HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Vũ Minh Đam

NGHIÊN CỨU PHƯƠNG PHÁP PHÁT HIỆN DỮ LIỆU RIÊNG TƯ ĐƯỢC SỬ DỤNG TRONG HUẤN LUYỆN MÔ HÌNH NGÔN NGỮ LỚN

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

Hà Nội – 2026

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Vũ Minh Đam

NGHIÊN CỨU PHƯƠNG PHÁP PHÁT HIỆN DỮ LIỆU RIÊNG TƯ ĐƯỢC SỬ DỤNG TRONG HUẤN LUYỆN MÔ HÌNH NGÔN NGỮ LỚN

CHUYÊN NGÀNH	HỆ THỐNG THÔNG TIN
MÃ SỐ	8.48.01.04

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. DƯƠNG TRẦN ĐỨC

Hà Nội – 2026

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi.

Các nội dung lý thuyết, phương pháp nghiên cứu, thuật toán và kết quả thực nghiệm được trình bày trong đề án là trung thực và chưa từng được công bố trong bất kỳ công trình nào khác.

Tôi xin chịu hoàn toàn trách nhiệm về tính trung thực và tính nguyên bản của đề án này trước Hội đồng bảo vệ và pháp luật.

Hà Nội, Ngày 20 tháng 06 năm 2026

Tác giả đề án

Vũ Minh Đam

LỜI CẢM ƠN

Để hoàn thành đề án thạc sĩ này, tôi đã nhận được sự hướng dẫn tận tình, sự hỗ trợ và động viên to lớn từ các thầy cô, gia đình, bạn bè và đồng nghiệp.

Trước tiên, tôi xin bày tỏ lòng biết ơn sâu sắc nhất đến TS. Dương Trần Đức, người đã trực tiếp định hướng, dìu dắt và truyền đạt cho tôi những kiến thức khoa học quý báu, cũng như dành nhiều thời gian góp ý, chỉnh sửa để tôi có thể hoàn thiện hướng nghiên cứu về Mô hình Ngôn ngữ Lớn (LLM) và hoàn thành tốt đề án này.

Tôi cũng xin trân trọng cảm ơn Học viện Công nghệ Bưu chính Viễn thông và Phòng Đào tạo Sau Đại học đã tạo điều kiện thuận lợi nhất về môi trường học tập, cơ sở vật chất và thủ tục hành chính trong suốt quá trình tôi học tập và nghiên cứu tại trường. Xin gửi lời cảm ơn đến quý thầy cô đã trực tiếp giảng dạy, cung cấp cho tôi nền tảng tri thức vững chắc làm hành trang cho con đường nghiên cứu.

Cuối cùng, tôi xin gửi lời tri ân sâu sắc đến gia đình, người thân và những người bạn đồng nghiệp đã luôn là điểm tựa tinh thần vững chắc, luôn động viên, chia sẻ và tạo mọi điều kiện tốt nhất để tôi tập trung hoàn thành chương trình cao học.

Dù đã nỗ lực hết mình, đề án chắc chắn vẫn không tránh khỏi những thiếu sót nhất định. Tôi rất mong nhận được sự góp ý quý báu từ Quý thầy cô trong Hội đồng đánh giá để đề án được hoàn thiện hơn.

Xin chân thành cảm ơn!

MỤC LỤC

LỜI CAM ĐOAN I
LỜI CẢM ƠN II
MỤC LỤC III
DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT V
DANH SÁCH CÁC BẢNG BIỂU VI
I. MỞ ĐẦU 1
1. LÝ DO CHỌN ĐỀ TÀI 1
2. TỔNG QUAN VỀ VẤN ĐỀ NGHIÊN CỨU 2
2.1. Sự vô hiệu của các phương pháp MIA truyền thống 2
2.2. Các hướng tiếp cận hiện tại trên LLM và điểm nghẽn 3
3. MỤC TIÊU NGHIÊN CỨU 3
3.1. Mục tiêu khoa học 4
3.2. Mục tiêu thực tiễn 4
4. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU 5
4.1. Đối tượng nghiên cứu 5
4.2. Phạm vi nghiên cứu 5
5. PHƯƠNG PHÁP NGHIÊN CỨU 6
5.1. Phương pháp nghiên cứu lý thuyết 6
5.2. Phương pháp nghiên cứu thực nghiệm 6
II. NỘI DUNG 8
CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 8
1.1. Mô hình ngôn ngữ lớn và cơ chế sinh văn bản 8
1.1.1. Mục tiêu tiền huấn luyện tự hồi quy 8
1.1.2. Giải mã và tính biến thiên của đầu ra 8
1.2. Hiện tượng ghi nhớ dữ liệu trong LLM 8
1.2.1. Ghi nhớ nguyên văn và ghi nhớ xấp xỉ 8
1.2.2. Các yếu tố ảnh hưởng đến ghi nhớ 9
1.2.3. Phân biệt ghi nhớ, truy xuất và kiến thức chung 9
1.3. Các kỹ thuật nền tảng hỗ trợ phương pháp đề xuất 9
1.3.1. Sử dụng mô hình mã hóa phụ trợ để biểu diễn đặc trưng ngữ nghĩa 9
1.3.2. Đánh giá đa chiều và phân loại bằng học máy 10
1.4. Chỉ số đánh giá 12
1.4.1. Chỉ số AUC 12
1.4.2. Chỉ số TPR@5%FPR 13
CHƯƠNG 2: TỔNG QUAN CÁC PHƯƠNG PHÁP HIỆN TẠI VÀ ĐIỂM NGHẼN 14
2.1. Phương pháp Trích xuất bằng câu lệnh 14
2.1.1. Định nghĩa nền tảng về sự ghi nhớ của LLM 14
2.1.2. Các kỹ thuật Trích xuất dữ liệu thực tế 15
2.1.3. Các điểm nghẽn và hạn chế của phương pháp 15

2.2. Phương pháp Suy luận thành viên 16
2.2.1. Phương pháp dựa trên tham chiếu và rào cản chi phí 16
2.2.2. Phương pháp không tham chiếu và rào cản sai số 17
2.3. Động lực đề xuất phương pháp mới 18
CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT 20
3.1. Khung tổng quan của hệ thống 20
3.2. Chuẩn bị và phân chia dữ liệu 21
3.3. Kích thích hành vi sinh văn bản của mô hình mục tiêu 22
3.4. Trích xuất vector đặc trưng hành vi đa chiều 23
3.5. Phân loại bằng mô hình học máy 24
CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ 26
4.1. Mục tiêu thực nghiệm 26
4.2. Dữ liệu thực nghiệm 26
4.2.1. Các tập dữ liệu đánh giá 26
4.2.2. Thiết lập cấu trúc mẫu dữ liệu 29
4.3. Thiết lập thực nghiệm 29
4.3.1. Cấu trúc mẫu và cấu hình sinh 29
4.3.2. Mô hình mục tiêu 29
4.3.3. Mô hình phụ trợ trích xuất đặc trưng 29
4.3.4. Bộ phân loại 29
4.3.5. Chỉ số đánh giá 30
4.3.6. Môi trường thực nghiệm 30
4.4. Các thực nghiệm cần tiến hành 31
4.4.1. Thực nghiệm 1: Đánh giá đóng góp của các nhóm đặc trưng 31
4.4.2. Thực nghiệm 2: Đánh giá ảnh hưởng của cấu hình độ dài Prefix/Suffix 33
4.4.3. Thực nghiệm 3: So sánh các bộ phân loại và kích thước dữ liệu huấn luyện 34
4.4.4. Thực nghiệm 4: Đánh giá hiệu suất tổng thể và so sánh với các phương pháp hiện có 35
4.5. Phân tích kết quả 36
4.6. Đánh giá chi phí và khả năng mở rộng 36
4.7. Hạn chế thực nghiệm 37
4.8. Kết luận chương ..... **Error! Bookmark not defined.**
KẾT LUẬN 38
TÀI LIỆU THAM KHẢO 40

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Từ viết tắt	Giải nghĩa theo tiếng anh	Giải nghĩa theo tiếng việt
AUC	Area Under the ROC Curve	Diện tích dưới đường cong ROC (Thước đo hiệu suất của mô hình phân loại nhị phân)
BERT	Bidirectional Encoder Representations from Transformers	Mô hình biểu diễn ngôn ngữ hai chiều từ Transformer
FPR	False Positive Rate	Tỷ lệ dương tính giả (Tỷ lệ nhận diện sai dữ liệu chưa học thành đã học)
LLM	Large Language Model	Mô hình ngôn ngữ lớn
MIA	Membership Inference Attack	Tấn công suy luận thành viên (Phương pháp xác định xem một mẫu dữ liệu có thuộc tập huấn luyện hay không)
ML	Machine Learning	Học máy
MLP	Multi-Layer Perceptron	Mạng nơ-ron truyền thẳng đa lớp
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
PPL	Perplexity	Độ hoang mang (Thước đo mức độ dự đoán từ tiếp theo của mô hình ngôn ngữ)
PrototypicalNet	Prototypical Networks	Mạng nơ-ron nguyên mẫu
RAG	Retrieval-Augmented Generation	Kiến trúc sinh văn bản tăng cường truy xuất
RLHF	Reinforcement Learning from Human Feedback	Học tăng cường từ phản hồi của con người
TPR@5%FPR	True Positive Rate at 5% FPR	Tỷ lệ dương tính thật khi giới hạn tỷ lệ dương tính giả ở mức 5%

DANH SÁCH CÁC BẢNG BIỂU

Bảng 4.1: Điểm số AUC các tập feature đem lại trên kiến trúc cơ bản đã đề xuất 32
Bảng 4.2: Điểm số TPR@5%FPR các tập feature đem lại trên kiến trúc cơ bản đã đề xuất 32
Bảng 4.3: Điểm số AUC khi thay đổi tỉ lệ token của Prefix/Suffix trên kiến trúc cơ bản đã đề xuất.....33
Bảng 4.4: Điểm số TPR@5%FPR khi thay đổi tỉ lệ token của Prefix/Suffix trên kiến trúc cơ bản đã đề xuất.....33
Bảng 4.5: Điểm số AUC tương ứng với từng bộ phân loại trên kiến trúc cơ bản đã đề xuất 34
Bảng 4.6: Điểm số TPR@5%FPR tương ứng với từng bộ phân loại trên kiến trúc cơ bản đã đề xuất.....34
Bảng 4.7: So sánh điểm số AUC của phương pháp kết hợp đề xuất trong đề án với các phương pháp nổi bật hiện có 35
Bảng 4.8: So sánh điểm số TPR@5%FPR của phương pháp kết hợp đề xuất trong đề án với các phương pháp nổi bật hiện có.....35

I. MỞ ĐẦU

1. Lý do chọn đề tài

Sự phát triển mạnh mẽ của các Mô hình Ngôn ngữ Lớn (Large Language Models - LLMs) đã mang lại hiệu suất vượt trội trong hàng loạt tác vụ xử lý ngôn ngữ tự nhiên (NLP) như trả lời câu hỏi, dịch máy và tóm tắt văn bản. Sức mạnh của các mô hình này chủ yếu đến từ việc sử dụng một lượng dữ liệu khổng lồ trong quá trình huấn luyện. Tuy nhiên, việc mở rộng quy mô dữ liệu không được kiểm soát chặt chẽ đã dẫn đến việc các tập dữ liệu huấn luyện (training datasets) thường bao hàm cả các thông tin nhạy cảm như dữ liệu cá nhân, sách, báo chí và các tài liệu có bản quyền. Hậu quả là, LLM có nguy cơ rò rỉ quyền riêng tư hoặc vi phạm bản quyền nghiêm trọng nếu mô hình sinh ra các đoạn văn bản giống hệt hoặc có độ tương đồng cao với dữ liệu gốc. Thực tế này đã làm nảy sinh nhiều tranh chấp pháp lý phức tạp, tiêu biểu như vụ kiện giữa The New York Times và OpenAI vào năm 2023 [1], hay các vụ kiện tập thể chống lại Stable Diffusion và Midjourney [2].

Để bảo vệ quyền sở hữu trí tuệ của người sáng tạo và quyền riêng tư của người dùng, việc phát triển các phương pháp phát hiện xem một đoạn dữ liệu cụ thể có được sử dụng để huấn luyện LLM hay không (Tấn công Suy luận Thành viên - Membership Inference Attack) trở thành một yêu cầu cấp thiết. Mặc dù các phương pháp Suy luận Thành viên (MIA) truyền thống đã đạt được nhiều thành tựu trên các mô hình học máy (ML) thông thường, nhưng chúng lại hoàn toàn thất bại hoặc kém hiệu quả khi áp dụng lên LLM. Nguyên nhân là do rào cản về không gian tham số khổng lồ, tính đa dạng trong đầu ra của LLM và đặc biệt là chi phí quá đắt đỏ để huấn luyện các "mô hình bóng" (shadow models) nhằm mô phỏng dữ liệu [3].

Hiện nay, các nghiên cứu phát hiện dữ liệu huấn luyện cho LLM đang vấp phải bốn điểm nghẽn kỹ thuật vô cùng lớn:

- Chi phí tính toán và yêu cầu tài nguyên quá đắt đỏ: Các phương pháp sử dụng mô hình tham chiếu (như huấn luyện mạng seq2seq để học ánh xạ giữa văn bản đầu vào và phần sinh tiếp) đòi hỏi lượng tài nguyên khổng lồ và thời gian huấn luyện dài, khiến chúng trở nên phi thực tế khi triển khai trên quy mô lớn.

- Tỷ lệ nhận diện sai (dương tính giả - false positive rate) rất cao: Các phương pháp không tham chiếu hiện hành phần lớn chỉ phân tích thông tin bề mặt, điển hình là việc sử dụng một thước đo đơn chiều duy nhất như Độ hoang mang (Perplexity). Việc này mang lại sự không chắc chắn rất lớn, khiến hệ

thống dễ dàng nhận diện nhầm các cụm từ phổ biến nhưng chưa từng được học thành dữ liệu huấn luyện.

- Sự nhiễu loạn từ kiến trúc RAG. Nhiều LLM hiện nay được tích hợp kiến trúc RAG để truy xuất thông tin từ cơ sở dữ liệu bên ngoài vào thời gian thực. Khi LLM sinh ra chính xác một đoạn văn bản có bản quyền, rất khó để phân biệt đoạn văn bản đó là hệ quả của việc mô hình đã ghi nhớ trong trọng số hay là do nó vừa truy xuất động từ nguồn ngoài, làm sai lệch kết quả phát hiện.

