TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHÓA LUẬN TỐT NGHIỆP

MÔ HÌNH PHÁT HIỆN ẢO GIÁC CỦA MÔ HÌNH NGÔN NGỮ LỚN TRONG NGỮ CẢNH DỊCH VỤ CÔNG

1 THÔNG TIN CHUNG

Giảng viên hướng dẫn:

- TS. Nguyễn Tiến Huy (Khoa Công nghệ Thông tin)
- ThS. Nguyễn Trần Duy Minh (Khoa Công nghệ Thông tin)

Nhóm sinh viên thực hiện:

1. Nguyễn Tiến Nhật (MSSV: 21120108)
2. Bùi Đình Bảo (MSSV: 21120201)

Loại đề tài: Ứng dụng

Thời gian thực hiện: Từ *01/2025* đến *07/2025*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu tổng quát

Mô hình ngôn ngữ lớn (Large Language Model - LLM) là các mô hình học sâu rất lớn, được đào tạo trước (pre-trained) dựa trên một lượng dữ liệu khổng lồ. Một trong những thành công quan trọng nhất của LLM là nó có khả năng hiểu và sinh ngôn ngữ tự nhiên như con người, mô hình ngôn ngữ lớn có thể thực hiện các tác vụ hoàn toàn khác nhau, ví dụ như trả lời câu hỏi, tóm tắt tài liệu, dịch ngôn ngữ và hoàn thành câu. Ngày nay, mô hình ngôn ngữ lớn đã và đang được ứng dụng rộng rãi trên nhiều lĩnh vực khác nhau như giáo dục, y tế, thương mại, truyền thông, quảng cáo,... Xét về khía cạnh dịch vụ công, mô hình ngôn ngữ lớn cũng cho thấy được tiềm năng vì nó có thể tạo ra xử lý và tạo sinh văn bản, từ đó có thể hỗ trợ được người dân nhanh chóng hơn trong các vấn đề về giấy tờ, hồ sơ liên quan, các bước thực hiện thủ tục hành chính.

Một trong những hạn chế lớn nhất của mô hình ngôn ngữ lớn đó là nó có thể tạo ra văn bản không chính xác, vô nghĩa hoặc không dựa trên dữ liệu thực tế. Hiện tượng này còn được gọi là ảo giác (hallucination) của mô hình ngôn ngữ lớn. Nguyên nhân của vấn đề này xuất phát từ nhiều tác động khác nhau, có thể là từ dữ liệu được đưa vào không đúng đắn hay là từ quá trình huấn luyện mô hình không thể kiểm soát được. Quay trở lại với ngữ cảnh dịch vụ công thì hiện tượng ảo giác này lại tạo ra nhiều thách thức hơn nữa khi tính chính xác và độ tin cậy đầu ra (output) của mô hình là vô cùng quan trọng.

Chính vì thế, sau quá trình tìm hiểu, nhóm em xây dựng một mô hình dùng để phát hiện ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công. Nền tảng phía sau mô hình này chính là một khung phương pháp hoàn chỉnh từ bước chuẩn bị dữ liệu cho đến khi đánh giá hiệu suất của các mô hình ngôn ngữ lớn khác nhau. Ý tưởng cho phương pháp này dựa trên kỹ thuật tạo lời nhắc (prompt engineering) từ đó ta sẽ tinh chỉnh và tạo ra output phù hợp cho tác vụ cần xử lý.

Những kết quả đạt được từ mô hình phát hiện ảo giác trên dự kiến sẽ bao gồm một bộ dữ liệu chất lượng cao, đảm bảo được khả năng đánh giá khách quan về hiệu suất đầu ra của các mô hình ngôn ngữ lớn, từ đó góp phần giúp cho quá trình tinh chỉnh mô hình đạt được hiệu quả tốt nhất. Đặc biệt, ngôn ngữ của bộ dữ liệu sẽ hoàn toàn là tiếng Việt, điều này giúp cho tiêu chuẩn đánh giá trở nên chuyên dụng và thiết thực hơn trong ngữ cảnh dịch vụ công ở nước ta. Áp dụng vào thực tiễn, nhóm em hi vọng mô hình có thể giúp khắc phục được hiện tượng ảo giác của mô hình ngôn ngữ lớn, với mục đích khiến cho câu trả lời trở nên minh bạch và đúng đắn hơn. Dựa vào đó, các trợ lý ảo trí tuệ nhân tạo dịch vụ công có thể được xây dựng để hỗ trợ người dân kịp thời và tốt hơn, nhằm nâng cao đời sống cộng đồng về lâu dài.

