# VIETPS-HALLU: BỘ DỮ LIỆU PHÁT HIỆN ẢO GIÁC CỦA MÔ HÌNH NGÔN NGỮ LỚN TRONG NGỮ CẢNH DỊCH VỤ CÔNG

**Bùi Đình Bảo[1], Nguyễn Tiến Nhật[1], Nguyễn Tiến Huy[1], Lê Thanh Tùng[1]**

[1] Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

*bao.buidinh.03@gmail.com, tiennhat5103@gmail.com, ntienhuy@fit.hcmus.edu.vn, lttung@fit.hcmus.edu.vn*

*Tác giả liên hệ: Nguyễn Tiến Huy (ntienhuy@fit.hcmus.edu.vn)*

***TÓM TẮT***— *Mô hình ngôn ngữ lớn (LLMs), như ChatGPT, đang ngày càng được ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm cả hành chính công nhờ khả năng xử lý và tạo sinh ngôn ngữ tự nhiên. Tuy nhiên, một trong những thách thức lớn trong việc triển khai LLM vào môi trường thực tế là hiện tượng "ảo giác", khi mô hình tạo ra thông tin không chính xác, sai lệch hoặc không có căn cứ thực tế. Nhằm góp phần giải quyết giải quyết vấn đề này trong bối cảnh dịch vụ công, bài báo này giới thiệu VietPS-Hallu (Vietnamese Public Service Hallucination Dataset for LLMs) - một bộ dữ liệu chuyên biệt nhằm đánh giá khả năng phát hiện ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công thông qua xây dựng một bộ khung hoàn chỉnh từ bước chuẩn bị dữ liệu cho đến đánh giá hiệu suất của các mô hình ngôn ngữ lớn khác nhau trên bộ dữ liệu đó. Chúng tôi cung cấp một bộ dữ liệu thực nghiệm có giá trị cho việc cải thiện khả năng nhận biết và giảm thiểu sinh ảo giác ở mô hình ngôn ngữ lớn, qua đó góp phần xây dựng và cải thiện độ tin cậy ở mô hình ngôn ngữ lớn trong các tác vụ hành chính, nơi đòi hỏi tính nhất quán và đúng đắn về mặt thông tin đầu ra.*

***Từ khóa***— *Mô hình ngôn ngữ lớn, dịch vụ công, phát hiện ảo giác, dữ liệu tiêu chuẩn.*

## I. GIỚI THIỆU

Trong những năm gần đây, chúng ta đã chứng kiến sự phát triển vượt bậc của công nghệ trí tuệ nhân tạo, đặc biệt là trong lĩnh vực xử lý ngôn ngữ tự nhiên. Trong số đó, các mô hình ngôn ngữ lớn (Large Language Models - LLMs) nổi lên như một công nghệ chủ chốt với khả năng không chỉ tiếp nhận thông tin mà còn có thể tạo ra văn bản, trả lời câu hỏi, tóm tắt tài liệu và hỗ trợ người dùng trong các tác vụ phức tạp. Nhờ những khả năng này, mô hình ngôn ngữ lớn đã và đang được ứng dụng rộng rãi trong y tế, giáo dục, truyền thông, quảng cáo,... Xét về khía cạnh dịch vụ công, mô hình ngôn ngữ lớn cũng cho thấy được tiềm năng to lớn nhờ khả năng xử lý và tạo sinh văn bản tự động, từ đó có thể hỗ trợ được người dân trong các vấn đề về giấy tờ, hồ sơ liên quan, các bước thực hiện thủ tục hành chính một cách nhanh chóng và thuận tiện hơn. Bên cạnh việc mang lại nhiều lợi ích, mô hình ngôn ngữ lớn cũng tồn tại một số rủi ro đáng kể, đặc biệt là hiện tượng tạo ra văn bản không chính xác, vô nghĩa, không dựa trên dữ liệu thực tế hoặc mang tính bịa đặt - còn được gọi là hiện tượng "ảo giác" (hallucination). Vấn đề này đặc biệt nghiêm trọng trong bối cảnh dịch vụ công, nơi mà độ chính xác và đáng tin cậy của đầu ra mô hình là vô cùng quan trọng. Nguyên nhân của hiện tượng ảo giác này có thể xuất phát từ nhiều yếu tố khác nhau, bao gồm dữ liệu được đưa vào không đầy đủ hoặc thiếu đúng đắn, cũng như những sai lệch trong quá trình huấn luyện mô hình mà không thể kiểm soát được. Vì vậy việc kiểm soát và giảm thiểu ảo giác trong LLM là điều kiện tiên quyết để có thể ứng dụng hiệu quả công nghệ này trong quản lý hành chính và phục vụ người dân.

Tuy nhiên, hiện nay vẫn còn thiếu hụt nghiêm trọng các bộ dữ liệu chuyên biệt nhằm đánh giá hiện tượng ảo giác trong các lĩnh vực đặc thù như dịch vụ công. Điều này khiến việc kiểm chứng đầu ra của mô hình trở nên khó khăn. Chính vì thế, nghiên cứu của chúng tôi tập trung xây dựng VietPS-Hallu, một bộ dữ liệu dùng để đánh giá hiện tượng ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công. Nền tảng phía sau bộ dữ liệu này chính là một khung phương pháp hoàn chỉnh từ bước chuẩn bị dữ liệu, thiết kế truy vấn (prompt engineering), sinh dữ liệu bằng mô hình ngôn ngữ lớn cho đến đánh giá hiệu suất của các mô hình ngôn ngữ lớn khác nhau. Điểm nổi bật của bộ dữ liệu là sự kết hợp giữa khả năng tạo sinh nhanh chóng của mô hình ngôn ngữ lớn và quy trình thẩm định có sự tham gia của con người nhằm đảm bảo tính chính xác, nhất quán và độ tin cậy của đầu ra. Phương pháp này giúp nâng cao chất lượng dữ liệu mà còn tiết kiệm đáng kể thời gian và chi phí gán nhãn.

| |
|---|
| **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: 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 VietPS-Hallu

## II. CÔNG TRÌNH LIÊN QUAN

Để giải quyết hiện tượng ảo giác trong mô hình ngôn ngữ lớn, nhiều công trình nghiên cứu trước đây đã từng đưa ra các phương pháp đánh giá hiện tượng ảo giác của các mô hình ngôn ngữ lớn thông qua việc xây dựng các bộ dữ liệu tiêu chuẩn (benchmark dataset). Chẳng hạn, BEGIN là một bộ dữ liệu phân loại các câu trả lời của hệ thống đối thoại thành 3 mức độ: fully attributable (hoàn toàn được hỗ trợ bởi tri thức), not fully attributable (tạm dịch là không hoàn toàn được hỗ trợ bởi tri thức) và generic (câu trả lời chung chung) [1]. Tương tự AIS, một bộ dữ liệu được phát triển nhằm xác định liệu các tài liệu nguồn có thực sự hỗ trợ đầu ra của các mô hình tạo sinh văn bản hay không [2].

Một số bộ dữ liệu khác tập trung vào đánh giá tính thực tế ở cấp độ câu, ví dụ như SelfCheckGPT-Wikibio - một bộ dữ liệu được tạo ra bằng cách tổng hợp các bài viết từ Wikipedia với GPT-3, được chú thích thủ công về tính thực tế, gây ra thách thức cho việc phát hiện ảo giác về tiểu sử cá nhân [3]. Trong khi đó, FELM - một bộ dữ liệu được chú thích về tính thực tế trên nhiều lĩnh vực khác nhau bao gồm kiến thức thế giới, khoa học và toán học [4]. Ngoài ra, còn có HaluEval kết hợp tạo sinh tự động với chú thích của con người để xây dựng tập dữ liệu nhằm đánh giá khả năng nhận diện ảo giác của các mô hình ngôn ngữ lớn [5].

Tuy những bộ dữ liệu kể trên có giá trị trong việc đánh giá hiện tượng ảo giác trong nhiều ngữ cảnh khác nhau, song phần lớn đều tập trung vào các lĩnh vực như y tế, khoa học, hoặc kiến thức tổng quát. Ngữ cảnh dịch vụ công, với đặc thù yêu cầu độ chính xác cao, tuân thủ pháp lý và sự đúng đắn tuyệt đối trong phản hồi, vẫn là một vùng trống chưa được khai thác nhiều. Thực tế, đã có những nỗ lực tích hợp LLM vào hệ thống hành chính công – điển hình là trợ lý ảo dịch vụ công, được triển khai nhằm hỗ trợ người dân tiếp cận thông tin trên hơn 15 nhóm chủ đề phổ biến [6]. Tuy nhiên, các hệ thống này vẫn gặp nhiều hạn chế, đặc biệt trong các tình huống yêu cầu sự chính xác pháp lý hoặc xử lý trường hợp đặc thù, nơi mà chỉ một sai sót nhỏ cũng có thể gây hậu quả lớn. Chính vì thế, việc phát triển một bộ dữ liệu chuyên biệt để đánh giá và giảm thiểu hiện tượng ảo giác trong ngữ cảnh dịch vụ công là hoàn toàn cần thiết. Bộ dữ liệu như vậy không chỉ giúp lấp đầy khoảng trống nghiên cứu, mà còn tạo nền tảng cho việc triển khai mô hình ngôn ngữ lớn một cách an toàn và đáng tin cậy hơn trong hành chính công.

## III. PHƯƠNG PHÁP XÂY DỰNG BỘ DỮ LIỆU ẢO GIÁC TRONG NGỮ CẢNH DỊCH VỤ CÔNG

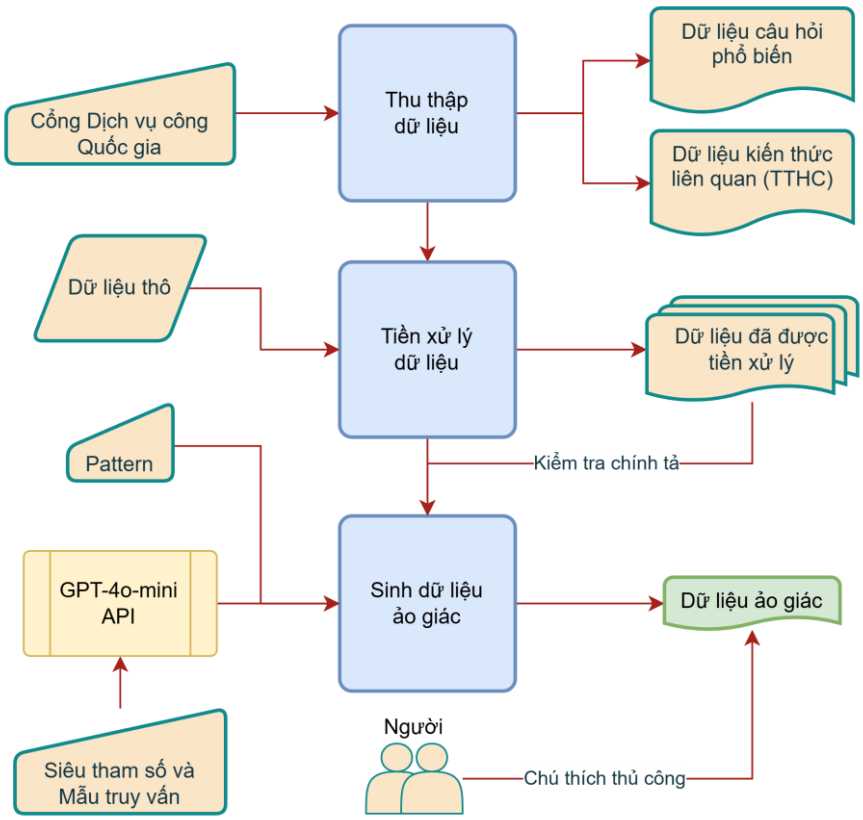
### *A. Tổng quan quy trình xây dựng bộ dữ liệu*

