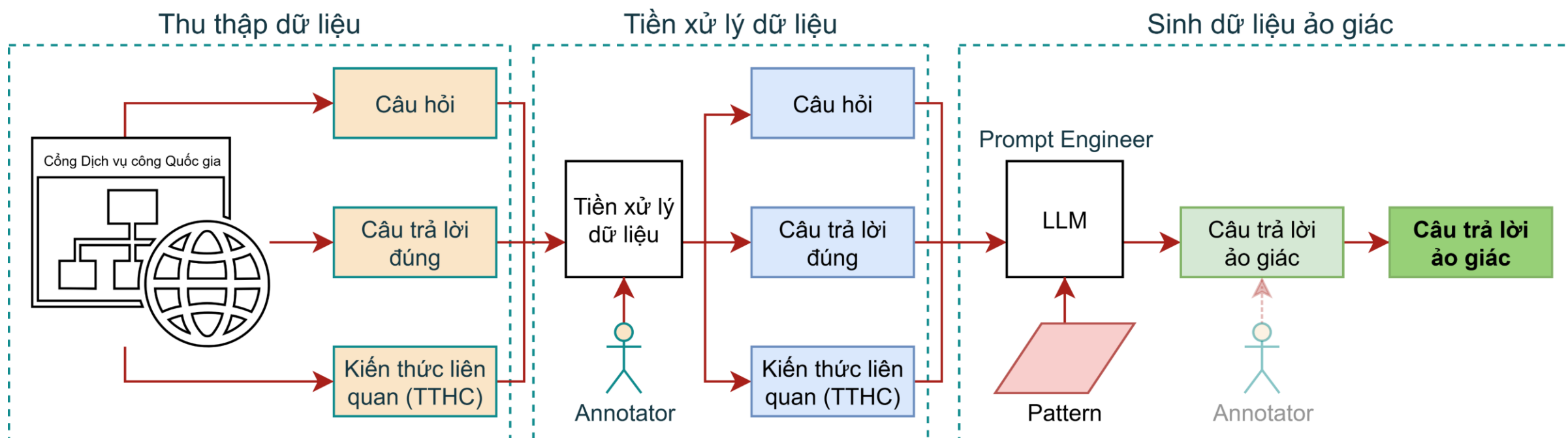
**: Số lượng mẫu dương = Số lượng mẫu âm*

- 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

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



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

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#:

Bảng 3: Accuracy (%) và False Negative 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 | FN | Mô hình | Accuracy | FN | Mô hình | Accuracy | FN |
| GPT-4o-mini | 51.72 | 431 | LLaMA-3 | 53.07 | 888 | Vistral | 50.91 | 3469 |
| Gemini-2.0-flash | 51.29 | 1128 | Mistral-v0.3 | 48.24 | 2050 | Qwen-Viet | 53.81 | 901 |
| DeepSeek-V3-0324 | 50.26 | 525 | Qwen-2.5 | 51.91 | 407 | | | |
| Claude-3.5-Haiku | 43.58 | 2762 | Vicuna-v1.5 | 50.59 | 12 | | | |
| | | | WizardLM-2 | 57.61 | 2052 | | | |
| Số lượng mẫu dương = Số lượng mẫu âm = 3717 | | | | | | | | |