- Rào cản căn chỉnh an toàn (RLHF): Các cơ chế bảo vệ bản quyền và quyền riêng tư trên LLM thương mại chủ động chặn việc sao chép nguyên văn, khiến phương pháp mồi nhử thất bại ở hơn 90% các trường hợp trên thực tế.

Nhằm giải quyết triệt để những hạn chế trên, giới nghiên cứu học thuật đang hướng tới việc khám phá các cấu trúc nội bộ của mô hình thay vì chỉ đo lường xác suất bề nổi, đồng thời kết hợp nhiều thước đo thành các vector đặc trưng đa chiều để đưa vào các mô hình phân loại học máy nhằm giảm thiểu sai số.

Xuất phát từ những vấn đề trên, đề án lựa chọn nghiên cứu bài toán phát hiện dữ liệu huấn luyện trong mô hình ngôn ngữ lớn theo hướng tiếp cận hộp đen, dựa trên hành vi phản hồi của mô hình. Thay vì yêu cầu mô hình sinh lại nguyên văn dữ liệu gốc hoặc truy cập vào thông tin nội bộ, đề án hướng tới việc đánh giá mức độ ghi nhớ xấp xỉ thông qua việc quan sát phần văn bản mô hình sinh ra khi được cung cấp một đoạn tiền tố. Cách tiếp cận này có ý nghĩa quan trọng trong việc hỗ trợ kiểm toán dữ liệu huấn luyện, bảo vệ quyền riêng tư, bảo vệ bản quyền và nâng cao tính minh bạch của các mô hình ngôn ngữ lớn.

2. Tổng quan về vấn đề nghiên cứu

Bài toán cốt lõi trong việc xác định dữ liệu huấn luyện của Mô hình Ngôn ngữ Lớn (LLM) về bản chất là một dạng Tấn công Suy luận Thành viên (Membership Inference Attack - MIA). Tuy nhiên, việc áp dụng bài toán này lên các LLM hiện đại đang vấp phải những thách thức kỹ thuật vô cùng lớn, tạo ra những khoảng trống nghiên cứu cần được giải quyết.

2.1. Sự vô hiệu của các phương pháp MIA truyền thống

Trong lĩnh vực học máy truyền thống, các phương pháp MIA đã đạt được độ chính xác cao nhờ việc phân tích sự chênh lệch về độ tự tin (confidence) giữa dữ liệu đã học và chưa học, thường thông qua việc huấn luyện các "mô hình bóng" (shadow models) để mô phỏng hành vi của mô hình mục tiêu. Tuy nhiên, phương pháp này hoàn toàn bất khả thi khi áp dụng lên LLM. Do không gian tham số của LLM lên tới hàng chục hay hàng trăm tỷ, chi phí tính toán để huấn luyện các mô

hình bóng là quá sức tưởng tượng và phi thực tế. Hơn nữa, đầu ra của các mô hình học máy truyền thống mang tính xác định (như nhãn phân loại), trong khi LLM sinh văn bản mang tính xác suất và ngẫu nhiên cao, khiến việc ánh xạ trực tiếp từ hành vi sinh văn bản sang một mẫu dữ liệu huấn luyện cụ thể trở nên vô cùng khó khăn.

2.2. Các hướng tiếp cận hiện tại trên LLM và điểm nghẽn

Để vượt qua các rào cản trên, giới nghiên cứu đã đề xuất hai hướng tiếp cận chính dành riêng cho LLM, nhưng cả hai đều đang bộc lộ những điểm yếu lớn:

- Hướng tiếp cận 1: Trích xuất dữ liệu ghi nhớ (Memorization Data Extraction). Phương pháp này sử dụng các kỹ thuật câu lệnh (Prompt Engineering) để "ép" LLM tự động xuất ra nguyên văn các đoạn dữ liệu mà nó đã ghi nhớ. Mặc dù cung cấp bằng chứng trực quan, phương pháp này có tỷ lệ thành công rất thấp do vấp phải các cơ chế bảo vệ an toàn của mô hình thương mại (tiêu biểu như Học tăng cường từ phản hồi con người - RLHF), vốn chủ động ngăn chặn mô hình xuất ra dữ liệu có bản quyền,. Việc yêu cầu LLM phải xuất ra trùng khớp 100% nguyên văn cũng là một tiêu chuẩn quá khắt khe đối với tính đa dạng trong sinh ngôn ngữ của LLM [4].

- Hướng tiếp cận 2: Suy luận thành viên (Membership Inference)

    - Suy luận thành viên không tham chiếu (Non-reference Membership Inference): Để tiết kiệm chi phí, phương pháp này phân tích đặc trưng bề mặt của đầu ra (như Độ hoang mang - PPL, hay Min-K% Prob). Tuy nhiên, việc chỉ phụ thuộc vào thước đo đơn chiều khiến hệ thống dễ nhầm lẫn văn bản phổ biến với dữ liệu đã học, dẫn đến tỷ lệ dương tính giả rất cao.

    - Suy luận dựa trên tham chiếu (Reference Membership Inference) sử dụng các mô hình phụ trợ làm hệ quy chiếu để phân biệt dữ liệu đã học và chưa học trong kịch bản hộp đen. Hướng tiếp cận này gồm hai kỹ thuật chính: thứ nhất là huấn luyện các "mô hình bóng" để mô phỏng LLM mục tiêu, nhưng cách này đòi hỏi chi phí tính toán khổng lồ và hoàn toàn phi thực tế. Hướng thứ hai là dùng các mô hình ngôn ngữ quy chiếu nhỏ gọn hơn (như GPT-2, Tinystories) nhằm so sánh độ chênh lệch của các thước đo như Độ hoang mang. Mặc dù giúp tiết kiệm tài nguyên đáng kể, kỹ thuật dùng mô hình nhỏ này vẫn đang ở giai đoạn sơ khai và chưa đạt được độ chính xác cao [3].

3. Mục tiêu nghiên cứu

Mục đích nghiên cứu của đề án là nghiên cứu bài toán phát hiện dữ liệu riêng tư được sử dụng trong quá trình huấn luyện hoặc tiền huấn luyện mô hình ngôn ngữ lớn, từ đó đánh giá khả năng xác định một văn bản cụ thể có từng xuất hiện trong dữ liệu huấn luyện của mô hình hay không. Trên cơ sở đó, đề án hướng tới việc xây

dựng và thử nghiệm một phương pháp phát hiện phù hợp với bối cảnh mô hình hộp đen, đặc biệt muốn đóng góp cho thử nghiệm, nghiên cứu phương pháp phát hiện dữ liệu liên quan đến dữ liệu tiếng Việt - một hướng nghiên cứu hiện còn thiếu các công trình chuyên sâu và hệ thống.

3.1. Mục tiêu khoa học

Về mặt khoa học, đề án hướng tới hệ thống hóa cơ sở lý thuyết liên quan đến mô hình ngôn ngữ lớn, dữ liệu huấn luyện, hiện tượng ghi nhớ dữ liệu và bài toán suy luận thành viên. Trên cơ sở đó, nghiên cứu làm rõ nguyên lý phát hiện dấu vết dữ liệu huấn luyện thông qua hành vi phản hồi của mô hình, đặc biệt trong bối cảnh người nghiên cứu không thể truy cập vào trọng số, logit hoặc xác suất token của mô hình mục tiêu.

Bên cạnh đó, đề án khảo sát và phân tích các phương pháp phát hiện dữ liệu huấn luyện hiện có như trích xuất dữ liệu ghi nhớ, perplexity, Min-K% Prob, mô hình tham chiếu và các hướng tiếp cận hộp đen. Việc tổng hợp này nhằm chỉ ra ưu điểm, hạn chế và khoảng trống của các phương pháp hiện hành, từ đó làm cơ sở đề xuất một hướng tiếp cận phù hợp hơn với các mô hình ngôn ngữ thương mại dạng hộp đen.

Từ khoảng trống nghiên cứu đã xác định, đề án mô hình hóa bài toán dưới dạng phân loại nhị phân member và non-member. Văn bản đầu vào được chia thành tiền tố và hậu tố; tiền tố được dùng để yêu cầu mô hình sinh phần tiếp nối, sau đó phần văn bản sinh ra được so sánh với hậu tố thật bằng các đặc trưng từ vựng và ngữ nghĩa. Ngoài ra, đề án xây dựng và thử nghiệm phương pháp trên dữ liệu tiếng Việt nhằm đánh giá khả năng áp dụng đối với một ngôn ngữ còn ít được nghiên cứu trong lĩnh vực phát hiện dữ liệu huấn luyện của mô hình ngôn ngữ lớn.

3.2. Mục tiêu thực tiễn

Về mặt thực tiễn, đề án hướng tới xây dựng một quy trình phát hiện dấu hiệu văn bản từng xuất hiện trong dữ liệu huấn luyện của mô hình ngôn ngữ lớn mà không cần truy cập dữ liệu gốc, trọng số hoặc xác suất token. Cách tiếp cận này phù hợp với các mô hình thương mại dạng hộp đen, nơi người dùng chỉ có thể gửi truy vấn và quan sát văn bản đầu ra.

Đề án đồng thời thử nghiệm phương pháp trên dữ liệu tiếng Việt và đánh giá bằng các chỉ số như Accuracy, AUC và TPR@5%FPR nhằm xác định khả năng phân biệt giữa văn bản có khả năng đã xuất hiện và chưa xuất hiện trong dữ liệu huấn luyện. Kết quả nghiên cứu có thể hỗ trợ kiểm toán dữ liệu, phát hiện nguy cơ

sử dụng tài liệu có bản quyền, rò rỉ thông tin và đánh giá mức độ minh bạch của mô hình ngôn ngữ lớn.

4. Đối tượng và phạm vi nghiên cứu

4.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề án là bài toán phát hiện liệu một văn bản có từng được sử dụng trong quá trình huấn luyện hoặc tiền huấn luyện mô hình ngôn ngữ lớn hay không. Bài toán được tiếp cận dưới góc độ suy luận thành viên, trong đó mục tiêu là phân biệt giữa các mẫu văn bản có khả năng thuộc và không thuộc tập dữ liệu huấn luyện của mô hình dựa trên hành vi phản hồi khi được truy vấn.

Đề án tập trung vào các mô hình ngôn ngữ lớn trong kịch bản hộp đen. Trong bối cảnh này, người nghiên cứu không thể truy cập vào dữ liệu huấn luyện gốc, trọng số, gradient, logit hoặc phân phối xác suất token mà chỉ có thể gửi truy vấn và quan sát văn bản đầu ra do mô hình sinh ra.

Trên cơ sở đó, nghiên cứu xem xét các đặc trưng hành vi như khả năng sinh tiếp văn bản, mức độ trùng khớp từ vựng và mức độ tương đồng ngữ nghĩa giữa phần văn bản do mô hình tạo ra với phần văn bản gốc. Các đặc trưng này được sử dụng làm đầu vào cho mô hình phân loại nhằm đánh giá khả năng một văn bản thuộc nhóm member hoặc non-member.

4.2. Phạm vi nghiên cứu

Về phạm vi mô hình, đề án tập trung vào các mô hình ngôn ngữ lớn trong kịch bản hộp đen. Trong bối cảnh này, người nghiên cứu không thể truy cập vào dữ liệu huấn luyện gốc, trọng số, gradient, logit hoặc xác suất token của mô hình mục tiêu mà chỉ có thể gửi truy vấn và quan sát văn bản đầu ra. Cách tiếp cận này phù hợp với điều kiện sử dụng thực tế của phần lớn các mô hình ngôn ngữ thương mại hiện nay.

Về phạm vi dữ liệu, đề án chỉ xem xét dữ liệu dạng văn bản, bao gồm các đoạn văn, bài viết web, tin tức, tài liệu học thuật, sách và những nội dung công khai khác. Các loại dữ liệu như hình ảnh, âm thanh, video, dữ liệu có cấu trúc và dữ liệu đa phương thức không thuộc phạm vi nghiên cứu. Việc giới hạn ở văn bản giúp đề án tập trung vào khả năng ghi nhớ và sinh tiếp chuỗi ngôn ngữ của mô hình.

Về phạm vi ngôn ngữ, nghiên cứu sử dụng các tập dữ liệu tiếng Anh để đối chiếu với những benchmark và phương pháp đã được công bố, đồng thời bổ sung các tập dữ liệu tiếng Việt để đánh giá khả năng áp dụng trong bối cảnh Việt Nam.

Dữ liệu tiếng Việt có những đặc điểm riêng về dấu thanh, tách từ, token hóa và cách biểu đạt, có thể ảnh hưởng đến hiệu quả phát hiện dấu vết dữ liệu huấn luyện.

Về phương pháp và thực nghiệm, đề án sử dụng phần tiền tố của văn bản để truy vấn mô hình, thu thập phần văn bản sinh tiếp và so sánh với hậu tố thật bằng các đặc trưng từ vựng và ngữ nghĩa. Các đặc trưng thu được được đưa vào bộ phân loại để phân biệt giữa nhóm văn bản có khả năng đã xuất hiện và chưa xuất hiện trong dữ liệu huấn luyện. Hiệu quả của phương pháp được đánh giá chủ yếu thông qua các chỉ số Accuracy, AUC và TPR@5%FPR.

5. Phương pháp nghiên cứu

Để thực hiện đề án, nghiên cứu sử dụng kết hợp phương pháp nghiên cứu lý thuyết và phương pháp nghiên cứu thực nghiệm. Việc kết hợp hai nhóm phương pháp này giúp đề tài vừa xây dựng được cơ sở khoa học cho bài toán phát hiện dữ liệu huấn luyện trong mô hình ngôn ngữ lớn, vừa kiểm chứng được khả năng áp dụng của phương pháp đề xuất trong điều kiện thực tế.

5.1. Phương pháp nghiên cứu lý thuyết

Phương pháp nghiên cứu lý thuyết được sử dụng để hệ thống hóa cơ sở khoa học về mô hình ngôn ngữ lớn, dữ liệu huấn luyện, hiện tượng ghi nhớ và bài toán suy luận thành viên. Thông qua việc khảo sát các công trình đã công bố, đề án làm rõ bản chất của bài toán xác định một văn bản có từng xuất hiện trong quá trình huấn luyện mô hình hay không.

Trên cơ sở đó, nghiên cứu phân tích và so sánh các nhóm phương pháp hiện có như trích xuất dữ liệu ghi nhớ, mô hình bóng, mô hình tham chiếu, perplexity, xác suất token, Min-K% Prob và các hướng tiếp cận hộp đen. Việc tổng hợp này nhằm chỉ ra ưu điểm, hạn chế và điều kiện áp dụng của từng phương pháp, đặc biệt trong trường hợp không thể truy cập dữ liệu huấn luyện, trọng số, logit hoặc xác suất token của mô hình mục tiêu.

Ngoài ra, đề án mô hình hóa bài toán dưới dạng phân loại nhị phân và xây dựng cơ sở lý thuyết cho hướng tiếp cận ghi nhớ xấp xỉ. Thay vì yêu cầu mô hình sinh lại nguyên văn dữ liệu gốc, nghiên cứu đánh giá mức độ tương đồng giữa văn bản sinh ra và văn bản thật thông qua các đặc trưng từ vựng, ngữ nghĩa và biểu diễn vector. Đây là nền tảng để xây dựng quy trình thực nghiệm và lựa chọn các chỉ số đánh giá phù hợp.

5.2. Phương pháp nghiên cứu thực nghiệm

Phương pháp nghiên cứu thực nghiệm được sử dụng để xây dựng, thử nghiệm và đánh giá hiệu quả của phương pháp phát hiện dữ liệu huấn luyện trong bối cảnh mô hình ngôn ngữ lớn hộp đen. Trọng tâm là kiểm tra khả năng xác định dấu hiệu một văn bản từng xuất hiện trong dữ liệu huấn luyện thông qua việc truy vấn mô hình và phân tích văn bản đầu ra.

Trước hết, đề án xây dựng bộ dữ liệu thực nghiệm gồm các văn bản tiếng Anh và tiếng Việt thuộc nhiều miền khác nhau. Mỗi văn bản được chia thành phần tiền tố và hậu tố; tiền tố được dùng để truy vấn mô hình, còn hậu tố thật được giữ lại làm dữ liệu đối chiếu. Trong quá trình thực nghiệm, hệ thống chỉ thu thập phần văn bản do mô hình sinh ra, không sử dụng trọng số, logit, gradient hoặc xác suất token.

Tiếp theo, phần văn bản sinh ra được so sánh với hậu tố thật để trích xuất các đặc trưng từ vựng và ngữ nghĩa, như mức độ trùng khớp chuỗi, độ tương đồng nội dung và biểu diễn vector từ mô hình mã hóa BERT. Các đặc trưng này được đưa vào bộ phân loại học máy nhằm dự đoán trạng thái member hoặc non-member.

Cuối cùng, hiệu quả của phương pháp được đánh giá bằng các chỉ số Accuracy, AUC và TPR@5%FPR, đồng thời phân tích ảnh hưởng của miền dữ liệu, cấu hình độ dài tiền tố–hậu tố và đặc điểm tiếng Việt. Qua đó, đề án kiểm chứng tính khả thi của phương pháp và đánh giá khả năng ứng dụng trong kiểm toán dữ liệu huấn luyện, bảo vệ quyền riêng tư và bản quyền.

II. NỘI DUNG

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1. Mô hình ngôn ngữ lớn và cơ chế sinh văn bản

1.1.1. Mục tiêu tiền huấn luyện tự hồi quy

LLM tự hồi quy học phân phối xác suất của chuỗi token bằng cách dự đoán token tiếp theo từ các token đứng trước. Với chuỗi x = $(x_1, x_2, \dots, x_t)$ xác suất chuỗi được phân rã thành tích các xác suất có điều kiện:

$$P\theta(x) = \prod_{t-1} T\, P\theta(x_t \mid x_1, x_2, \dots, x_{t-1})$$

Quá trình tối ưu làm tăng xác suất của các mẫu xuất hiện trong dữ liệu huấn luyện. Tuy nhiên, xác suất cao không đồng nghĩa trực tiếp với ghi nhớ. Một chuỗi có thể dễ dự đoán vì cấu trúc phổ biến, còn một chuỗi hiếm có thể được tái tạo do xuất hiện lặp lại hoặc do mô hình có đủ năng lực lưu giữ chi tiết. Đây là lý do các phương pháp chỉ dựa trên một điểm số đơn lẻ dễ gặp dương tính giả.

1.1.2. Giải mã và tính biến thiên của đầu ra

Sau tiền huấn luyện, văn bản được sinh bằng một chiến lược giải mã như greedy decoding, beam search hoặc lấy mẫu theo temperature/top-p. Với cùng một Prefix, đầu ra có thể thay đổi nếu sử dụng lấy mẫu. Ngược lại, temperature = 0 thường làm kết quả ổn định hơn nhưng vẫn không bảo đảm tuyệt đối giống nhau giữa các phiên bản dịch vụ hoặc thời điểm gọi API. Do đó, mọi thực nghiệm hộp đen cần ghi lại model identifier, ngày truy vấn và tham số sinh.

Khác biệt quan trọng so với mô hình phân loại không phải là mọi mô hình truyền thống đều "xác định", mà là đầu ra của LLM có cấu trúc chuỗi, không gian đáp án rất lớn và có nhiều tiếp nối hợp lý. Vì vậy, dấu hiệu thành viên cần được đánh giá trên nhiều khía cạnh của văn bản sinh ra thay vì chỉ kiểm tra một nhãn hoặc một giá trị confidence.

1.2. Hiện tượng ghi nhớ dữ liệu trong LLM

1.2.1. Ghi nhớ nguyên văn và ghi nhớ xấp xỉ

Ghi nhớ nguyên văn xảy ra khi mô hình tái tạo đúng hoặc gần đúng chuỗi đã xuất hiện trong huấn luyện. Carlini và cộng sự phân biệt ghi nhớ có thể trích xuất, trong đó tồn tại một prompt khiến mô hình sinh ra chuỗi mục tiêu, và ghi nhớ có thể khám phá, trong đó một Prefix của chính văn bản được dùng để yêu cầu mô hình hoàn thành Suffix [5][10].

Trong thực tế, mô hình có thể không sao chép từng token nhưng vẫn sinh ra phần tiếp nối giữ gần nguyên ý, cấu trúc hoặc thực thể của văn bản gốc. Đề án gọi đây là ghi nhớ xấp xỉ ở cấp hành vi. Khái niệm này cần được dùng thận trọng: tương đồng ngữ nghĩa có thể bắt nguồn từ ghi nhớ, nhưng cũng có thể chỉ phản ánh khả năng khái quát và dự đoán hợp lý của mô hình.

1.2.2. Các yếu tố ảnh hưởng đến ghi nhớ

Mức độ ghi nhớ chịu ảnh hưởng bởi nhiều yếu tố: tần suất lặp lại và mức độ trùng lặp của dữ liệu; độ hiếm và tính độc nhất của chuỗi; quy mô mô hình; số bước huấn luyện; chất lượng làm sạch dữ liệu; và chiến lược giải mã [4], [10]. Vì vậy, không nên quy hiện tượng ghi nhớ chỉ cho việc "học nhiều epoch" hoặc suy ra trực tiếp số lần một tài liệu được mô hình quan sát.

Văn bản công thức, lời mở đầu phổ biến hoặc điều khoản pháp lý có thể được mô hình hoàn thành tốt dù chưa từng thấy đúng tài liệu đó. Ngược lại, một tài liệu thực sự thuộc tập huấn luyện có thể không được tái tạo nếu chỉ xuất hiện một lần, bị cắt nhỏ trong quá trình tiền xử lý, hoặc bị cơ chế căn chỉnh sau tiền huấn luyện hạn chế. Hai trường hợp này lần lượt tạo ra dương tính giả và âm tính giả.

1.2.3. Phân biệt ghi nhớ, truy xuất và kiến thức chung

Một phản hồi giống văn bản gốc không luôn chứng minh rằng nội dung được lưu trong trọng số. Nếu hệ thống mục tiêu tích hợp truy xuất tăng cường (RAG), kết quả có thể đến từ kho dữ liệu bên ngoài tại thời điểm suy luận. Tương tự, fine-tuning hoặc cập nhật sau mốc tiền huấn luyện có thể làm thay đổi phạm vi tri thức. Trong kiểm toán hộp đen, cần ghi rõ có bật duyệt web/RAG hay không và chỉ diễn giải kết quả trong phạm vi cấu hình đã kiểm soát

1.3. Các kỹ thuật nền tảng hỗ trợ phương pháp đề xuất

Để khắc phục những hạn chế về chi phí tính toán và nguy cơ nhận diện sai của các phương pháp Tấn công Suy luận Thành viên hiện hành, phương pháp đề xuất trong đề án được xây dựng theo hướng khai thác hành vi sinh văn bản của mô hình ngôn ngữ lớn trong kịch bản hộp đen. Thay vì phụ thuộc vào mô hình bóng quy mô lớn, truy cập nội bộ mô hình hoặc một thước đo đơn lẻ như Perplexity, đề án tập trung vào việc thu thập các tín hiệu hành vi từ đầu ra sinh tiếp văn bản, biểu diễn các tín hiệu này thành vector đặc trưng đa chiều và sử dụng mô hình học máy để phân loại member/non-member. Cách tiếp cận này giúp giảm chi phí triển khai, phù hợp với các LLM chỉ truy cập qua API, đồng thời cải thiện tính ổn định so với các phương pháp đánh giá đơn chiều.

1.3.1. Sử dụng mô hình mã hóa phụ trợ để biểu diễn đặc trưng ngữ nghĩa

Trong các phương pháp Tấn công Suy luận Thành viên (MIA) hiện tại, hướng tiếp cận dựa trên tham chiếu (reference-based methods) thường sử dụng các "mô hình bóng" (shadow models) hoặc các mô hình ngôn ngữ quy chiếu nhỏ (như GPT-2, TinyStories, hay Phi-1.5) để hỗ trợ quá trình phân tích [6]. Mục đích của các mô hình tham chiếu này là mô phỏng lại hành vi tính toán xác suất hoặc độ hoang mang (perplexity) của mô hình ngôn ngữ lớn (LLM) mục tiêu nhằm tìm ra sự chênh lệch . Tuy nhiên, phương pháp này đòi hỏi chi phí tính toán đắt đỏ và thường mang lại sai số lớn do các mô hình nhỏ không thể mô phỏng chính xác năng lực của một LLM khổng lồ, khiến chúng trở nên chưa hoàn thiện và thiếu khả thi trong thực tế.

Để khắc phục rào cản này, thay vì sử dụng các mô hình tham chiếu sinh văn bản để tính toán và mô phỏng xác suất, đề án đề xuất sử dụng BERT (Bidirectional Encoder Representations from Transformers) đóng vai trò như một bộ mã hóa phụ trợ hoàn toàn độc lập.

Khác với các LLM (như GPT hay LLaMA) vốn là các mô hình tự hồi quy chuyên về "sinh" văn bản, BERT là một mô hình biểu diễn ngôn ngữ đã được huấn luyện sẵn (Pre-trained Encoder) hoạt động theo cơ chế hai chiều (Bidirectional) . Nhờ cơ chế này, BERT không tập trung vào việc dự đoán từ tiếp theo, mà xuất sắc trong việc đọc hiểu ngữ cảnh sâu của câu từ. Chức năng cốt lõi của BERT là "phiên dịch" các đoạn văn bản ngôn ngữ tự nhiên thành các vector đặc trưng (embeddings) trong không gian toán học nhiều chiều, nơi các câu văn có ý nghĩa tương đồng sẽ có vị trí hình học gần nhau.

Do bị giới hạn ở bối cảnh hộp đen (không thể truy cập vào xác suất token, logit hay tham số nội bộ của LLM mục tiêu), hệ thống bắt buộc phải đánh giá hành vi sinh văn bản thông qua đầu ra bề mặt . Thay vì dùng mô hình tham chiếu để tính toán sự chênh lệch xác suất một cách tốn kém, đề án dùng BERT để chuyển đổi phần văn bản do LLM sinh ra (Generated Suffix) và phần văn bản tham chiếu (True Suffix) thành các vector nhúng.

Dựa trên các biểu diễn vector đa chiều này, hệ thống dễ dàng đo lường khoảng cách và độ tương đồng về mặt ý nghĩa (thông qua Cosine Similarity, L2 Distance, v.v.) . Việc thay thế mô hình tham chiếu bằng bộ mã hóa BERT không chỉ giúp loại bỏ chi phí huấn luyện các mô hình bóng đắt đỏ, mà còn giúp phát hiện hiệu quả hiện tượng "ghi nhớ xấp xỉ" (approximate memorization) – tức là khi LLM sinh ra một nội dung có ý nghĩa tương đương với dữ liệu huấn luyện dù không sử dụng chính xác các từ vựng gốc

1.3.2. Đánh giá đa chiều và phân loại bằng học máy

Một điểm nghẽn kỹ thuật lớn của nhiều phương pháp Membership Inference Attack không tham chiếu hiện nay là việc phụ thuộc vào một thước đo duy nhất để quyết định trạng thái thành viên của mẫu dữ liệu. Ví dụ, các phương pháp dựa trên Perplexity thường giả định rằng nếu một văn bản có độ hoang mang thấp thì mô hình có khả năng đã từng thấy văn bản đó trong quá trình huấn luyện. Tuy nhiên, giả định này không phải lúc nào cũng đúng. Một đoạn văn bản phổ biến, có cấu trúc quen thuộc hoặc thuộc miền tri thức thường gặp có thể được mô hình dự đoán tốt dù chưa từng nằm trong tập huấn luyện. Khi đó, hệ thống dễ nhận diện nhầm một mẫu non-member thành member, làm tăng tỷ lệ dương tính giả.

Để khắc phục hạn chế này, hướng tiếp cận của đề án không dựa vào một chỉ số đơn lẻ, mà sử dụng chiến lược đánh giá đa chiều. Thay vì chỉ quan sát một giá trị như Perplexity, hệ thống trích xuất nhiều đặc trưng khác nhau từ hành vi sinh văn bản của mô hình mục tiêu. Cụ thể, mỗi mẫu văn bản được tách thành hai phần: Prefix và True Suffix. Phần Prefix được đưa vào mô hình ngôn ngữ lớn để yêu cầu sinh tiếp văn bản, tạo ra Generated Suffix. Sau đó, hệ thống so sánh Generated Suffix với True Suffix bằng nhiều thước đo tương đồng khác nhau. Các thước đo được sử dụng phản ánh nhiều khía cạnh của mức độ tái tạo văn bản, bao gồm:

Đặc trưng từ vựng (lexical features):

- ROUGE-L: đo mức độ trùng khớp theo chuỗi con chung dài nhất giữa văn bản sinh ra và phần suffix thật.
- Token F1: đo mức độ trùng lặp token giữa hai đoạn văn bản.
- LCS ratio: đo tỷ lệ đoạn token liên tiếp dài nhất được mô hình tái tạo đúng so với suffix thật.
- N-gram overlap: đo mức độ trùng khớp theo cụm token liên tiếp, đặc biệt hữu ích để phát hiện hiện tượng mô hình sinh lại các cụm từ giống dữ liệu gốc.

Đặc trưng ngữ nghĩa (Semantic features):

Bên cạnh các đặc trưng từ vựng, nghiên cứu còn sử dụng các đặc trưng ngữ nghĩa nhằm đo mức độ tương đồng về mặt ý nghĩa giữa văn bản do mô hình sinh ra và phần suffix thật. Cụ thể, generated suffix và true suffix được mã hóa thành các vector biểu diễn bằng mô hình BERT. Từ hai vector này, hệ thống tính các độ đo ngữ nghĩa sau:

- Cosine similarity: đo mức độ tương đồng về hướng giữa hai vector biểu diễn. Giá trị càng cao cho thấy generated suffix và true suffix càng gần nhau về mặt ngữ nghĩa.

- L2 distance: đo khoảng cách Euclidean giữa hai vector biểu diễn. Giá trị càng nhỏ cho thấy hai đoạn văn bản có biểu diễn ngữ nghĩa càng gần nhau.
- Dot product: đo tích vô hướng giữa hai vector biểu diễn, phản ánh mức độ tương quan trực tiếp giữa hai biểu diễn embedding

Các đặc trưng ngữ nghĩa này giúp phát hiện những trường hợp mô hình không sinh lại chính xác từng token nhưng vẫn tạo ra nội dung có ý nghĩa gần với suffix thật. Điều này bổ sung cho các đặc trưng từ vựng, vốn chủ yếu đo sự trùng khớp bề mặt giữa hai đoạn văn bản.

Mỗi thước đo riêng lẻ chỉ phản ánh một góc nhìn nhất định. Chẳng hạn, Token F1 có thể cao khi hai đoạn văn bản dùng nhiều từ giống nhau, nhưng không nhất thiết bảo toàn thứ tự. Ngược lại, ROUGE-L và LCS ratio nhạy hơn với thứ tự xuất hiện của token. Trong khi đó, N-gram overlap giúp phát hiện các cụm từ hoặc đoạn ngắn được mô hình tái tạo gần như nguyên văn. Vì vậy, việc kết hợp nhiều thước đo giúp hệ thống đánh giá hành vi ghi nhớ của mô hình một cách toàn diện hơn.

Cách tiếp cận này phù hợp với mục tiêu của đề án vì nó không yêu cầu truy cập vào trọng số, gradient, logit hay xác suất token bên trong mô hình mục tiêu. Toàn bộ quá trình suy luận chỉ dựa trên đầu ra văn bản sinh bởi mô hình, phù hợp với kịch bản hộp đen. Đồng thời, nhờ sử dụng vector đặc trưng đa chiều, hệ thống có khả năng giảm rủi ro nhận diện nhầm các văn bản phổ biến thành dữ liệu huấn luyện. Đây là cải tiến quan trọng so với các phương pháp đo lường đơn chiều, vốn dễ bị nhiễu bởi đặc điểm ngôn ngữ, độ phổ biến của văn bản hoặc khả năng mô hình sinh tốt với các mẫu chưa từng học.

Tóm lại, đánh giá đa chiều và phân loại bằng học máy là nền tảng quan trọng của phương pháp đề xuất. Thay vì đưa ra kết luận dựa trên một metric duy nhất, đề án xây dựng một biểu diễn hành vi gồm nhiều thước đo tương đồng giữa Generated Suffix và True Suffix, sau đó sử dụng mô hình học máy để học ranh giới phân biệt member và non-member. Cách làm này giúp phương pháp có khả năng khái quát tốt hơn, phù hợp hơn với bối cảnh mô hình ngôn ngữ lớn dạng hộp đen và trực tiếp khắc phục hạn chế của các kỹ thuật MIA đơn chiều trước đây.

1.4. Chỉ số đánh giá

1.4.1. Chỉ số AUC

AUC là diện tích dưới đường cong ROC, phản ánh khả năng xếp hạng điểm thành viên của mẫu dương cao hơn mẫu âm trên toàn bộ ngưỡng. Có thể diễn giải AUC như xác suất một member ngẫu nhiên nhận điểm cao hơn một non-member ngẫu nhiên. AUC = 0,5 tương đương xếp hạng ngẫu nhiên; giá trị càng gần 1 càng tốt.

$$AUC = P(r(x^+) > r(x^-)).$$

1.4.2. Chỉ số TPR@5%FPR

TPR@5%FPR là tỷ lệ phát hiện đúng khi ngưỡng được chọn sao cho tỷ lệ dương tính giả không vượt 5%. Chỉ số này quan trọng trong kiểm toán vì cảnh báo sai có thể gây chi phí kiểm tra lớn hoặc dẫn đến kết luận nhạy cảm. Một phương pháp có AUC vừa phải nhưng TPR@5%FPR cao vẫn có thể hữu ích trong kịch bản ưu tiên độ chính xác của cảnh báo.

$$TPR@5\%FPR = TPR(\tau^*) \text{ với } FPR(\tau^*) \leq 0,05.$$

Để đánh giá đáng tin cậy, nên báo cáo trung bình và độ lệch chuẩn trên nhiều seed, khoảng tin cậy bootstrap và số lượng mẫu ở mỗi lớp. Báo cáo hiện tại chưa có đầy đủ các thống kê này; do đó, khác biệt nhỏ giữa các cấu hình cần được diễn giải thận trọng.

CHƯƠNG 2: TỔNG QUAN CÁC PHƯƠNG PHÁP HIỆN TẠI VÀ ĐIỂM NGHẼN

Trước khi đi sâu vào phân tích các kỹ thuật phát hiện dữ liệu, việc đầu tiên và quan trọng nhất là phải xác định rõ Kiến thức tiên nghiệm (Prior Knowledge) mà người kiểm tra có được đối với Mô hình Ngôn ngữ Lớn (LLM) mục tiêu [3]. Theo cấu trúc phân tích tiêu chuẩn, kiến thức tiên nghiệm được chia thành bốn khía cạnh cốt lõi: thông tin về tập dữ liệu huấn luyện, thông tin về cấu trúc và tham số mô hình, thông tin về phương pháp huấn luyện, và dạng đầu ra mà mô hình cung cấp (văn bản sinh ra hay phân phối xác suất token)

Dựa trên mức độ truy cập vào các thông tin này, các kịch bản kiểm tra được phân loại thành hộp trắng (white-box), hộp xám (gray-box) và hộp đen (black-box) [3]. Trong điều kiện thực tế hiện nay, hầu hết các LLM thương mại đều hoạt động dưới kịch bản hộp đen khép kín, nơi người dùng hoàn toàn không có quyền truy cập vào tập huấn luyện gốc, trọng số hay thuật toán nội bộ, mà chỉ có thể tương tác duy nhất thông qua các câu lệnh (prompts) và quan sát văn bản đầu ra [3]. Chính sự giới hạn khắt khe về kiến thức tiên nghiệm này đã định hình và thúc đẩy giới nghiên cứu tập trung vào hai hướng tiếp cận chủ đạo để giải quyết bài toán phát hiện dữ liệu huấn luyện: phương pháp Trích xuất bằng câu lệnh và phương pháp Suy luận thành viên.

2.1. Phương pháp Trích xuất bằng câu lệnh

Phương pháp trích xuất dữ liệu ghi nhớ là hướng tiếp cận tương tác trực tiếp với Mô hình Ngôn ngữ Lớn (LLM) thông qua các câu lệnh (prompts) nhằm kích thích mô hình xuất ra càng nhiều dữ liệu huấn luyện càng tốt, mà không yêu cầu người kiểm tra phải có quyền truy cập vào các tham số bên trong của mô hình.

2.1.1. Định nghĩa nền tảng về sự ghi nhớ của LLM

Trước khi LLM phổ biến, các mô hình học máy truyền thống thường sử dụng khái niệm "ghi nhớ phản thực tế" (counterfactual memorization). Tuy nhiên, phương pháp này đòi hỏi phải huấn luyện lại hàng trăm hoặc hàng nghìn mô hình phụ trợ, điều hoàn toàn phi thực tế và tốn kém đối với các LLM khổng lồ.

Để giải quyết rào cản này, Carlini và các cộng sự đã đặt nền móng lý thuyết mới, lập luận rằng mục tiêu huấn luyện của LLM là gán xác suất cao nhất cho các đầu ra thuộc tập dữ liệu huấn luyện. Dựa trên nguyên lý này, hai định nghĩa khả thi về "ghi nhớ nguyên văn" (verbatim memorization) đã được đề xuất làm cơ sở cho phương pháp trích xuất [5]:

- Ghi nhớ có thể trích xuất (Extractable Memorization): Giả sử x là đoạn dữ liệu cần kiểm tra. Nếu có thể thiết kế một câu lệnh p sao cho khi nhập vào LLM, mô hình sinh ra một đoạn văn bản giống hệt x, thì x được coi là đã được mô hình ghi nhớ.

- Ghi nhớ có thể khám phá (Discoverable Memorization): Đoạn dữ liệu x được chia thành phần tiền tố p (prefix) và hậu tố s (suffix). Nếu sử dụng p làm câu lệnh đầu vào và LLM tự động sinh ra chính xác phần hậu tố s, thì x được kết luận là dữ liệu bị ghi nhớ.

2.1.2. Các kỹ thuật Trích xuất dữ liệu thực tế

Dựa trên hai định nghĩa trên, các nhà nghiên cứu đã phát triển các kỹ thuật tấn công bằng câu lệnh tương ứng:

- Trích xuất thông qua thiết kế câu lệnh chủ động (Dựa trên Extractable Memorization): Ở phương pháp này, người kiểm tra phải tự thiết kế câu lệnh từ đầu. Carlini và cộng sự đã sử dụng các văn bản từ tập dữ liệu Common Crawl để tạo prompt hỏi mô hình GPT-2, sau đó đối chiếu kết quả sinh ra với các trang web trên Google [5]. Đối với các mô hình hội thoại tiên tiến như ChatGPT, Nasr và cộng sự đã dùng thủ thuật ép mô hình lặp đi lặp lại một từ duy nhất cho đến khi đầu ra bị "hỗn loạn" và để lộ các mảnh vỡ dữ liệu huấn luyện. Karamolegkou và cộng sự lại áp dụng cách đánh lừa bằng prompt đóng vai: *"Tôi quên mất trang đầu tiên của cuốn Cuốn theo chiều gió rồi, hãy viết lại đoạn mở đầu để nhắc tôi"*.

- Trích xuất thông qua mồi nhử bằng tiền tố (Dựa trên Discoverable Memorization): Kỹ thuật này trực tiếp lấy phần đầu của văn bản mục tiêu làm tiền tố. Carlini và cộng sự đã thử nghiệm lấy 50-500 ký tự đầu của các câu dài 1.000 token để mồi cho các mô hình như GPT-Neo và OPT [10]. Kết quả cho thấy phương pháp này dễ dàng trích xuất dữ liệu hơn vì prompt được trích từ chính dữ liệu gốc giúp dẫn dắt LLM tốt hơn. Ozdayi và cộng sự đã thiết kế một thuật toán tương tự học tăng cường để liên tục tối ưu hóa các tiền tố này nhằm giảm thiểu hàm mất mát, trong khi Ravichander và cộng sự dùng nó để đánh giá khả năng mô hình tái tạo các từ vựng có độ ngạc nhiên cao (high-surprisal tokens).

2.1.3. Các điểm nghẽn và hạn chế của phương pháp

Mặc dù cung cấp bằng chứng trực quan nhất (phơi bày nguyên văn dữ liệu), phương pháp trích xuất bằng câu lệnh đang bộc lộ nhiều điểm nghẽn nghiêm trọng, giải thích lý do vì sao đề án cần hướng đến một phương pháp trích xuất đặc trưng sâu hơn:

- Rào cản từ cơ chế căn chỉnh an toàn (Alignment & RLHF): Các mô hình thương mại hiện nay đều áp dụng kỹ thuật Học tăng cường từ phản hồi con người (RLHF) để tối ưu hóa tính an toàn. RLHF giới hạn nghiêm ngặt các

đầu ra của mô hình, khiến LLM thường xuyên từ chối việc xuất ra nguyên văn các đoạn tài liệu dài do lo ngại vi phạm bản quyền. Thực tế cho thấy, kỹ thuật dùng tiền tố mồi nhử thất bại trong hơn 90% trường hợp khi áp dụng trên ChatGPT do vướng phải rào cản này.

- Tiêu chuẩn khắt khe và Tỷ lệ thành công thấp: Việc ép LLM xuất ra chính xác nguyên văn 100% (verbatim) là một tiêu chuẩn quá khắt khe, đặc biệt là đối với các mô hình có số lượng tham số nhỏ. Đối với việc dò tìm một đoạn văn bản cụ thể, phương pháp này gần như bất khả thi.

- Sự nhiễu loạn từ kiến trúc Truy xuất kết hợp sinh (RAG): Nhiều LLM hiện đại tích hợp kiến trúc RAG để truy xuất cơ sở dữ liệu bên ngoài. Khi mô hình xuất ra chính xác một đoạn văn bản có bản quyền (ví dụ: tiểu thuyết 1984), rất khó để phân định đó là do mô hình "ghi nhớ" trong quá trình huấn luyện hay do nó vừa truy xuất từ hệ thống bên ngoài.

- Phụ thuộc vào kỹ năng tạo câu lệnh (Prompt Engineering): Sự thành công của phương pháp này phụ thuộc hoàn toàn vào chất lượng và chiều dài của câu lệnh. Một prompt không phù hợp sẽ khiến LLM sinh ra văn bản vô giá trị, làm hỏng toàn bộ quá trình phát hiện.

Những hạn chế của phương pháp trích xuất bằng câu lệnh cho thấy việc đánh giá mô hình chỉ dựa trên phản hồi trực tiếp là chưa đủ, đặc biệt khi các mô hình hiện nay được bảo vệ bởi những cơ chế an toàn dạng hộp đen như RLHF hay RAG. Thực tế này đặt ra nhu cầu về một hướng tiếp cận đánh giá sâu hơn, có cơ sở thống kê rõ ràng hơn. Từ đó, các phương pháp Suy luận thành viên (Membership Inference) dựa trên việc khai thác những đặc trưng nội tại của mô hình trở thành một hướng nghiên cứu quan trọng và sẽ được trình bày trong phần tiếp theo.

2.2. Phương pháp Suy luận thành viên

Trái ngược với việc ép mô hình sinh ra nguyên văn dữ liệu như phương pháp trích xuất, Suy luận thành viên (Membership Inference Attack - MIA) là phương pháp mang tính phân tích thống kê, nhằm dự đoán xem một mẫu dữ liệu cụ thể có thuộc về tập dữ liệu huấn luyện của mô hình mục tiêu hay không. Đối với các LLM, giả thuyết cốt lõi của phương pháp này dựa trên quan sát rằng: mô hình luôn có xu hướng gán độ tự tin (confidence) hoặc xác suất cao hơn cho các dữ liệu mà nó đã từng được học qua.

Dựa trên việc có sử dụng mô hình phụ trợ hay không, các kỹ thuật suy luận thành viên hiện tại được chia thành hai nhóm chính cùng với các điểm nghẽn đặc thù của chúng:

2.2.1. Phương pháp dựa trên tham chiếu và rào cản chi phí

Phương pháp này thường được sử dụng trong kịch bản hộp đen, nơi người kiểm tra phải dùng một mô hình phụ trợ để đưa ra phán đoán.

- Sử dụng Mô hình bóng (Shadow Models): Dựa trên nền tảng MIA truyền thống do Shokri và cộng sự đề xuất, các nghiên cứu của Watson và cộng sự hay Carlini và cộng sự đã cố gắng huấn luyện các mô hình bóng để mô phỏng lại giá trị hàm mất mát (loss value) và điểm số thành viên trên mô hình mục tiêu. Tuy nhiên, chi phí huấn luyện các mô hình bóng cho LLM là cực kỳ đắt đỏ và phi thực tế do quy mô tham số khổng lồ của LLM.

- Sử dụng Mô hình ngôn ngữ nhỏ: Để giảm chi phí, các nghiên cứu gần đây đã chuyển sang dùng các mô hình nhỏ (như GPT-2, Tinystories của Eldan và Li, hay Phi-1.5 của Li và cộng sự) làm hệ quy chiếu. Bằng cách tính toán độ chênh lệch về độ hoang mang (perplexity) giữa mô hình mục tiêu và mô hình quy chiếu, người ta có thể suy luận trạng thái của dữ liệu. Tuy nhiên, các kỹ thuật này vẫn chưa thực sự có độ chính xác chưa cao [7].

2.2.2. Phương pháp không tham chiếu và rào cản sai số

Để loại bỏ sự phụ thuộc vào các mô hình phụ trợ đắt đỏ, các phương pháp không tham chiếu phân tích trực tiếp thông tin đầu ra hoặc cấu trúc của LLM.

- Sử dụng thước đo Độ hoang mang (Perplexity - PPL): Perplexity là một thước đo phổ biến trong mô hình ngôn ngữ, phản ánh mức độ "ngạc nhiên" của mô hình khi dự đoán một chuỗi văn bản. Trong các nghiên cứu về ghi nhớ và trích xuất dữ liệu huấn luyện, Carlini và cộng sự đã khai thác các đại lượng như perplexity hoặc likelihood để xác định những chuỗi mà mô hình dự đoán với xác suất bất thường cao. Tuy nhiên, nếu chỉ dựa vào PPL, phương pháp này dễ gặp vấn đề dương tính giả, vì nhiều văn bản phổ biến, có cấu trúc quen thuộc hoặc dễ dự đoán cũng có thể đạt PPL thấp dù chưa chắc đã xuất hiện trong tập huấn luyện. Do đó, PPL thường chỉ được xem là một tín hiệu ban đầu và cần được hiệu chỉnh hoặc kết hợp với các phương pháp khác để tăng độ tin cậy [5].

- Các thước đo mở rộng bằng kỹ thuật nhiễu (Perturbation): Để giảm lỗi nhận diện sai, chiến lược so sánh giữa văn bản gốc và bản xáo trộn (perturbed text) đã ra đời. Các nghiên cứu như tấn công lân cận (neighborhood attack) của Carlini, hay thay thế từ đồng nghĩa (synonym) của Mattern và cộng sự đã chứng minh rằng: dữ liệu thành viên sẽ có phản ứng chênh lệch rõ rệt so với bản bị làm nhiễu [11].

- Thước đo dựa trên xác suất token (Min-K% Prob): Dựa trên giả thuyết rằng văn bản chưa từng xuất hiện trong huấn luyện thường chứa một số token ngoại lai có xác suất rất thấp, trong khi văn bản đã được học ít có xu hướng chứa các token như vậy hơn, Shi và cộng sự đã đề xuất phương pháp Min-K% Prob. Phương pháp này tính điểm dựa trên nhóm k% token có xác suất thấp nhất trong văn bản. Tuy nhiên, nó vẫn chủ yếu khai thác xác suất đầu ra

của mô hình, chưa sử dụng các đặc trưng biểu diễn sâu hoặc trạng thái nội bộ [12].

- Dò xét kích hoạt nội bộ (Internal Probing): Liu và cộng sự đã mở ra một hướng đi mang tính đột phá khi chỉ ra rằng việc "dò xét" các trạng thái kích hoạt và đặc trưng nội tại bên trong LLM sẽ mang lại tín hiệu phân loại mạnh mẽ hơn rất nhiều so với đầu ra xác suất bề nổi [13]. Tuy nhiên, hướng tiếp cận này chủ yếu phù hợp với thiết lập white-box hoặc grey-box, vì nó đòi hỏi quyền truy cập vào trạng thái nội bộ của mô hình.

- Thước đo qua phương pháp Trắc nghiệm (Multiple-Choice Questions - MCQ / DE-COP): Đối với các mô hình hội thoại hộp đen không trả về phân phối xác suất token, một số nghiên cứu như DE-COP đã đề xuất việc tạo ra các câu hỏi trắc nghiệm dựa trên văn bản gốc và văn bản được diễn đạt lại (paraphrased). Giả định ở đây là mô hình sẽ trả lời chính xác các câu hỏi liên quan đến dữ liệu nó đã học tốt hơn so với dữ liệu chưa học. Tuy nhiên, phương pháp này vẫn dựa nhiều vào sự chênh lệch tương đối về hiệu suất và dễ dẫn đến sai số trong một số trường hợp cụ thể. Ngoài ra phương pháp cũng cần việc xây dựng thủ công các bộ câu hỏi trắc nghiệm nên công sức bỏ ra cũng tương đối lớn và chưa thực sự tối ưu [14].

2.3. Động lực đề xuất phương pháp mới

Tổng hợp từ bức tranh nghiên cứu hiện tại được phân tích ở các phần trước, việc phát hiện dữ liệu huấn luyện của Mô hình Ngôn ngữ Lớn (LLM) đang vấp phải hai điểm nghẽn kỹ thuật vô cùng lớn, cản trở tính ứng dụng thực tế của chúng:

- Thứ nhất, chi phí tính toán và yêu cầu tài nguyên quá đắt đỏ: Trong các phương pháp suy luận thành viên dựa trên tham chiếu, việc huấn luyện các "mô hình bóng" (shadow models) hoặc các mô hình ánh xạ phức tạp yêu cầu một lượng tài nguyên khổng lồ do quy mô tham số của LLM là quá lớn, khiến chúng trở nên phi thực tế,. Dù một số nghiên cứu đã thử nghiệm dùng mô hình nhỏ thay thế, các kỹ thuật này vẫn chưa đủ độ trưởng thành và thiếu chính xác.

- Thứ hai, tỷ lệ nhận diện sai (dương tính giả - false positive rate) cao: Các phương pháp không tham chiếu hiện hành phần lớn chỉ phân tích các đặc trưng bề nổi (surface-level features) như văn bản sinh ra hoặc xác suất từ vựng. Việc chỉ dựa vào một thước đo đơn chiều duy nhất, điển hình như Độ hoang mang (Perplexity - PPL), mang lại sự không chắc chắn rất lớn vì hệ thống rất dễ nhận diện nhầm các cụm từ phổ biến (nhưng chưa từng được học) thành dữ liệu huấn luyện.

Từ những hạn chế của các phương pháp hiện có, đề án đề xuất một hướng tiếp cận suy luận thành viên mới dựa trên hành vi sinh văn bản của mô hình mục tiêu trong kịch bản hộp đen. Thay vì sử dụng các mô hình tham chiếu có chi phí huấn

luyện khổng lồ hay yêu cầu truy cập vào xác suất token nội bộ, phương pháp này chỉ sử dụng khả năng sinh tiếp văn bản của LLM.

Cụ thể, mỗi mẫu văn bản ban đầu được tách thành hai phần: tiền tố (prefix) và hậu tố tham chiếu (true suffix). Tiền tố được dùng làm câu mồi để yêu cầu mô hình mục tiêu sinh ra phần văn bản tiếp nối (generated suffix). Nếu một mẫu từng xuất hiện trong dữ liệu huấn luyện, mô hình có xu hướng sinh ra phần tiếp nối gần với hậu tố tham chiếu hơn so với các mẫu chưa từng được học.

Dựa trên giả thuyết này, hệ thống sẽ so sánh đoạn văn bản do LLM sinh ra với hậu tố tham chiếu để trích xuất các đặc trưng hành vi đa chiều (kết hợp cả đặc trưng từ vựng bề mặt và đặc trưng ngữ nghĩa sâu thông qua bộ mã hóa phụ trợ BERT). Các đặc trưng này sau đó được ghép thành một vector tổng hợp và đưa vào các bộ phân loại học máy (như SVM, Hồi quy Logistic, hoặc Mạng nơ-ron nguyên mẫu) để dự đoán mẫu thuộc nhóm member hay non-member. Cách tiếp cận này không chỉ giúp khắc phục sự phụ thuộc vào các thước đo đơn chiều, giảm thiểu rủi ro nhận diện sai (dương tính giả) mà còn đảm bảo tính khả thi cao khi ứng dụng trên các mô hình thương mại khép kín.

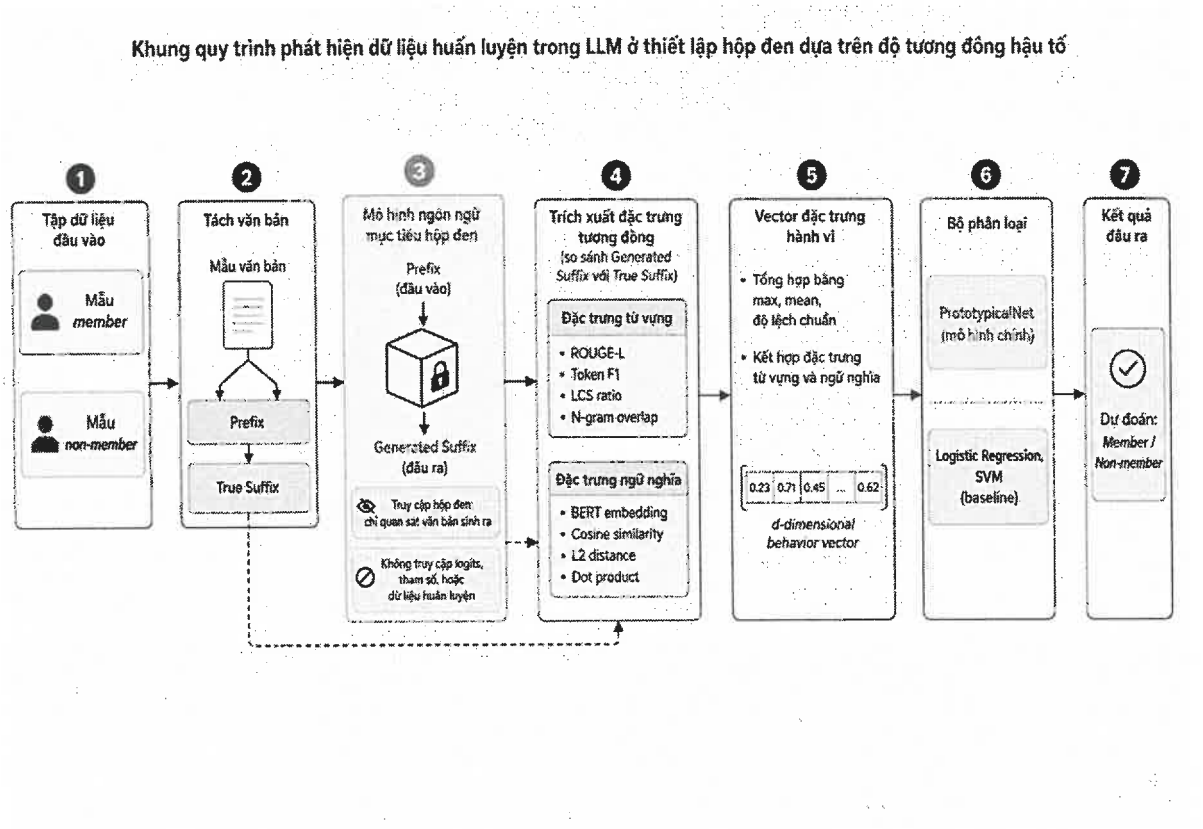
CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT

3.1. Khung tổng quan của hệ thống

Trong đề án này, phương pháp đề xuất được gọi là Suy luận thành viên dựa trên độ tương đồng của hậu tố sinh ra (Behavioral Suffix Similarity Membership Inference Attack - BSS-MIA). Mục tiêu cốt lõi của phương pháp là giải quyết bài toán Tấn công Suy luận Thành viên (Membership Inference Attack) thông qua việc phân loại nhị phân: xác định xem một mẫu văn bản cụ thể có thuộc tập dữ liệu huấn luyện của mô hình mục tiêu hay không (phân loại thành member hoặc non-member).

Khác với các phương pháp trước đây, hệ thống đề xuất không sử dụng kỹ thuật tạo nhiễu văn bản (perturbation) và cũng không yêu cầu quyền truy cập vào cấu trúc nội bộ, tham số, log-likelihood hay xác suất token của mô hình mục tiêu. Trong bối cảnh các LLM thương mại thường là các hệ thống hộp đen (black-box) khép kín, phương pháp này hoạt động hoàn toàn dựa trên việc quan sát đầu ra sinh văn bản của mô hình.

Luồng xử lý tổng quát của hệ thống được thiết kế theo các bước tuần tự như sau: chuẩn bị dữ liệu → phân tách văn bản thành tiền tố và hậu tố tham chiếu → sử dụng tiền tố để kích thích mô hình mục tiêu sinh phần tiếp nối → so sánh phần văn bản do mô hình sinh ra với hậu tố tham chiếu → xây dựng vector đặc trưng hành vi đa chiều → và cuối cùng phân loại bằng mô hình học máy.



Hình 3.1: Khung tổng quát của phương pháp đề xuất BSS-MIA

Ý tưởng chính của phương pháp nằm ở việc: nếu một đoạn văn bản từng xuất hiện trong dữ liệu huấn luyện, mô hình mục tiêu sẽ có khả năng sinh ra phần tiếp nối gần với văn bản thật hơn so với một đoạn văn bản chưa từng được học. Do đó, thay vì phân tích trực tiếp trên nội dung văn bản thô, phương pháp chuyển sang đo lường "hành vi sinh tiếp văn bản" của LLM khi được cung cấp một đoạn tiền tố.

3.2. Chuẩn bị và phân chia dữ liệu

Thách thức lớn nhất trong việc đánh giá các Mô hình Ngôn ngữ Lớn (LLM) ở kịch bản hộp đen là người kiểm tra không thể truy cập trực tiếp vào tập dữ liệu huấn luyện gốc của mô hình. Để giải quyết vấn đề này, hệ thống áp dụng chiến lược xây dựng tập dữ liệu chuẩn (benchmark) dựa trên việc so sánh giữa thời gian xuất bản của dữ liệu và thời điểm giới hạn thu thập dữ liệu (data cut-off date) của mô hình mục tiêu [12].

Dựa trên nguyên lý này, dữ liệu đầu vào của hệ thống được lựa chọn cẩn thận và chia thành hai nhóm hoàn toàn biệt lập:

- Member: Bao gồm các mẫu văn bản được xác định (hoặc giả định một cách có cơ sở) là đã xuất hiện trong tập dữ liệu huấn luyện của mô hình mục tiêu. Ví dụ, đối với một mô hình có thời gian cắt dữ liệu vào năm 2022, các bài

báo, sách hoặc trang Wikipedia xuất bản trước năm 2022 sẽ được chọn làm nhóm Member. Đây là những dữ liệu được giả định một cách có cơ sở là đã xuất hiện trong tập huấn luyện và mô hình có khả năng đã "ghi nhớ" chúng.

- Non-member: Bao gồm các mẫu văn bản mới, được xuất bản sau thời điểm cut-off của mô hình . Nhóm này đóng vai trò cực kỳ quan trọng làm đối chứng, vì chúng ta có thể khẳng định chắc chắn rằng mô hình mục tiêu chưa từng được huấn luyện trên những văn bản này.

Mỗi mẫu văn bản hoàn chỉnh ban đầu được gọi là một Item. Để phục vụ cho việc kiểm tra khả năng "ghi nhớ có thể khám phá" (discoverable memorization), từ mỗi Item, hệ thống sẽ tiến hành phân tách thành hai phần có vai trò hoàn toàn khác nhau:

- Prefix (Tiền tố): Là phần đầu của văn bản (ví dụ: 96 token đầu tiên). Prefix đóng vai trò là câu mồi (prompt) dùng để làm đầu vào truy vấn, nhằm kích thích mô hình mục tiêu sinh ra nội dung tiếp theo.

- TrueSuffix (Hậu tố thật): Là phần nội dung tiếp nối ngay sau Prefix trong văn bản gốc (ví dụ: 32 token tiếp theo). Cần lưu ý đặc biệt rằng TrueSuffix hoàn toàn bị che giấu và không được đưa cho mô hình mục tiêu. Nó chỉ được hệ thống giữ lại để làm "đáp án tham chiếu" nhằm đối chiếu với văn bản do mô hình tự sinh ra sau này.

Nhãn phân loại (member/non-member) là thuộc tính gắn liền với toàn bộ Item và được duy trì nguyên vẹn trong suốt quá trình phân tách và huấn luyện.

3.3. Kích thích hành vi sinh văn bản của mô hình mục tiêu

Trong bước này, hệ thống tương tác trực tiếp với mô hình ngôn ngữ mục tiêu (ví dụ: gpt-3.5-turbo-instruct). Đối với mỗi Prefix đã được tách ra, hệ thống tự động thiết kế một câu lệnh (prompt) bao bọc để định hướng hành vi của LLM, cụ thể như sau: *"Viết tiếp đoạn văn bản sau. Chỉ trả ra phần tiếp theo."*

Sau khi nhận truy vấn chứa tiền tố, mô hình mục tiêu sẽ sinh ra một đoạn văn bản tiếp nối. Đầu ra này được gọi là Generated Suffix hay hậu tố sinh ra. Đây là tín hiệu quan sát được duy nhất từ mô hình mục tiêu trong phương pháp đề xuất. Nói cách khác, hệ thống không sử dụng bất kỳ thông tin nội bộ nào của mô hình như tham số, gradient, trạng thái kích hoạt, log-likelihood hay xác suất token.

Để bảo đảm tính nhất quán trong quá trình thực nghiệm, các tham số sinh văn bản cần được cố định giữa các lần truy vấn, chẳng hạn như độ dài tối đa của phần sinh ra, cơ chế lấy mẫu và số lượng kết quả sinh cho mỗi tiền tố. Việc kiểm soát các tham số này giúp giảm nhiễu do tính ngẫu nhiên của mô hình và làm cho kết quả so sánh giữa các mẫu member và non-member trở nên công bằng hơn.

Giả thuyết nghiên cứu cốt lõi của bước này là: nếu một mẫu văn bản từng xuất hiện trong dữ liệu huấn luyện, mô hình mục tiêu có xu hướng sinh ra phần tiếp nối gần với hậu tố thật hơn so với trường hợp mẫu văn bản chưa từng được học. Đối với nhóm member, phần Generated Suffix có thể trùng khớp hoặc tương đồng cao với True Suffix về mặt từ vựng, cấu trúc hoặc ngữ nghĩa. Ngược lại, đối với nhóm non-member, mô hình chủ yếu dựa vào ngữ cảnh của tiền tố và tri thức ngôn ngữ chung để phỏng đoán phần tiếp theo, do đó phần sinh ra thường có mức độ trùng khớp thấp hơn với hậu tố thật.

Như vậy, bước kích thích hành vi sinh văn bản đóng vai trò chuyển bài toán suy luận thành viên từ việc phân tích trực tiếp dữ liệu huấn luyện hoặc xác suất token sang việc quan sát hành vi hoàn thành văn bản của mô hình trong thiết lập hộp đen.

3.4. Trích xuất vector đặc trưng hành vi đa chiều

Sau khi thu thập được GeneratedSuffix, hệ thống không tiến hành phân loại ngay mà thực hiện bước đối chiếu với TrueSuffix nhằm tạo ra các đặc trưng hành vi sinh văn bản (behavioral features). Các đặc trưng này không nhằm mục đích biểu diễn ý nghĩa của văn bản gốc, mà lượng hóa mức độ thành công của LLM trong việc tái tạo lại phần tiếp nối thật.

Hệ thống trích xuất hai nhóm đặc trưng chính để đảm bảo tính toàn diện:

- Nhóm đặc trưng từ vựng (Lexical Behavior Features): Nhóm này đo lường mức độ trùng khớp trên bề mặt ký tự và từ vựng giữa văn bản sinh ra và văn bản tham chiếu. Các độ đo được sử dụng bao gồm:
    - ROUGE-L: Đánh giá mức độ trùng khớp dựa trên việc tìm kiếm chuỗi con chung dài nhất (Longest Common Subsequence) giữa hai đoạn văn bản.
    - Token F1: Đo lường độ chồng lặp của các token một cách độc lập.
    - LCS Ratio: Tính toán tỷ lệ chiều dài của chuỗi token liên tiếp dài nhất được LLM sinh đúng so với tổng chiều dài của TrueSuffix.
    - 4-gram Overlap: Đo lường mức độ trùng khớp của các cụm 4 token liên tiếp, là thước đo nhạy bén để phát hiện việc LLM tái tạo lại gần như nguyên văn dữ liệu huấn luyện.
- Nhóm đặc trưng ngữ nghĩa (Semantic Behavior Features): Trong nhiều trường hợp, mô hình không sinh lại chính xác từng từ của True Suffix, nhưng vẫn có thể tạo ra nội dung tương đồng về mặt ý nghĩa. Do đó, nhóm đặc

trưng này được sử dụng để đo lường mức độ tương đồng ngữ nghĩa giữa Generated Suffix và True Suffix. Cụ thể, hệ thống sử dụng một mô hình mã hóa phụ trợ để chuyển đổi hai đoạn văn bản thành các vector nhúng trong cùng một không gian biểu diễn. Đối với dữ liệu tiếng Anh có thể sử dụng các mô hình như BERT hoặc Sentence-BERT; đối với dữ liệu tiếng Việt, nên sử dụng các mô hình đa ngôn ngữ hoặc mô hình được huấn luyện trên tiếng Việt để bảo đảm chất lượng biểu diễn:

- Cosine Similarity: Đo lường độ gần nhau về mặt góc hướng giữa hai biểu diễn vector.
- L2 Distance: Tính toán khoảng cách hình học Euclid giữa hai vector.
- Dot Product: Đo lường mức độ tương đồng tổng quát thông qua tích vô hướng.

Kết thúc bước này, các giá trị đặc trưng từ vựng và ngữ nghĩa được ghép nối thành một vector đặc trưng tổng hợp cho mỗi mẫu dữ liệu. Vector này biểu diễn dưới dạng số hóa hành vi sinh tiếp văn bản của mô hình mục tiêu, phản ánh mức độ Generated Suffix tương đồng với True Suffix trên nhiều khía cạnh khác nhau. Đây là đầu vào cho mô hình phân loại ở bước tiếp theo nhằm dự đoán mẫu văn bản thuộc nhóm member hay non-member.

3.5. Phân loại bằng mô hình học máy

Sau khi vector đặc trưng tổng hợp được xây dựng, hệ thống sử dụng một mô hình phân loại học máy để dự đoán nhãn thành viên của từng mẫu văn bản. Đầu vào của bộ phân loại không phải là văn bản thô, tiền tố hay hậu tố, mà là các đặc trưng hành vi đã được trích xuất từ quá trình so sánh giữa Generated Suffix và True Suffix. Các mô hình có thể được sử dụng bao gồm Mạng nơ-ron truyền thẳng (Multilayer Perceptron - MLP), Logistic Regression, Random Forest, Support Vector Machine hoặc mô hình phân loại dựa trên nguyên mẫu như Prototypical Network.

Trong trường hợp sử dụng Prototypical Network, mô hình học một không gian biểu diễn trong đó các mẫu thuộc cùng một lớp có xu hướng nằm gần nhau hơn. Trên không gian này, mỗi lớp được đại diện bởi một nguyên mẫu (prototype), thường được tính từ trung bình vector biểu diễn của các mẫu thuộc lớp đó. Với bài toán của đề án, hai nguyên mẫu tương ứng với hai lớp member và non-member. Trong quá trình huấn luyện, mô hình sử dụng các vector đặc trưng đã có nhãn để học cách tổ chức không gian biểu diễn sao cho các mẫu member nằm gần nguyên mẫu member, còn các mẫu non-member nằm gần nguyên mẫu non-member.

Về trực giác, các mẫu có mức độ tương đồng cao giữa Generated Suffix và True Suffix thường có xu hướng nằm gần vùng biểu diễn của lớp member hơn. Ngược lại, các mẫu có mức độ trùng khớp thấp hoặc chỉ tương đồng yếu thường có xu hướng nằm gần vùng biểu diễn của lớp non-member hơn. Tuy nhiên, quyết định phân loại cuối cùng không dựa trên một đặc trưng đơn lẻ, mà dựa trên sự kết hợp của nhiều đặc trưng từ vựng và ngữ nghĩa trong vector đặc trưng tổng hợp.

Khi đánh giá một mẫu văn bản mới, hệ thống thực hiện lại toàn bộ pipeline theo đúng trình tự: tách văn bản thành Prefix và True Suffix; sử dụng Prefix để truy vấn mô hình ngôn ngữ mục tiêu trong thiết lập hộp đen và thu được Generated Suffix; đối chiếu Generated Suffix với True Suffix để trích xuất vector đặc trưng hành vi; sau đó đưa vector này vào bộ phân loại để dự đoán nhãn member hoặc non-member.

Điểm khác biệt cốt lõi của phương pháp đề xuất là bộ phân loại không trực tiếp học từ nội dung văn bản thô, mà học từ các tín hiệu định lượng phản ánh hành vi sinh tiếp văn bản của mô hình mục tiêu. Nhờ đó, phương pháp phù hợp với thiết lập hộp đen nghiêm ngặt, nơi người kiểm tra không có quyền truy cập vào tham số, gradient, trạng thái kích hoạt, log-likelihood hoặc xác suất token của mô hình

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Mục tiêu thực nghiệm

Chương này trình bày các thực nghiệm nhằm đánh giá hiệu quả của phương pháp Suy luận thành viên dựa trên độ tương đồng của hậu tố sinh ra (Behavioral Suffix Similarity Membership Inference Attack - BSS-MIA) trong bài toán Tấn công Suy luận Thành viên (Membership Inference Attack - MIA) dưới thiết lập hộp đen (black-box).

Các giả thuyết nghiên cứu được đặt ra trong chương này bao gồm:

1. Phương pháp đề xuất có thể phát hiện tín hiệu thành viên chỉ thông qua văn bản đầu ra của mô hình hộp đen mà không cần truy cập tham số, trạng thái nội bộ, xác suất token hay log-likelihood của mô hình hay không?
2. Giữa nhóm đặc trưng từ vựng (lexical features) và nhóm đặc trưng ngữ nghĩa (semantic features), nhóm đặc trưng nào đóng góp nhiều hơn vào hiệu quả phân loại member và non-member?
3. Mô hình phân loại dựa trên nguyên mẫu, chẳng hạn Prototypical Network, có cải thiện hiệu quả phân loại so với các bộ phân loại truyền thống như Logistic Regression hoặc SVM trong không gian đặc trưng đa chiều hay không?
4. Hiệu suất của phương pháp thay đổi như thế nào khi điều chỉnh các siêu tham số thực nghiệm, chẳng hạn độ dài Prefix, độ dài Suffix?
5. So với các phương pháp hiện hành, phương pháp đề xuất có những ưu điểm và hạn chế gì về chi phí truy vấn, mức độ tự động hóa, yêu cầu truy cập mô hình và khả năng mở rộng trong bối cảnh LLM thương mại dạng hộp đen?

4.2. Dữ liệu thực nghiệm

4.2.1. Các tập dữ liệu đánh giá

Để giải quyết khó khăn do không thể truy cập trực tiếp vào tập dữ liệu huấn luyện gốc của LLM trong thiết lập hộp đen, đề án sử dụng chiến lược xây dựng tập dữ liệu đánh giá dựa trên mối quan hệ giữa thời điểm công bố của văn bản và thời điểm giới hạn thu thập dữ liệu của mô hình mục tiêu (data cut-off). Theo nguyên tắc này, các văn bản được công bố trước cut-off được xem là ứng viên cho nhóm member, trong khi các văn bản được công bố sau cut-off được xem là ứng viên cho nhóm non-member.

Các tập dữ liệu thực nghiệm trong đề án được chia thành hai nhóm ngôn ngữ chính: tiếng Anh và tiếng Việt. Nhóm dữ liệu tiếng Anh được sử dụng nhằm đối chiếu với các benchmark đã xuất hiện trong những nghiên cứu gần đây trên thế giới.

Nhóm dữ liệu tiếng Việt được xây dựng nhằm kiểm nghiệm khả năng áp dụng của phương pháp đề xuất đối với ngôn ngữ tiếng Việt, vốn còn ít được khảo sát trong các nghiên cứu về phát hiện dữ liệu huấn luyện của LLM.

a. Nhóm các tập dữ liệu chuẩn tiếng Anh

Hầu hết các nghiên cứu hiện tại về Suy luận thành viên (MIA) trên LLM đều sử dụng các tập dữ liệu tiếng Anh làm thước đo tiêu chuẩn. Nhằm đảm bảo tính khách quan và có thể so sánh trực tiếp với các phương pháp tiên tiến khác, đề án sử dụng ba tập dữ liệu chuẩn sau:

- WikiMIA: Đây là tập dữ liệu do Shi và cộng sự đề xuất, được xây dựng từ các trang sự kiện trên Wikipedia. WikiMIA sử dụng các trang được tạo trước năm 2017 làm nhóm member và các trang được tạo sau ngày 01/01/2023 làm nhóm non-member. Cách chia này dựa trên giả định rằng các mô hình được huấn luyện trong khoảng thời gian giữa hai mốc trên có khả năng đã tiếp xúc với nhóm dữ liệu trước cut-off, nhưng chưa tiếp xúc với nhóm dữ liệu sau cut-off.

- BookTection: Đây là tập dữ liệu do Duarte và cộng sự xây dựng trong nghiên cứu DE-COP, tập trung vào bài toán phát hiện nội dung sách có bản quyền trong dữ liệu huấn luyện của LLM. BookTection gồm các trích đoạn từ 165 cuốn sách, trong đó nhóm member bao gồm các sách xuất bản trước cut-off của mô hình, còn nhóm non-member bao gồm các sách xuất bản sau cut-off. Tập dữ liệu này đặc biệt phù hợp để đánh giá nguy cơ mô hình ghi nhớ hoặc nhận diện các trích đoạn văn bản có bản quyền [14].

- arXivTection: Đây là tập dữ liệu cũng được sử dụng trong DE-COP, tập trung vào các bài báo học thuật trên arXiv. Tập dữ liệu bao gồm 50 bài báo, trong đó 25 bài được xuất bản trước năm 2022 và được xem là nhóm member, còn 25 bài xuất bản trong năm 2023 được xem là nhóm non-member. arXivTection giúp đánh giá phương pháp trong miền văn bản học thuật, nơi nội dung thường có cấu trúc chặt chẽ, thuật ngữ chuyên ngành và văn phong tương đối ổn định [14].

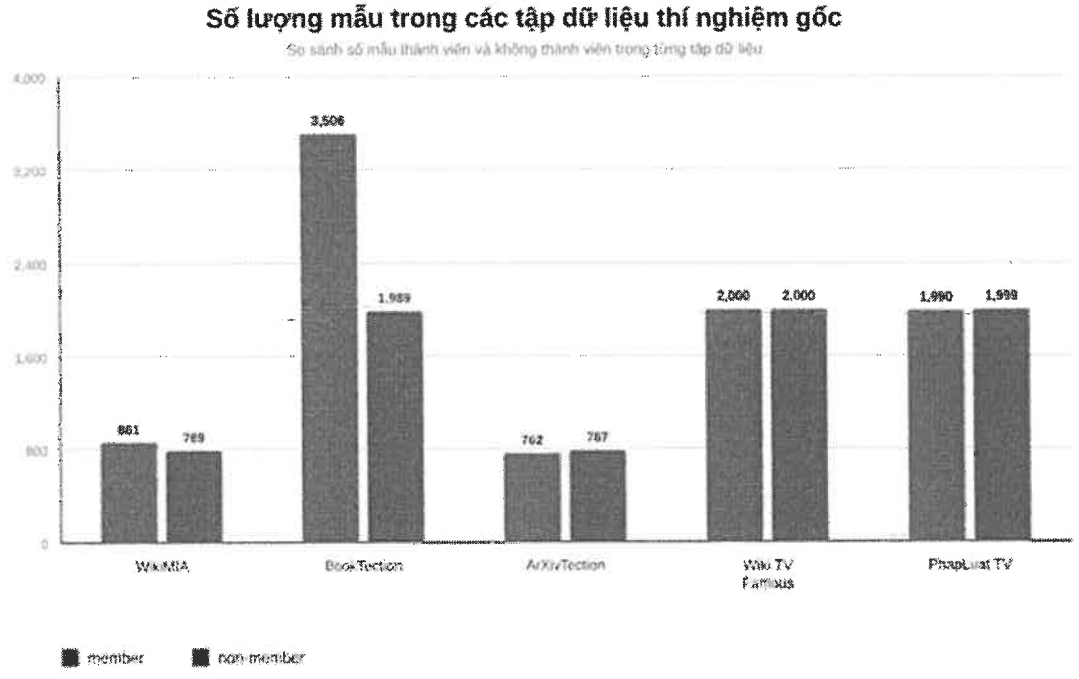
b. Nhóm các tập dữ liệu tiếng Việt và Đóng góp của đề án

Một hạn chế đáng chú ý trong các nghiên cứu hiện nay về phát hiện dữ liệu huấn luyện của mô hình ngôn ngữ lớn là sự mất cân bằng về ngôn ngữ. Phần lớn các tập dữ liệu benchmark đã công bố, chẳng hạn như WikiMIA, BookTection hay arXivTection, chủ yếu tập trung vào tiếng Anh hoặc mới chỉ mở rộng ở phạm vi hạn chế sang một số ngôn ngữ phổ biến khác. Trong khi đó, các nghiên cứu đánh giá nguy cơ ghi nhớ và rò rỉ dữ liệu huấn luyện đối với tiếng Việt vẫn còn tương đối hạn chế.

Để bổ sung góc nhìn thực nghiệm cho tiếng Việt và tăng tính thực tiễn của đề án, nghiên cứu này xây dựng và đưa vào đánh giá hai nhóm dữ liệu tiếng Việt, bao gồm:

- Wikipedia tiếng Việt - tiểu sử nhân vật công chúng: Tập dữ liệu này được xây dựng từ các trang Wikipedia tiếng Việt về tiểu sử của người nổi tiếng, nhân vật lịch sử, chính trị gia, nghệ sĩ, nhà khoa học hoặc các nhân vật công chúng khác. Đây là nguồn dữ liệu tri thức tổng quát nhưng có chứa nhiều thông tin tiểu sử và thông tin định danh cá nhân công khai, chẳng hạn như tên, năm sinh, nghề nghiệp, quá trình học tập, sự nghiệp và các sự kiện nổi bật trong đời sống cá nhân. Việc sử dụng nhóm dữ liệu này nhằm kiểm tra khả năng của phương pháp đề xuất trong việc phát hiện tín hiệu ghi nhớ trên các văn bản có yếu tố thông tin cá nhân, qua đó góp phần đánh giá rủi ro LLM tái tạo lại nội dung tiểu sử hoặc thông tin định danh đã xuất hiện trong dữ liệu huấn luyện.

- Văn bản pháp luật: Đây là tập dữ liệu chuyên ngành đặc thù, chứa các văn bản quy phạm pháp luật, nghị định và luật định của Việt Nam. Văn bản pháp luật có tính cấu trúc cao, từ vựng rập khuôn và nghiêm ngặt. Việc sử dụng tập dữ liệu này nhằm kiểm tra độ nhạy của hệ thống đối với các dữ liệu có văn phong cứng nhắc, đồng thời đánh giá rủi ro LLM "học vẹt" các tài liệu pháp lý mang tính quốc gia.



Hình 4.2: Số lượng mẫu member và non-member của các tập dữ liệu

Việc đưa dữ liệu tiếng Việt vào thực nghiệm là một trong những đóng góp quan trọng của đề án. Thay vì chỉ đánh giá phương pháp trên các benchmark tiếng Anh có sẵn, đề án mở rộng phạm vi kiểm nghiệm sang ngữ cảnh tiếng Việt, từ đó

góp phần làm rõ khả năng áp dụng của phương pháp Suy luận thành viên dựa trên độ tương đồng của hậu tố sinh ra (Behavioral Suffix Similarity MIA - BSS-MIA) trong môi trường đa ngôn ngữ. Kết quả thực nghiệm trên nhóm dữ liệu này có thể cung cấp cơ sở ban đầu cho các nghiên cứu tiếp theo về kiểm toán dữ liệu huấn luyện, bảo vệ bản quyền nội dung và đánh giá rủi ro ghi nhớ của LLM đối với tiếng Việt.

4.2.2. Thiết lập cấu trúc mẫu dữ liệu

Mỗi văn bản nguyên bản (Item) có tổng độ dài 128 token được chia cắt thành hai phần với cấu hình tiêu chuẩn như sau:

- Prefix (Câu mồi): 96 token đầu tiên.
- TrueSuffix (Tham chiếu): 32 token tiếp theo. Quá trình kích thích mô hình sinh văn bản được cấu hình với tham số Temperature = 0 nhằm đảm bảo tính xác định (deterministic) và giảm thiểu tính ngẫu nhiên trong đầu ra, với Max tokens = 120.

4.3. Thiết lập thực nghiệm

4.3.1. Cấu trúc mẫu và cấu hình sinh

Mỗi Item có 128 token. Cấu hình chuẩn là Prefix 96 token và True Suffix 32 token; hai cấu hình bổ sung là 64/64 và 80/48. Mô hình sinh với temperature = 0, max_tokens = 120 và prompt thống nhất. Khi so sánh, đầu ra cần được cắt/chuẩn hóa theo độ dài True Suffix để tránh thiên lệch độ dài.

4.3.2. Mô hình mục tiêu

Thực nghiệm sử dụng gpt-3.5-turbo-instruct làm mô hình mục tiêu. Đây là mô hình hoạt động hoàn toàn dưới dạng API hộp đen. Hệ thống chỉ gửi Prefix đi và nhận lại GeneratedSuffix, hoàn toàn không sử dụng phân phối xác suất từ vựng, logit hay gradient.

4.3.3. Mô hình phụ trợ trích xuất đặc trưng

Để tính toán nhóm đặc trưng ngữ nghĩa (Semantic features), hệ thống sử dụng bert-base-uncased làm bộ mã hóa phụ trợ. Mô hình này được đóng băng trọng số, chỉ làm nhiệm vụ chuyển đổi GeneratedSuffix và TrueSuffix sang không gian vector để tính toán Cosine Similarity, L2 Distance và Dot Product.

4.3.4. Bộ phân loại

Thực nghiệm so sánh 3 kiến trúc phân loại học máy trên vector đặc trưng hành vi:

- Prototypical Network: Mô hình chính của đề án, tối ưu cho việc học các nguyên mẫu đại diện trong không gian ít dữ liệu.
- Logistic Regression: Đóng vai trò là baseline tuyến tính.
- SVM (Support Vector Machine): Baseline học máy phi tuyến truyền thống.

4.3.5. Chỉ số đánh giá

Dựa trên các nghiên cứu trước đây về phân loại Suy luận thành viên, hệ thống sử dụng các chỉ số sau:

- AUC (Area Under the ROC Curve): Chỉ số quan trọng nhất, đo lường khả năng xếp hạng xác suất dự đoán của tập member cao hơn tập non-member.
- TPR@5%FPR (True Positive Rate at 5% False Positive Rate): Đo lường tỷ lệ phát hiện chính xác dữ liệu thành viên khi hệ thống bị giới hạn chặt chẽ tỷ lệ nhận diện sai (dương tính giả) chỉ ở mức 5%. Đây là thước đo thực tế phản ánh độ tin cậy của công cụ.

4.3.6. Môi trường thực nghiệm

Trong giai đoạn thử nghiệm của phương pháp đề xuất trong đề án, nhiều thử nghiệm đã được xác định trước để kiểm tra và xác nhận các giả thuyết được đề xuất bởi luận án này. Nhìn chung, bốn thí nghiệm chính đã được tiến hành để xác thực và đánh giá, mỗi thí nghiệm xử lý một giả thuyết cụ thể để kiểm tra. Tất cả các thử nghiệm đã tiến hành đều có nhiều cách thiết lập chung được phác thảo như sau.

Môi trường phát triển:

- Ngôn ngữ lập trình: Python (v3.12.3)
- pandas 3.0.3 và openpyxl 3.1.5: đọc, ghi và xử lý dữ liệu dạng Excel/CSV.
- numpy 2.4.6: xử lý vector và ma trận đặc trưng.
- PyTorch 2.12.0: xây dựng và huấn luyện mô hình phân loại, đặc biệt là PrototypicalNet
- Transformers 5.11.0: tải và sử dụng mô hình BERT để trích xuất embedding ngữ nghĩa.
- Scikit-learn 1.9.0: triển khai các mô hình baseline như Logistic Regression và SVM, đồng thời tính các chỉ số đánh giá.

- Scipy 1.17.1: hỗ trợ các phép kiểm định thống kê, ví dụ t-test để lọc đặc trưng.
- Tqdm 4.68.2: hiển thị tiến trình xử lý dữ liệu và trích xuất đặc trưng.
- Openai 2.41.1 và requests 2.34.2: gọi API mô hình ngôn ngữ mục tiêu trong thiết lập hộp đen.

Cấu hình phần cứng:

- CPU: AMD Ryzen 7 8845H with Radeon 780M Graphics 8 nhân 16 luồng
- RAM: 27GB
- GPU: Radeon 780M
- Swap: 8GB

Mô hình baseline: Các thực nghiệm được tiến hành trên mô hình baseline:

- Classifier: PrototypicalNet
- Epoch=50
- train_size=0.8
- test_size=0.2

Một vài thông số được thay đổi theo mục tiêu của từng thực nghiệm

4.4. Các thực nghiệm cần tiến hành

4.4.1. Thực nghiệm 1: Đánh giá đóng góp của các nhóm đặc trưng

Mục tiêu: Xác định nhóm đặc trưng hành vi nào cung cấp tín hiệu mạnh nhất để phân định trạng thái thành viên. Việc này giúp lý giải mô hình mục tiêu có xu hướng ghi nhớ nguyên văn (verbatim) hay ghi nhớ xấp xỉ (approximate). Thiết lập: Thực nghiệm cô lập các nhóm đặc trưng đầu vào:

- Chỉ sử dụng đặc trưng từ vựng (Lexical-only: ROUGE-L, Token F1, LCS ratio, 4-gram overlap).
- Chỉ sử dụng đặc trưng ngữ nghĩa (Semantic-only: Cosine, L2 distance, Dot product).
- Kết hợp toàn bộ đặc trưng (Full behavior vector).

Bộ dữ liệu / Đặc trưng	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
Từ vựng	0.521	0.723	0.620	0.591	0.656
Ngữ nghĩa	0.573	0.733	0.651	0.655	0.662
Kết hợp	0.685	0.741	0.703	0.701	0.703

Bảng 4.1: Điểm số AUC các tập feature đem lại trên kiến trúc cơ bản đã đề xuất

Bộ dữ liệu / Đặc trưng	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
Từ vựng	0.152	0.352	0.201	0.271	0.251
Ngữ nghĩa	0.206	0.415	0.212	0.285	0.268
Kết hợp	0.221	0.430	0.352	0.368	0.372

Bảng 4.2: Điểm số TPR@5%FPR các tập feature đem lại trên kiến trúc cơ bản đã đề xuất

Từ kết quả trên có thể rút ra một số nhận xét chính. Thứ nhất, nhóm đặc trưng từ vựng cho kết quả thấp nhất trên toàn bộ các tập dữ liệu. Điều này cho thấy mức độ trùng khớp bề mặt giữa Generated Suffix và True Suffix chưa đủ để phân biệt rõ ràng giữa member và non-member. Nói cách khác, mô hình mục tiêu không nhất thiết sinh lại nguyên văn văn bản gốc, ngay cả khi mẫu đó từng xuất hiện trong dữ liệu huấn luyện.

Thứ hai, nhóm đặc trưng ngữ nghĩa đạt kết quả cao hơn nhóm đặc trưng từ vựng trên tất cả các bộ dữ liệu. Điều này cho thấy tín hiệu thành viên không chỉ nằm ở sự sao chép nguyên văn, mà còn thể hiện qua mức độ tương đồng về ý nghĩa giữa phần sinh ra và hậu tố tham chiếu.

Thứ ba, vector đặc trưng kết hợp đạt kết quả tốt nhất trên toàn bộ các tập dữ liệu, xét cả theo AUC và TPR@5%FPR. Kết quả này cho thấy việc kết hợp đặc trưng từ vựng và đặc trưng ngữ nghĩa giúp mô hình khai thác đầy đủ hơn các biểu hiện hành vi của mô hình mục tiêu.

Nhìn chung, đặc trưng ngữ nghĩa là nguồn tín hiệu đơn lẻ quan trọng nhất, trong khi vector đặc trưng kết hợp là cấu hình hiệu quả nhất tổng thể. Điều này phù hợp với giả thuyết rằng hành vi ghi nhớ của LLM trong thiết lập hộp đen không chỉ biểu hiện qua tái tạo nguyên văn, mà còn qua khả năng sinh ra nội dung tương đồng về mặt ngữ nghĩa với văn bản gốc.

4.4.2. Thực nghiệm 2: Đánh giá ảnh hưởng của cấu hình độ dài Prefix/Suffix

Mục tiêu: Phân tích xem việc điều chỉnh tỷ lệ cung cấp ngữ cảnh (Prefix) và phần đối chiếu (TrueSuffix) ảnh hưởng như thế nào đến khả năng kích thích vùng nhớ của mô hình mục tiêu. Thiết lập: Cố định tổng độ dài mẫu (Item) ở 128 token, thực nghiệm tiến hành cắt lớp với các tỷ lệ khác nhau:

- 64 token Prefix / 64 token TrueSuffix.
- 80 token Prefix / 48 token TrueSuffix.
- 96 token Prefix / 32 token TrueSuffix (Cấu hình thiên về cung cấp nhiều ngữ cảnh mồi).

Bộ dữ liệu / Tỉ lệ Prefix/Suffix	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
64/64	0.631	0.723	0.620	0.694	0.656
80/48	0.665	0.733	0.691	0.721	0.722
96/32	0.685	0.741	0.703	0.701	0.703

Bảng 4.3: Điểm số AUC khi thay đổi tỉ lệ token của Prefix/Suffix trên kiến trúc cơ bản đã đề xuất

Bộ dữ liệu / Tỉ lệ Prefix/Suffix	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
64/64	0.202	0.401	0.330	0.321	0.351
80/48	0.215	0.415	0.341	0.375	0.375
96/32	0.221	0.430	0.352	0.368	0.372

Bảng 4.4: Điểm số TPR@5%FPR khi thay đổi tỉ lệ token của Prefix/Suffix trên kiến trúc cơ bản đã đề xuất

Kết quả thực nghiệm cho thấy việc tăng độ dài Prefix nhìn chung giúp cải thiện hiệu quả suy luận thành viên so với cấu hình cân bằng 64/64. Cả hai cấu hình 80/48 và 96/32 đều cho kết quả tốt hơn 64/64 trên hầu hết các tập dữ liệu, cho thấy việc cung cấp thêm ngữ cảnh mồi có thể giúp mô hình mục tiêu sinh ra phần tiếp nối mang nhiều tín hiệu phân biệt hơn giữa member và non-member.

Xét theo AUC, cấu hình 96/32 đạt kết quả cao nhất trên WikiMIA, BookTection và ArXivTection, trong khi cấu hình 80/48 đạt kết quả tốt nhất trên hai tập dữ liệu tiếng Việt. Xu hướng tương tự cũng xuất hiện ở TPR@5%FPR, khi 96/32 tốt hơn trên ba tập dữ liệu tiếng Anh, còn 80/48 lại cao hơn trên Tiểu sử - tiếng Việt và Pháp luật - tiếng Việt.

Nhìn chung, kết quả cho thấy việc tăng tỷ lệ Prefix là có lợi, nhưng không phải lúc nào cấu hình có Prefix dài nhất cũng tối ưu. Khi phần True Suffix quá ngắn, lượng thông tin dùng để đối chiếu có thể bị giảm, đặc biệt ở một số miền dữ liệu tiếng Việt. Do đó, tỷ lệ Prefix/Suffix cần được lựa chọn tùy theo miền dữ liệu và mục tiêu đánh giá

4.4.3. Thực nghiệm 3: So sánh các bộ phân loại và kích thước dữ liệu huấn luyện

Mục tiêu: Thực nghiệm này nhằm so sánh hiệu quả của các bộ phân loại khác nhau trên không gian đặc trưng đa chiều, từ đó xác định mô hình phù hợp nhất cho bài toán suy luận thành viên trong điều kiện dữ liệu gán nhãn hạn chế. Thay vì giả định trước rằng một mô hình cụ thể vượt trội hơn các mô hình còn lại, thực nghiệm tập trung đánh giá khách quan khả năng phân loại của từng phương pháp khi số lượng mẫu huấn luyện thay đổi.

Bộ dữ liệu / Bộ phân loại	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
PrototypicalNet	0,716	0,738	0,719	0,772	0,710
Logistic Regression	0,708	0,762	0,714	0,716	0,720
SVM	0,748	0,751	0,744	0,703	0,734

Bảng 4.5: Điểm số AUC tương ứng với từng bộ phân loại trên kiến trúc cơ bản đã đề xuất

Bộ dữ liệu / Bộ phân loại	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
PrototypicalNet	0,401	0,462	0,421	0,472	0,409
Logistic Regression	0,316	0,378	0,333	0,365	0,328
SVM	0,354	0,371	0,362	0,347	0,346

Bảng 4.6: Điểm số TPR@5%FPR tương ứng với từng bộ phân loại trên kiến trúc cơ bản đã đề xuất

Kết quả ở Bảng 4.5 cho thấy SVM đạt AUC trung bình cao nhất với 0,736, cao hơn nhẹ so với PrototypicalNet ở mức 0,731 và Logistic Regression ở mức 0,724. Tuy nhiên, không có bộ phân loại nào đạt kết quả tốt nhất trên toàn bộ các tập dữ liệu. SVM đạt AUC cao nhất trên WikiMIA, arXivTection và Pháp luật tiếng Việt; Logistic Regression đạt kết quả cao nhất trên BookTection; trong khi PrototypicalNet đạt kết quả tốt nhất trên tập Tiểu sử tiếng Việt. Kết quả này cho thấy hiệu quả của bộ phân loại phụ thuộc nhất định vào đặc điểm của từng miền dữ liệu..

Đối với TPR@5%FPR, PrototypicalNet đạt kết quả cao nhất trên cả năm tập dữ liệu, với giá trị trung bình 0,433, cao hơn SVM ở mức 0,356 và Logistic Regression ở mức 0,344. Điều này cho thấy PrototypicalNet có ưu thế rõ hơn trong điều kiện yêu cầu kiểm soát tỷ lệ dương tính giả ở mức thấp. Vì vậy, mặc dù SVM nhỉnh hơn về khả năng phân biệt tổng thể theo AUC, PrototypicalNet vẫn là lựa chọn phù hợp hơn đối với bài toán kiểm toán dữ liệu, nơi việc hạn chế cảnh báo sai được ưu tiên.

Nhìn chung, PrototypicalNet là bộ phân loại phù hợp nhất trong thực nghiệm này. Kết quả này gợi ý rằng việc mô hình hóa hai lớp member và non-member thông qua các nguyên mẫu trong không gian đặc trưng hành vi có thể đem lại lợi thế so với các bộ phân loại tuyến tính hoặc truyền thống như Logistic Regression và SVM, đặc biệt trong điều kiện dữ liệu gán nhãn hạn chế.

4.4.4. Thực nghiệm 4: Đánh giá hiệu suất tổng thể và so sánh với các phương pháp hiện có

Mục tiêu: Kết hợp các kết quả thực nghiệm đem lại kết quả cao nhất trước đó đồng thời so sánh với các phương pháp nghiên cứu trước đó để kiểm nghiệm liệu phương pháp của đề án đề xuất có thực sự hiệu quả so với các phương pháp tiền nghiên cứu không:

Bộ dữ liệu / Phương pháp	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
DE-COP	0.525	0.720	0.506	x	x
Name-cloze	0.531	0.544	0.527	x	x
BSS-MIA	**0.651**	**0.743**	**0.712**	**0.692**	**0.683**

Bảng 4.7: So sánh điểm số AUC của phương pháp kết hợp đề xuất trong đề án với các phương pháp nổi bật hiện có

Bộ dữ liệu / Phương pháp	WikiMIA	BookTection	ArXivTection	Tiểu sử - Tiếng Việt	Pháp Luật - Tiếng Việt
DE-COP	0.092	0.385	0.018	x	x
Name-cloze	0.079	0.148	0.096	x	x
BSS-MIA	**0.230**	**0.437**	**0.352**	**0.291**	**0.305**

Bảng 4.8: So sánh điểm số TPR@5%FPR của phương pháp kết hợp đề xuất trong đề án với các phương pháp nổi bật hiện có

Nhìn chung, thực nghiệm tổng thể cho thấy BSS-MIA đạt hiệu quả tốt hơn các phương pháp hiện có trên các bộ dữ liệu có kết quả đối sánh. Xét theo AUC, BSS-MIA đạt kết quả cao nhất trên WikiMIA, BookTection và ArXivTection, lần lượt

đạt 0.651, 0.743 và 0.712. Xu hướng này cũng được thể hiện rõ ở chỉ số TPR@5%FPR, khi BSS-MIA tiếp tục vượt DE-COP và Name-cloze trên cả ba bộ dữ liệu.

Đối với hai bộ dữ liệu tiếng Việt, do các phương pháp đối sánh chưa có kết quả công bố tương ứng, thực nghiệm chủ yếu dùng để đánh giá khả năng áp dụng của BSS-MIA trong bối cảnh tiếng Việt. Kết quả cho thấy phương pháp vẫn duy trì hiệu quả tương đối ổn định trên Tiểu sử - tiếng Việt và Pháp luật - tiếng Việt, qua đó gợi ý khả năng mở rộng sang các miền dữ liệu ngoài tiếng Anh.

Lợi thế chính của BSS-MIA nằm ở việc chỉ khai thác văn bản đầu ra của mô hình mục tiêu trong thiết lập hộp đen, không yêu cầu truy cập xác suất token, log-likelihood, tham số hoặc trạng thái nội bộ của mô hình. Điều này giúp phương pháp phù hợp hơn với bối cảnh kiểm toán các LLM thương mại, nơi người đánh giá thường chỉ có thể tương tác với mô hình thông qua API sinh văn bản.

4.5. Phân tích kết quả

Từ các số liệu thu thập được, phân tích sẽ tập trung vào các khía cạnh sau:

1. Đánh giá tổng thể: Chỉ số AUC và TPR@5%FPR duy trì ở mức cao chứng minh rằng vector đặc trưng hành vi cung cấp tín hiệu phân biệt tốt giữa dữ liệu đã học và chưa học, khắc phục được sự phụ thuộc vào xác suất bề nổi.

2. So sánh Lexical và Semantic: Semantic có đóng góp lớn, mô hình thiên về ghi nhớ xấp xỉ - sinh ra các cấu trúc từ ngữ khác nhưng giữ nguyên ý nghĩa gốc. Sự kết hợp cả hai (Multidimensional feature vectors) tạo ra độ nhận diện chính xác nhất.

3. Phân tích Lỗi Dương tính giả (False Positive): Các mẫu Non-member bị nhận diện nhầm thường chứa các chuỗi từ vựng dễ đoán, phổ biến hoặc có văn phong rập khuôn.

4. Phân tích Lỗi Âm tính giả (False Negatives): Một số mẫu Member không được nhận diện có thể do dữ liệu đó chỉ được huấn luyện 1 chu kỳ (1-epoch) khiến mô hình chưa đủ độ nhớ, hoặc do cơ chế an toàn RLHF của OpenAI chủ động ngắt hành vi sao chép văn bản.

4.6. Đánh giá chi phí và khả năng mở rộng

So với các phương pháp dựa trên tham chiếu (cần huấn luyện Shadow models đắt đỏ) hoặc DE-COP (cần tạo tập hợp các câu hