6

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

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(HƯỚNG NGHIÊN CỨU)

Tên đề tài: Mô Hình Phát Hiện Ảo Giác Của Mô Hình Ngôn Ngữ Lớn Trong Ngữ Cảnh Dịch Vụ Công

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

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

1. **Chủ đề và ý tưởng nghiên cứu:**

Hiện tượng ảo giác (hallucination) của các mô hình ngôn ngữ lớn (LLMs) là vấn đề "nóng", đặc biệt trong ngữ cảnh dịch vụ công, nơi độ chính xác và khả năng kiểm chứng thông tin rất quan trọng. Nhóm tập trung xây dựng bộ dữ liệu tiếng Việt chuyên biệt cho phát hiện ảo giác trong dịch vụ công—một lĩnh vực vẫn còn rất thiếu tập dữ liệu đánh giá.

2. **Phương pháp nghiên cứu:**

Dựa trên phương pháp HaluEval được đề xuất ở tiếng Anh, đề tài đề xuất pipeline ba giai đoạn: (i) thu thập dữ liệu, (ii) tiền xử lý & kiểm tra thủ công, (iii) sinh và chú thích phản hồi ảo giác. Bộ dữ liệu được đánh giá trên các mô hình ngôn ngữ lớn nhằm đánh giá mức độ ảo giác của các mô hình cho ngữ cảnh dịch vụ công tiếng Việt.

3. **Đóng góp Khoa học và thực tiễn:**

Khóa luận xây dựng tập dữ liệu ảo giác với 3717 mẫu dữ liệu thuộc nhiều chủ đề được thu thập từ cổng dịch vụ công quốc gia. Dữ liệu cũng chia ra 4 dạng ảo giác phổ biến nhằm đánh giá chi tiết các loại ảo giác mà mô hình gặp phải. Bộ dữ liệu và quy trình tạo dữ liệu này là nền tảng để phát triển trợ lý ảo hành chính an toàn, giảm rủi ro thông tin sai lệch cho người dân.

4. **Quá trình thực hiện và quản lý dự án:**

Nhóm sinh viên làm việc nghiêm túc và có trách nhiệm.

5. **Báo cáo viết:**

Báo cáo được chia thành 5 chương dài 72 trang với bố cục hợp lý, và mạch lạc.

6. **Trình bày trước hội đồng:**

Tốt

7. **Công bố khoa học/ ứng dụng thực tế:**

Không có

Đánh giá xếp loại: Đạt yêu cầu của một khóa luận đại học

TP.HCM, ngày …. tháng …. năm …….

Giảng viên hướng dẫn

(*Ký và ghi rõ họ tên*)

Nguyễn Tiến Huy **Nguyễn Trần Duy Minh**

b

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

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(HƯỚNG NGHIÊN CỨU)

Tên đề tài : 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

Giảng viên phản biện: TS. Nguyễn Ngọc Thảo

1. Chủ đề và ý tưởng nghiên cứu:

Khóa luận tập trung vào hiện tượng "ảo giác" (hallucination) của các mô hình ngôn ngữ lớn (LLM) – một vấn đề quan trọng nhưng còn ít được nghiên cứu trong ngữ cảnh tiếng Việt, đặc biệt là dịch vụ công. Nhóm tác giả đề xuất xây dựng một mô hình phát hiện ảo giác dựa trên kỹ thuật tạo dữ liệu có giám sát kết hợp với đánh giá nhiều mô hình ngôn ngữ lớn hiện nay. Ý tưởng mang tính thời sự, có giá trị ứng dụng trong bối cảnh Việt Nam đang thúc đẩy chuyển đổi số.

2. Phương pháp nghiên cứu:

Khóa luận xây dựng một quy trình phát hiện ảo giác của mô hình ngôn ngữ lớn trong ngữ cảnh dịch vụ công thông qua năm bước chính. Đầu tiên, nhóm thu thập hơn dữ liệu từ Cổng Dịch vụ công Quốc gia, gồm câu hỏi, câu trả lời và thủ tục hành chính liên quan. Sau đó, dữ liệu được tiền xử lý để loại bỏ trùng lặp, chuẩn hóa văn bản và đảm bảo tính đầy đủ thông tin. Tiếp theo, nhóm thiết kế hệ thống prompt theo bốn loại lỗi phổ biến (hiểu sai ngữ cảnh, mâu thuẫn tri thức, quá chung/chi tiết, suy luận sai) để tạo ra các phản hồi ảo giác từ mô hình LLM. Các phản hồi được gán nhãn thủ công bởi người có chuyên môn để tạo thành bộ dữ liệu tiêu chuẩn. Cuối cùng, nhóm đánh giá khả năng phát hiện ảo giác của nhiều mô hình ngôn ngữ lớn mã nguồn mở và đóng (GPT-4o, WizardLM-2, Claude, DeepSeek...) thông qua các độ đo như accuracy, phân tích theo lĩnh vực và pattern lỗi. Phương pháp có tính hệ thống, thực nghiệm rõ ràng, giúp xác định mô hình phù hợp cho ứng dụng trợ lý ảo trong dịch vụ công.

3. Đóng góp Khoa học và thực tiễn:

Khóa luận đóng góp bộ dữ liệu tiếng Việt đầu tiên về phát hiện ảo giác trong ngữ cảnh dịch vụ công – VietPS-Hallu, với hơn 3.700 mẫu gán nhãn. Ngoài ra, khóa luận đề xuất quy trình xây dựng dữ liệu có thể mở rộng và đánh giá thực nghiệm trên nhiều mô hình để lựa chọn phương án hiệu quả. Các kết quả có thể ứng dụng trong việc phát triển trợ lý ảo hành chính công đáng tin cậy và có thể mở rộng sang các lĩnh vực khác.

4. Báo cáo viết:

Báo cáo viết gồm 74 trang chia thành 05 chương và 04 phụ lục, trình bày mạch lạc, rõ ràng, có cấu trúc hợp lý, và ngôn ngữ học thuật phù hợp. Các bảng số liệu, hình ảnh minh họa được thiết lập có mục đích rõ ràng, được chú thích đầy đủ.

5. Trình bày trước hội đồng:

Nhóm sinh viên trình bày tốt, nắm vững nội dung và trả lời các câu hỏi phản biện một cách rõ ràng và đầy đủ.

6. Công bố khoa học/ ứng dụng thực tế:

Chưa có.

Đánh giá xếp loại: Đạt yêu cầu của Khóa luận tốt nghiệp bậc Đại học Chương trình Chuẩn.

TP.HCM, ngày 01 tháng 8 năm 2025

Giảng viên phản biện

Nguyễn Ngọc Thảo