2.2 Mục tiêu nghiên cứu

Đứng trước sự bùng nổ của các mô hình ngôn ngữ lớn, việc tận dụng sức mạnh của nó là một điều cần thiết nhằm tự động hóa các tác vụ, mang đến khả năng phản hồi tức thời cho những khó khăn của con người. Tuy nhiên, việc áp dụng công nghệ này vẫn đi kèm theo nhiều thách thức đặt ra, một trong số đó là hiện tượng ảo giác của mô hình tạo sinh ngôn ngữ, khiến cho câu trả lời đầu ra trở nên sai lệch, vô lý hoặc xa rời thực tế. Điều này là vấn đề rất nhạy cảm trong ngữ cảnh dịch vụ công khi những vấn đề pháp lý và chứng thực thông tin là không thể bỏ qua. Đồng thời, hầu hết các mô hình ngôn ngữ lớn hiện chỉ đang phục vụ cho tiếng Anh là chủ yếu, trong khi đó hỗ trợ cho tiếng Việt cụ thể hơn là các mô hình huấn luyện, đánh giá và các bộ dữ liệu tiêu chuẩn cho tiếng Việt vẫn còn đang rất hiếm trong hàng loạt các lĩnh vực nói chung, chứ không riêng gì ngữ cảnh dịch vụ công. Vì vậy, mô hình phát hiện ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công được đề xuất nhằm khắc phục những nhược điểm trên.

Với đề tài nghiên cứu này, nhóm em mong muốn đem lại:

- Một quy trình hoàn chỉnh dùng để đánh giá ảo giác của mô hình ngôn ngữ

lớn trong ngữ cảnh dịch vụ công, hỗ trợ tiếng Việt.

- Một bộ dữ liệu phát hiện ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công, từ đó giúp tinh chỉnh mô hình đồng thời có thể làm một tiêu chuẩn (benchmark) phát hiện ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công, hỗ trợ tiếng Việt.
- Một bảng thống kê so sánh hiệu suất của các mô hình ngôn ngữ lớn phổ biến hiện nay trong ngữ cảnh dịch vụ công, cụ thể hơn là các chủ đề trong ngữ cảnh dịch vụ công, từ đó đưa ra mô hình phù hợp nhất cho tác vụ này.

Từ những kết quả trên, những ý nghĩa thực tiễn và ảnh hưởng tích cực của mô hình không chỉ dừng lại ở việc có thể khắc phục được hiện tượng ảo giác của mô hình ngôn ngữ lớn trong bối cảnh dịch vụ công, mà còn hứa hẹn có thể đề ra phương pháp chung để cải thiện hiệu suất của mô hình tạo sinh ngôn ngữ cho những lĩnh vực khác. Đi đôi với việc giảm thiểu hiện tượng ảo giác của mô hình ngôn ngữ lớn, ta còn có thể tinh chỉnh những mô hình này từ đó tạo ra những trợ lý ảo trí tuệ nhân tạo hỗ trợ con người tốt hơn về những vấn đề xung quanh dịch vụ công.

2.3 Phạm vi của đề tài

Đối với phạm vi của đề tài, các yếu tố liên quan là:

- Đối tượng nghiên cứu của đề tài bao gồm các mô hình ngôn ngữ lớn (LLM) như GPT4o-mini, DeepSeek, Gemini hoặc các mô hình tương tự, đặc biệt là khi chúng được áp dụng vào các hệ thống dịch vụ công.
- Thực thể liên quan bao gồm:
  - **Người dùng dịch vụ công**: Là những người trực tiếp tương tác với các hệ thống AI trong các dịch vụ công như y tế, hành chính, bảo hiểm xã hội,...

  - **Cơ quan nhà nước và các dịch vụ công**: Các tổ chức, cơ quan nhà nước triển khai các mô hình ngôn ngữ lớn để cung cấp các dịch vụ tự động hoặc trợ giúp thông qua các chatbot, trợ lý ảo.
  - **Mô hình ngôn ngữ lớn (LLM)**: Các mô hình AI được sử dụng để xử lý, tạo ra phản hồi, và hỗ trợ người dùng trong các dịch vụ công, đặc biệt là các mô hình có khả năng gây ra hiện tượng ảo giác.

- Về tập dữ liệu, có 2 tập dữ liệu chính là:
  - **Tập dữ liệu lấy từ trang web chính thống**: hơn 9000 mẫu gồm câu hỏi thường gặp, câu trả lời, bộ ngành và thủ tục hành chính được lấy từ trang 'dichvucong.gov.vn' của chính phủ Việt Nam để đảm bảo tính đúng đắn của tập dữ liệu. Sau khi trải qua bước tiền xử lý dữ liệu thì tập dữ liệu sau cùng còn gần 7000 mẫu.
  - **Tập dữ liệu câu trả lời ảo giác**: gồm gần 7000 mẫu câu trả lời ảo giác, được sinh ra từ mô hình ngôn ngữ lớn (LLM) dựa trên tập dữ liệu câu hỏi thường gặp.

2.4 Cách tiếp cận dự kiến

2.4.1 Các công trình liên quan

Các nghiên cứu trước đây đã từng đề cập đến vấn đề phát hiện ảo giác của các mô hình ngôn ngữ lớn nói chung bao gồm các tiêu chuẩn (benchmark) hay cụ thể hơn là các bộ dữ liệu dùng để phục vụ như một nền tảng giúp hiển thị được các ảo giác trong mô hình ngôn ngữ lớn như:

- **BEGIN**, một bộ dữ liệu phân loại các câu trả lời của hệ thống đối thoại thành 3 loại là: fully attributable (tạm dịch là 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].

- **AIS**, một bộ dữ liệu đánh giá liệu các tài liệu nguồn có hỗ trợ đầu ra của các mô hình tạo sinh văn bản hay không [2].
- **SelfCheckGPT-Wikibio**, một bộ dữ liệu ở cấp độ câu 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].
- **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].
- **HaluEval**, một bộ dữ liệu kết hợp tạo sinh tự động với chú thích của con người để đánh giá khả năng nhận diện ảo giác của các mô hình ngôn ngữ lớn [5].

Các nghiên cứu trên đã chứng minh được tính hiệu quả khi có thể đánh giá được các mô hình ngôn ngữ lớn khác nhau trong việc tạo ra ảo giác trên một số những lĩnh vực nhất định hay trải dài trên nhiều khía cạnh khác nhau.

Xét đến ngữ cảnh dịch vụ công, mô hình ngôn ngữ lớn cũng đã được áp dụng trong thực tế, **trợ lý ảo dịch vụ công** đã được xây dựng để hỗ trợ trên 15 chủ đề thiết yếu nhất mà người dân gặp phải khó khăn [6].

2.4.2 Điểm hạn chế tương ứng

Dù đã có nhiều tiêu chuẩn (bộ dữ liệu) trước đó nói về việc phát hiện ảo giác của mô hình ngôn ngữ lớn, các nghiên cứu chỉ mới dừng lại ở việc đánh giá trên ngữ cảnh tổng thể, chung chung trên toàn bộ mô hình ngôn ngữ lớn. Các lĩnh vực liên quan như y tế, giáo dục thường được chú trọng nhiều hơn trong khi lĩnh vực dịch vụ công vẫn còn khá sơ khai.

Các bộ dữ liệu trước đây chỉ được tập trung xây dựng cho tiếng Anh và một số ngoại ngữ phổ biến khác. Chính vì thế, việc sử dụng những tiêu chuẩn đánh giá này chỉ tối ưu nhất khi áp dụng cho tiếng Anh mà thôi.

Ở Việt Nam, mô hình ngôn ngữ lớn vẫn đang trong quá trình phát triển và được ứng dụng rộng rãi trên nhiều lĩnh vực khác nhau. Trong bối cảnh dịch vụ công, mô hình ngôn ngữ lớn cũng được áp dụng nhưng vẫn chưa được phổ biến với người dân. Các trợ lý ảo dịch vụ công này chỉ việc đưa ra câu trả lời dựa trên tài liệu thủ tục hành chính nhất định mà không hề quan tâm đến khả năng tạo sinh ảo giác của nó.

Chính vì thế, mô hình phát hiện ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công được đề xuất để đánh giá chất lượng của các trợ lý ảo dịch vụ công, đồng thời có thể lấy đó làm cơ sở để cải tiến mô hình trong tương lai.

2.4.3 Phương pháp chính

Quá trình nghiên cứu bắt đầu bằng việc thu thập dữ liệu từ Cổng dịch vụ công quốc gia Việt Nam, bao gồm các câu hỏi thường gặp về dịch vụ công trong nhiều lĩnh vực khác nhau. Các câu trả lời được tham chiếu từ các tài liệu thủ tục hành chính chính thống, giúp đảm bảo tính chính xác. Bộ dữ liệu thô được xây dựng bao gồm câu hỏi, lĩnh vực liên quan, câu trả lời chính xác và tài liệu tham chiếu để làm cơ sở cho việc xử lý tiếp theo.

Dữ liệu sau khi thu thập được tiến hành tiền xử lý, bao gồm xử lý chuỗi để chuẩn hóa văn bản, loại bỏ các trùng lặp về câu hỏi và câu trả lời, đồng thời bổ sung thông tin cho các câu hỏi thiếu thủ tục hành chính. Sau đó, dựa theo bài báo HaluEval, một khung prompt được thiết kế theo cấu trúc in-context learning nhằm tạo ra các ảo giác có kiểm soát, giúp đánh giá khả năng tạo ảo giác của mô hình ngôn ngữ lớn.

Cuối cùng, API của ChatGPT được sử dụng để tạo các câu trả lời có ảo giác theo khung prompt đã xây dựng, từ đó hình thành bộ dữ liệu tiêu chuẩn để đánh giá mô hình. Quá trình đánh giá giúp phát hiện và đo lường mức độ ảo giác của các mô hình ngôn ngữ lớn khác nhau, so sánh hiệu suất và xác định mô hình phù hợp nhất trong ngữ cảnh dịch vụ công.

2.5 Kết quả dự kiến

Qua những động lực nghiên cứu và phương pháp đã đặt ra, kết quả mà nhóm em mong muốn đạt được là:

- Một quy trình phù hợp nhất cho tác vụ đánh giá ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công.
- Một bộ dữ liệu tiếng Việt làm tiêu chuẩn tương ứng.
- Một bảng thống kê so sánh hiệu suất của các mô hình ngôn ngữ lớn, dựa trên bộ dữ liệu đã đưa ra.

2.6 Kế hoạch thực hiện

Giai đoạn	Công việc	Người thực hiện
01/01/2025 - 20/01/2024	Tìm hiểu, nghiên cứu các công trình liên quan đến ảo giác của mô hình ngôn ngữ lớn	Nhật, Bảo
20/01/2025 - 10/02/2025	Tìm hiểu nghiên cứu các công trình liên quan đến ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công	Nhật, Bảo
10/02/2025 - 20/02/2025	Thu thập dữ liệu chính thống về khía cạnh dịch vụ công	Nhật, Bảo
20/02/2025 - 01/03/2025	Tiền xử lý và tạo khung prompt dựa theo các công trình liên quan	Nhật, Bảo
01/03/2025 - 15/03/2025	Tạo ra các câu trả lời ảo giác tương ứng với khung prompt và bộ dữ liệu	Nhật, Bảo
15/03/2025 - 01/04/2025	Tiến hành đánh giá bộ dữ liệu thu được và tiền xử lý bổ sung	Nhật, Bảo
01/04/2025 - 01/05/2025	Áp dụng cho các mô hình ngôn ngữ lớn trong việc phát hiện ra ảo giác và thống kê so sánh	Nhật, Bảo
01/05/2025 - 01/06/2025	Tiến hành tối ưu bộ dữ liệu tiêu chuẩn, khiến cho mô hình ngôn ngữ lớn càng khó nhận diện ảo giác hơn trong ngữ cảnh dịch vụ công	Nhật, Bảo
01/06/2025 - 01/07/2025	Viết cuốn luận khóa luận tốt nghiệp và báo cáo	Nhật, Bảo

Bảng 1: Kế hoạch thực hiện

Tài liệu

[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."

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày 28 tháng 3 năm 2025
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)