Việc đánh giá hiện tượng ảo giác trong các mô hình ngôn ngữ lớn (LLMs) đòi hỏi một bộ dữ liệu được thiết kế bài bản và chính xác. Trong ngữ cảnh dịch vụ công, bộ dữ liệu còn đòi hỏi cần phải phản ánh được thực tế các vấn đề của công dân, có thể truy xuất được nguồn gốc của những thông tin liên quan và mức độ phù hợp về mặt pháp lý. Để giải quyết vấn đề trên, chúng tôi đề xuất một quy trình xây dựng bộ dữ liệu ảo giác có giám sát, hỗ trợ tiếng Việt. Quy trình sẽ gồm có 3 công đoạn chính, bao gồm: (i) Tìm kiếm và thu thập dữ liệu; (ii) Tiền xử lý dữ liệu cùng với kiểm tra chính tả; (iii) Sinh dữ liệu ảo giác kết hợp với chú thích thủ công. Sơ đồ tổng quát của quy trình được thể hiện ở Hình 1, thiết kế quy trình theo từng bước rõ ràng giúp đảm bảo tính tự động hóa, khả năng tái lập, và đặc biệt là có thể đánh giá chính xác các mô hình ngôn ngữ lớn trong ngữ cảnh có yêu cầu đặc thù cao như dịch vụ công, nổi trội với khả năng hỗ trợ tiếng Việt.

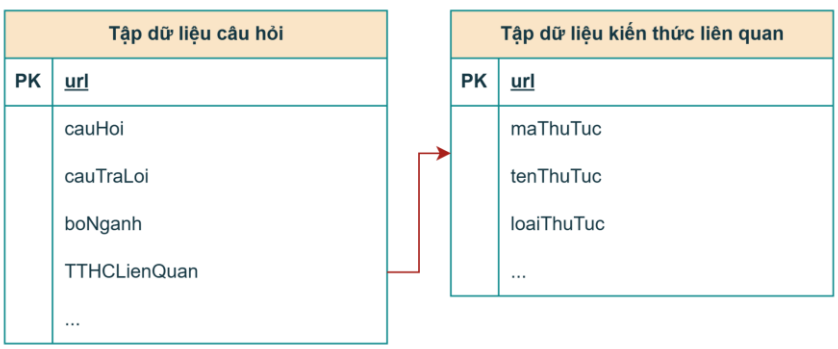
### *B. Tìm kiếm và thu thập dữ liệu*

Để có được một bộ dữ liệu chất lượng cho quá trình đánh giá hiện tượng ảo giác của các mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công, việc tìm kiếm và thu thập dữ liệu đóng vai trò nền tảng quan trọng trong toàn bộ quy trình xây dựng bộ dữ liệu ảo giác. Sau quá trình tìm hiểu và nghiên cứu các nguồn dữ liệu khác nhau, chúng tôi đã quyết định lựa chọn Cổng Dịch vụ công Quốc gia, vốn cung cấp thông tin chính thống và được cập nhật thường xuyên về các thủ tục hành chính (TTHC) trên cả nước, làm nguồn dữ liệu ban đầu để thu thập. Cổng Dịch vụ công Quốc gia được chủ quản bởi cơ quan Văn phòng Chính phủ nước Cộng hòa Xã hội chủ nghĩa Việt Nam, giúp kết nối và cung cấp thông tin về dịch vụ công mọi lúc, mọi nơi đến với mọi công dân.

Tiếp đến với công đoạn lựa chọn nguồn dữ liệu, chúng tôi tiến hành thu thập những câu hỏi thường gặp nhất trong ngữ cảnh dịch vụ công. Mỗi câu hỏi không chỉ có câu trả lời tương ứng mà còn cung cấp thêm những thông tin tham chiếu cần thiết về những thủ tục hành chính liên quan dùng để hỗ trợ trả lời câu hỏi đó, bộ/ngành tương ứng với câu hỏi,... Sau khi đã thu thập được bộ dữ liệu câu hỏi, chúng tôi dựa vào những đường dẫn đến thủ tục hành chính ở từng câu hỏi để tiếp tục thu thập những thông tin chi tiết về các thủ tục hành chính đó. Hình 2 thể hiện mối quan hệ giữa các tập dữ liệu trong bộ dữ liệu sau khi thu thập. Ở đây, thông tin về các thủ tục hành chính liên quan ở mỗi câu hỏi đóng vai trò như kiến thức liên quan đến câu hỏi đó, và kiến thức liên quan này cũng là một thành phần rất quan trọng trong bộ dữ liệu nhằm có thể sinh ra được ảo giác và đánh giá hiện tượng ảo giác trong các mô hình ngôn ngữ lớn về sau.




Hình 1: Quy trình xây dựng bộ dữ liệu ảo giác trong ngữ cảnh dịch vụ công




Hình 2: Lược đồ quan hệ các tập dữ liệu có trong quy trình

Dữ liệu thô sau khi thu thập từ Cổng Dịch vụ công Quốc gia bao gồm hai tập dữ liệu sau:

- Tập dữ liệu câu hỏi, chứa câu hỏi và câu trả lời và tham chiếu đến các thủ tục hành chính (kiến thức liên quan đến câu hỏi đó), gồm tổng cộng 9452 câu hỏi.
- Tập dữ liệu kiến thức liên quan, chứa thủ tục hành chính và các trường thông tin chi tiết của thủ tục hành chính đó, tổng cộng gồm 2695 thủ tục hành chính.

### *C. Tiền xử lý dữ liệu*

Trước khi sinh ra bộ dữ liệu ảo giác nhằm hỗ trợ đánh giá hiện tượng ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công, tiền xử lý dữ liệu cũng là một bước trung gian quan trọng trong quy trình, nhằm cung cấp một đầu vào sạch và tối ưu, giúp nâng cao chất lượng của bộ dữ liệu tiêu chuẩn.

Sở dĩ cần phải có thêm giai đoạn tiền xử lý dữ liệu trong quy trình là bởi vì các câu hỏi phổ biến và câu trả lời tương ứng trên Cổng Dịch vụ công Quốc gia vẫn còn mắc nhiều lỗi về thiếu dữ liệu, trùng lặp dữ liệu, sai chính tả hoặc ngữ nghĩa,… Tương tự như các giai đoạn tiền xử lý văn bản trong quy trình khoa học dữ liệu, từ bộ dữ liệu thô đã có từ trước đó, chúng tôi đã tiến hành xóa bỏ những dữ liệu bị trùng lặp. Dựa vào độ tương đồng Cosine giữa các cặp câu hỏi và câu trả lời, chúng tôi đặt ngưỡng cho phép là 0.95, nghĩa là trên mức này thì các câu được xem là trùng lặp và bị loại bỏ. Bên cạnh đó, với những câu bị trùng lặp với nhau thì ưu tiên giữ lại những câu có đầy đủ bộ/ngành và số thủ tục hành

chính lớn hơn, giúp dữ liệu được đảm bảo toàn vẹn hơn. Đối với dữ liệu bị thiếu, chúng tôi chỉ giữ lại những câu hỏi có đầy đủ câu trả lời, bộ/ngành và kiến thức liên quan. Ngoài ra, một tập dữ liệu các câu hỏi và câu trả lời bị thiếu bộ/ngành tương ứng cũng được giữ lại, phòng trường hợp quá ít dữ liệu để đánh giá về sau. Để bộ dữ liệu được đảm bảo tính chính xác hơn nữa, chúng tôi cũng đã thực hiện quan sát và kiểm tra chính tả thủ công ở một số những cặp câu hỏi và câu trả lời có trong bộ dữ liệu.

Trải qua các bước tiền xử lý dữ liệu cùng với kiểm tra chính tả, bộ dữ liệu câu hỏi còn lại 3717 mẫu dữ liệu và bộ dữ liệu thủ tục hành chính liên quan còn lại 1820 mẫu dữ liệu. Ngoài ra, các câu hỏi trong bộ dữ liệu còn trải dài trên 14 bộ/ngành khác nhau. Từ đó, những phân loại bộ/ngành có thể hỗ trợ nhằm đưa ra những phân tích sâu và chi tiết hơn về hiện tượng ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công.

### *D. Sinh dữ liệu ảo giác*

Đây là phần quan trọng nhất trong quy trình xây dựng bộ dữ liệu ảo giác trong ngữ cảnh dịch vụ công. Mục đích của giai đoạn này là nhằm có thể tạo ra các câu trả lời ảo giác tự động dựa trên các câu hỏi, câu trả lời đúng và kiến thức liên quan đã được thu thập và tiền xử lý trước đó.

1. Lựa chọn model và siêu tham số

Trước tiên, chúng tôi đã khảo sát và lựa chọn ra những mô hình ngôn ngữ lớn phù hợp nhất để thực hiện quá trình sinh dữ liệu ảo giác. Các mô hình trí tuệ nhân tạo hiện nay cho thấy hiệu suất cao trong tác vụ sinh văn bản bao gồm: ChatGPT, DeepSeek, Claude, Gemini. Tuy nhiên, tiêu chí của bộ dữ liệu không phụ thuộc vào việc lựa chọn mô hình ngôn ngữ lớn mà chỉ cần đảm bảo rằng dữ liệu được sinh ra thực sự là dữ liệu ảo giác. Chính vì thế, chúng tôi quyết định lựa chọn GPT-4o-mini làm mô hình chính để thực hiện tác vụ trong công đoạn này. GPT-4o-mini là một mô hình trí tuệ nhân tạo được phát triển bởi OpenAI, được tối ưu hóa để hoạt động hiệu quả mà vẫn đảm bảo được những hạn chế về tài nguyên với chi phí API phải chăng. Việc lựa chọn mô hình GPT-4o-mini để sinh ra dữ liệu ảo giác là đủ đáp ứng điều kiện cho những tác vụ văn bản cơ bản mà không yêu cầu suy luận nâng cao trong ngữ cảnh dịch vụ công.

Kế tiếp việc lựa chọn mô hình, những siêu tham số quan trọng cũng cần được xem xét đến khi nhằm có thể tối ưu đầu ra của mô hình ngôn ngữ lớn trong việc sinh câu trả lời ảo giác. Ở đây, chúng tôi lựa chọn hai siêu tham số tác động lớn nhất đến mức độ ảo giác của câu trả lời, từ đó ảnh hưởng đáng kể đến chất lượng của bộ dữ liệu, đó là temperature và top_p. Temperature điều chỉnh mức độ ngẫu nhiên trong việc chọn từ tiếp theo trong văn bản trong khi top_p sẽ điều chỉnh sự đa dạng của văn bản bằng cách chọn từ trong một tập hợp sao cho tổng xác suất của các từ này đạt một ngưỡng nhất định. Chúng tôi lựa chọn temperature=1 và top_p=1 cho giai đoạn sinh ảo giác, với mong đợi rằng kết quả đầu ra vừa đảm bảo được tính ảo giác đồng thời khiến cho câu trả lời trở nên khó nhận diện ảo giác hơn. Ngoài ra, chúng tôi cũng lựa chọn thêm max_output_tokens=512, nhằm đảm bảo cho các câu trả lời ảo giác không quá ngắn so với câu trả lời đúng của cùng một câu hỏi (số lượng tokens đủ lớn).

2. Thiết kế truy vấn và phân loại (pattern)

Một yếu tố then chốt để sinh dữ liệu ảo giác hiệu quả là thiết kế truy vấn (prompt) phù hợp để hướng dẫn mô hình ngôn ngữ lớn (LLM) như ChatGPT tạo ra phản hồi sai lệch nhưng vẫn hợp ngữ nghĩa và bối cảnh. Trong nghiên cứu này, chúng tôi xây dựng mẫu truy vấn gồm bốn thành phần chính: mô tả mục tiêu, phân loại ảo giác và thể hiện ảo giác và dữ liệu đầu vào. Cấu trúc truy vấn này đóng vai trò định hướng cho mô hình hiểu được yêu cầu sinh phản hồi ảo giác trong ngữ cảnh dịch vụ công. Cụ thể hơn, khi áp dụng cấu trúc truy vấn này vào trong mô hình GPT-4o-mini, các phần mô tả mục tiêu và phân loại ảo giác sẽ được đưa vào phần truy vấn hệ thống (system prompt). Mặt khác, phần thể hiện ảo giác và dữ liệu sẽ được đưa vào phần truy vấn người dùng (user prompt) như Bảng 2.

Đối với phần truy vấn hệ thống: **Mô tả mục tiêu** giúp xác định vai trò của mô hình trong quá trình sinh dữ liệu, làm rõ rằng phản hồi cần được tạo ra phải "có vẻ hợp lý" nhưng thực chất "không được hỗ trợ bởi tri thức đã cho". Phần mô tả mục tiêu được đưa vào truy vấn hệ thống, đảm bảo sự rõ ràng, cô đọng, không lan man, giúp mô hình ngôn ngữ lớn có thể hiểu được những tác vụ cần xử lý, từ đó có thể sinh ra dữ liệu ảo giác một cách chính xác và tin cậy. **Phân loại ảo giác (pattern)** cũng là một phần không thể bỏ qua trong thiết kế truy vấn, cho phép kiểm soát chất lượng và đa dạng của các phản hồi sinh ra. Dựa trên phân tích đặc điểm ngôn ngữ và hiện tượng ảo giác phổ biến trong mô hình trí tuệ nhân tạo, chúng tôi xác định ra bốn phân loại ảo giác (pattern) được ngẫu nhiên chia ra cho các câu hỏi như sau: (i) Hiểu sai ngữ cảnh hoặc mục đích của câu hỏi; (ii) Mâu thuẫn giữa câu trả lời và kiến thức liên quan; (iii) Câu trả lời quá chung chung hoặc quá cụ thể; (iv) Trả lời lập luận sai từ kiến thức liên quan. Việc phân chia pattern đồng đều cho mỗi câu hỏi như trên cho phép đưa ra những phân tích sâu hơn về hiện tượng ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công trên nhiều khía cạnh và lĩnh vực khác nhau.

Đối với phần truy vấn người dùng: **Thể hiện ảo giác** đóng vai trò như huấn luyện ngữ cảnh giúp mô hình học được cách tạo ra phản hồi ảo giác đúng yêu cầu (In-context learning). Sau khi áp dụng những chiến lược không ví dụ (zero-shot) đến nhiều ví dụ (few-shot), chúng tôi nhận thấy rằng việc việc đưa những mô tả mục tiêu mà không cần đưa câu trả lời mẫu cụ thể (zero-shot) vẫn cho thấy hiệu suất ổn định trong việc sinh ra câu trả lời ảo giác. **Dữ liệu (đã được tiền xử lý trước đó)** được đưa vào truy vấn như một phần quan trọng nhất để giúp cho mô hình ngôn ngữ lớn có thể sinh ra những phản hồi ảo giác.

3. Kết quả và chú thích thủ công

Quá trình sinh ảo giác tự động được thực hiện bằng mô hình GPT-4o-mini trên bộ dữ liệu đã được tiền xử lý, sau khi đã chuẩn bị những thành phần quan trọng như siêu tham số, truy vấn và phân loại ảo giác (pattern). Bất cứ tác vụ sinh văn bản tự động nào cũng vậy, việc tận dụng mô hình ngôn ngữ lớn cũng sẽ tiềm ẩn những rủi ro, đó là không thể chắc chắn được dữ liệu (câu trả lời) đó có thật sự là ảo giác hay không và cũng không thể chắc chắn được dữ liệu (câu trả lời) được sinh ra đó đã đúng định dạng mong muốn hay chưa. Chính vì thế, chúng tôi đã tiến hành chú thích dữ liệu ảo giác (Human Annotation).

Cụ thể, sau khi sinh ra các phản hồi ảo giác từ mô hình GPT-4o-mini, dữ liệu các câu trả lời tương ứng thu được gồm có 3717 mẫu. Để có thể giúp cho quá trình chú thích dữ liệu được diễn ra một cách tin cậy và đảm bảo, nhóm nghiên cứu đã học hỏi thêm một số những kiến thức miền trong ngữ cảnh dịch vụ công, nhằm có thể kiểm tra dữ liệu ảo giác một cách hiệu quả và đúng đắn nhất. Chúng tôi đã quan sát bộ dữ liệu ảo giác sau khi được xây dựng và rút ra nhận xét rằng các câu trả lời được sinh ra là hoàn toàn ảo giác. Xét đến ngữ cảnh dịch vụ công, tính chuẩn xác và tin cậy về mặt pháp lý là vô cùng quan trọng, cho nên chỉ cần một chút sai sót về số liệu, ngữ nghĩa hay bối cảnh cũng có thể được xem là hiện tượng ảo giác trong mô hình ngôn ngữ lớn. Với tiêu chí đó, chúng tôi đã tiến hành chú thích dữ liệu ảo giác trên toàn bộ 3717 mẫu và đảm bảo những phản hồi này là những câu trả lời ảo giác thực sự.

| 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}<br><br>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}<br><br>#Câu hỏi#: {question}<br><br>#Câu trả lời đúng#: {right_answer}<br><br>#Câu trả lời ảo giác#: |

Bảng 2: Mẫu truy vấn cho quá trình sinh ảo giác

Bộ dữ liệu sau khi được xây dựng bao gồm 3717 mẫu ảo giác được tạo ra từ mô hình GPT-4o-mini. Với những nỗ lực không chỉ sử dụng mô hình ngôn ngữ lớn để sinh câu trả lời tự động mà còn tận dụng tài nguyên con người trong việc chú thích thủ công, chúng tôi tin rằng bộ dữ liệu ảo giác trong ngữ cảnh dịch vụ công này hoàn toàn có thể làm tiêu chuẩn để phát hiện ảo giác của những mô hình lớn trong ngữ cảnh dịch vụ công, nổi bật với khả năng hỗ trợ tiếng Việt.

## IV. THỰC NGHIỆM ĐÁNH GIÁ VÀ PHÂN TÍCH KẾT QUẢ

### *A. Phương pháp đánh giá*

Từ bộ dữ liệu ảo giác trong ngữ cảnh dịch vụ công đã xây dựng, chúng tôi thực hiện đánh giá hiệu quả phát hiện thông tin ảo giác của các mô hình ngôn ngữ lớn (LLM) thông qua hai nhóm mô hình: mã nguồn đóng và mã nguồn mở. Đối với mô hình mã nguồn đóng (hoặc mô hình quy mô lớn chỉ truy cập qua API), các đại diện được lựa chọn gồm: GPT-4o-mini, DeepSeek-v3-0324, Gemini-2.0-flash và Claude-3.5-Haiku. Với mô hình mã nguồn mở (7B), sau quá trình thử nghiệm chất lượng đầu ra, chúng tôi chọn các mô hình: Llama-3, Mistral-v0.3, Qwen-2.5, Vicuna-v1.5, WizardLM-2. Bên cạnh đó, để đánh giá khả năng thích ứng với tiếng Việt, hai mô hình đã được tinh chỉnh thêm gồm: Vistral và Qwen-Viet.

Để đảm bảo tính đồng nhất và công bằng trong đánh giá, chúng tôi sử dụng chung các siêu tham số quan trọng (temperature, top_p) và một mẫu truy vấn thống nhất (như ở Bảng 3). Mỗi truy vấn gồm một câu hỏi và một câu trả lời; nhiệm vụ của mô hình là xác định liệu câu trả lời có chứa thông tin ảo giác hay không. Với mỗi cặp câu hỏi, gồm một câu trả lời đúng và một câu trả lời chứa ảo giác, chúng tôi đánh giá độc lập để xây dựng ma trận nhầm lẫn (confusion matrix). Trong đó, pattern chỉ được đưa vào đối với câu trả lời ảo giác nhằm tham chiếu với giai đoạn sinh câu trả lời trước đó. Ngoài ra, chúng tôi thực hiện hai phương án đánh giá: (1) không cung cấp kiến thức liên quan để kiểm tra khả năng vốn có của mô hình, và (2) đưa thêm thông tin chuyên ngành vào truy vấn để kiểm tra khả năng nhận diện ảo giác trong điều kiện có hỗ trợ tri thức.

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

### *B. Kết quả và nhận xét*

Trong thiết lập không cung cấp kiến thức liên quan, 11 mô hình được đánh giá khả năng phát hiện thông tin ảo giác chỉ dựa vào câu hỏi và câu trả lời. Bảng 4 cho thấy rõ sự khác biệt giữa các mô hình, đặc biệt nổi bật là một số mô hình mã nguồn mở cho kết quả vượt trội hơn cả mô hình mã nguồn đóng. Cụ thể, Vicuna-v1.5 (mã nguồn mở) đạt precision gần tuyệt đối (99.68%) và F1-score cao nhất (66.86%), dù accuracy chỉ ở mức trung bình (50.59%). Qwen-Viet (mã nguồn mở, tinh chỉnh tiếng Việt) cũng đạt F1-score ấn tượng 62.12%, vượt qua cả GPT-4o-mini (64.68%) và DeepSeek (63.32%). Trong khi đó, Claude-3.5-Haiku lại thể hiện yếu nhất với F1-score chỉ 31.29%. Các mô hình như Llama-3, WizardLM-2 và Qwen-2.5 đều cho thấy sự cân bằng tốt giữa precision và recall, khẳng định tiềm năng của các mô hình mã nguồn mở trong phát hiện ảo giác, ngay cả khi không có kiến thức nền hỗ trợ.

| Mô hình | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| GPT-4o-mini | 51.72 | 88.40 | 50.99 | 64.68 |
| Gemini-2.0-flash | 51.29 | 69.65 | 50.94 | 58.85 |
| DeepSeek-V3-0324 | 50.26 | 85.88 | 50.15 | 63.32 |
| Claude-3.5-Haiku | 43.58 | 25.69 | 40.01 | 31.29 |
| Llama-3 | 53.07 | 76.11 | 52.10 | 61.86 |
| Mistral-v0.3 | 48.24 | 44.85 | 48.11 | 46.42 |
| Qwen-2.5 | 51.91 | 89.05 | 51.10 | 64.93 |
| Vicuna-v1.5 | 50.59 | **99.68** | 50.30 | **66.86** |
| WizardLM-2 | **57.61** | 44.79 | **60.24** | 51.38 |
| Vistral | 50.91 | 6.67 | 57.94 | 11.97 |
| Qwen-Viet | 53.81 | 75.76 | 52.65 | 62.12 |

Bảng 4: Bảng thống kê đánh giá các mô hình khi không truyền kiến thức liên quan (Đơn vị: %)

Bảng 5 trình bày độ chính xác trung bình của các mô hình ở Bảng 4 trong việc phát hiện thông tin ảo giác theo từng lĩnh vực bộ/ngành. Kết quả cho thấy accuracy dao động rất hẹp, quanh mức 50%, phản ánh tính đồng đều và cho thấy rằng các mô hình hiện tại khó khăn chung trong việc phân biệt thông tin ảo giác, bất kể lĩnh vực cụ thể. Dù có một vài lĩnh vực đạt độ chính xác cao hơn đôi chút như Bộ Lao động – Thương binh và Xã hội (52.12%), Bộ Quốc phòng (52.07%), hay Bộ Giao thông vận tải (51.66%), mức chênh lệch là không đáng kể. Ngược lại, các bộ như Bộ Tài chính (50.09%) hay Thanh tra Chính phủ (50.38%) vẫn nằm trong ngưỡng trung bình.

| 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 5: Accuracy theo từng bộ/ngành (Đơn vị: %)

Bảng 6 thể hiện độ chính xác trung bình của các mô hình ở Bảng 4 theo bốn phân loại ảo giác là khá đồng đều, dao động quanh mức 50%. Nhóm P-II đạt cao nhất (52.73%), trong khi P-IV thấp nhất (48.48%). Hai nhóm còn lại, P-I (51.59%) và P-III (51.97%) chênh lệch không đáng kể. Điều này cho thấy mô hình xử lý các phân loại ảo giác với mức độ hiệu quả tương đối gần nhau, phản ánh khó khăn mang tính phổ quát khi không có kiến thức hỗ trợ.

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

Bảng 6: Accuracy theo từng phân loại ảo giác (Đơn vị: %)

Kết quả trong Bảng 7 cho thấy khi truyền toàn bộ kiến thức liên quan vào prompt với số lượng mẫu dương bằng số lượng mẫu âm, các mô hình mã nguồn đóng có xu hướng đánh giá hầu hết các mẫu là ảo giác. Điều này thể hiện qua precision rất cao (gần tuyệt đối), nhưng recall và accuracy gần như giữ nguyên hoặc giảm nhẹ, cho thấy mô hình dự đoán ảo giác một cách bảo thủ và ít phân biệt hơn. Dù F1-score được cải thiện nhờ precision tăng, nhưng cách truyền kiến thức hiện tại vẫn chưa tối ưu, làm giảm độ cân bằng trong đánh giá. Vì vậy, bài toán truyền tri thức sao cho hiệu quả vẫn còn là một thách thức. Cần nghiên cứu thêm các hướng tiếp cận như tinh chỉnh mô hình (fine-tuning) hoặc kết hợp với truy xuất tri thức (RAG) để cải thiện khả năng hiểu ngữ cảnh và ra quyết định chính xác hơn.

| Mô hình | Có tri thức | Số lượng mẫu dương (âm) | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| GPT-4o-mini | Không | 3717 | **51.72** | 88.40 | **50.99** | 64.68 |
| | Có | 1000 | 50.00 | **99.40** | 50.00 | **66.53** |
| Gemini-2.0-flash | Không | 3717 | 51.29 | 69.65 | 50.94 | 58.85 |
| | Có | 1000 | 49.85 | 92.40 | 49.92 | 64.82 |
| DeepSeek-V3-0324 | Không | 3717 | 50.26 | 85.88 | 50.15 | 63.32 |
| | Có | 1000 | 49.90 | 98.10 | 49.95 | 66.19 |
| Claude-3.5-Haiku | Không | 3717 | 43.58 | 25.69 | 40.01 | 31.29 |
| | Có | 1000 | 45.55 | 33.50 | 44.14 | 38.09 |

Bảng 7: So sánh hiệu suất khi không truyền và khi truyền kiến thức liên quan trên những mô hình mã nguồn đóng (Đơn vị: %)

## V. THẢO LUẬN

Mặc dù nghiên cứu đã xây dựng được một quy trình tạo dữ liệu tương đối hoàn chỉnh và đóng góp một bộ dữ liệu có giá trị về hiện tượng ảo giác trong lĩnh vực dịch vụ công, song vẫn còn tồn tại một số điểm cần cải thiện trong các giai đoạn tiếp theo. Trước hết, do quy trình hiện tại chủ yếu được vận hành tự động với sự can thiệp giới hạn của con người - chỉ ở bước kiểm tra dữ liệu đầu vào và đánh giá đầu ra sau khi mô hình tạo sinh - nên nguy cơ dữ liệu bị lỗi thời là điều khó tránh khỏi. Trong bối cảnh thủ tục hành chính công luôn có khả năng thay đổi theo các quy định pháp luật

mới, việc sử dụng một bộ dữ liệu tĩnh dễ dẫn đến nguy cơ mô hình đánh giá không phản ánh đúng thực tiễn. Tuy nhiên, điểm mạnh của quy trình này nằm ở khả năng mở rộng và tự động hóa. Trong tương lai, việc tích hợp hệ thống thu thập và cập nhật dữ liệu định kỳ từ các nguồn chính thống như Cổng Dịch vụ công Quốc gia có thể giúp duy trì độ chính xác và kịp thời của tập dữ liệu, đồng thời phản ánh đầy đủ những thay đổi trong chính sách hoặc thủ tục hành chính. Bên cạnh đó, các mô hình ngôn ngữ lớn được sử dụng trong nghiên cứu hiện chưa được tinh chỉnh trên tập dữ liệu chuyên biệt về dịch vụ công. Điều này phần nào ảnh hưởng đến khả năng nắm bắt ngữ cảnh đặc thù, cấu trúc câu hỏi và thông tin hành chính, dẫn đến kết quả đánh giá có thể chưa phản ánh đúng bản chất của hiện tượng ảo giác. Việc mô hình chưa được đào tạo theo hướng đặc thù cũng có thể làm giảm độ chính xác trong nhận diện các nội dung sai lệch hay thiếu căn cứ.

Để cải thiện chất lượng nghiên cứu trong giai đoạn tiếp theo, chúng tôi đang xem xét hai hướng tiếp cận chính. Thứ nhất là xây dựng hệ thống cập nhật dữ liệu động, đảm bảo tính thời sự và phù hợp với các quy định hành chính mới nhất. Thứ hai là triển khai quá trình tinh chỉnh mô hình trên chính tập dữ liệu chuyên biệt đã xây dựng. Cách tiếp cận này có tiềm năng nâng cao đáng kể khả năng hiểu ngữ cảnh và nội dung hành chính công của mô hình, từ đó cải thiện hiệu quả phát hiện và đánh giá nội dung ảo giác. Những điều chỉnh này không chỉ giúp tăng cường độ chính xác của nghiên cứu mà còn góp phần nâng cao tính ứng dụng thực tiễn của các mô hình ngôn ngữ lớn trong môi trường dịch vụ công, nơi tính đúng đắn và nhất quán của thông tin là yếu tố then chốt.

## VI. KẾT LUẬN

Nhằm góp phần giải quyết bài toán ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công, bài báo đã đề xuất và xây dựng một bộ dữ liệu chuyên biệt tập trung vào hiện tượng ảo giác ở dịch vụ công. Quy trình bao gồm các bước thu thập dữ liệu, tiền xử lý, tạo câu trả lời ảo giác cùng với việc đánh giá thủ công nhằm đảm bảo chất lượng của bộ dữ liệu ở đầu vào lẫn đầu ra. Từ đó, chúng ta có thể dùng bộ dữ liệu này để đánh giá mức độ nhận biết ảo giác trên nhiều mô hình ngôn ngữ lớn khác nhau.

Kết quả thực nghiệm cho thấy khả năng nhận biết câu trả lời ảo giác ở dịch vụ công của các mô hình nếu chưa được truyền kiến thức liên quan còn hạn chế. Do đó, bộ dữ liệu bài báo cung cấp chứa cả tri thức chuyên ngành lẫn câu trả lời ảo giác sẽ có thể đóng vai trò quan trọng trong việc huấn luyện các mô hình ngôn ngữ lớn trong tương lai, qua đó giúp tăng độ chính xác và độ tin cậy của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công.

## VII. TÀI LIỆU THAM KHẢO

[1] N. Dziri, H. Rashkin, T. Linzen, and D. Reitter, “Evaluating attribution in dialogue systems: The BEGIN benchmark,” Transactions of the Association for Computational Linguistics, vol. 10, pp. 1066–1083, 2022.

[2] H. Rashkin, V. Nikolaev, M. Lamm, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter, “Measuring attribution in natural language generation models,” CoRR, vol. abs/2112.12870, 2021.

[3] N. Miao, Y. W. Teh, and T. Rainforth, “Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning,” ArXiv preprint, vol. abs/2308.00436, 2023.

[4] S. Chen, Y. Zhao, J. Zhang, I.-C. Chern, S. Gao, P. Liu, and J. He, “Felm: Benchmarking factuality evaluation of large language models,” ArXiv preprint, vol. abs/2310.00741, 2023.

[5] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “Halueval: A large-scale hallucination evaluation benchmark for large language models,” CoRR, vol. abs/2305.11747, 2023.

[6] DVC AI, “Dvc ai: Hỗ trợ dịch vụ hành chính công.”

# VIETPS-HALLU: A DATASET FOR HALLUCINATION DETECTION IN LARGE LANGUAGE MODELS WITHIN THE PUBLIC SERVICE CONTEXT

**Dinh Bao Bui, Tien Nhat Nguyen, Huy Tien Nguyen, Tung Le**

***Abstract*** *— Large Language Models (LLMs), such as ChatGPT, are increasingly being applied across various domains, including public administration, due to their ability to process and generate natural language. However, one major challenge in deploying LLMs in real-world settings is the phenomenon of "hallucination", when the model produces inaccurate, misleading, or unsubstantiated information. In order to contribute to solving this problem in the context of public services, this paper introduces VietPS-Hallu (Vietnamese Public Service Hallucination Dataset for LLMs) - a specialized dataset to evaluate the hallucination detection ability of large language models in the context of public services through building a complete framework from data preparation to performance evaluation of different large language models on that dataset. We provide a valuable experimental dataset for improving the ability to recognize and reduce hallucinations in large language models, thereby contributing to building and improving the reliability of large language models in administrative tasks, where consistency and correctness in terms of output information are required.*

***Key words*** *— Large Language Models (LLMs), public services, hallucination detection, benchmark dataset.*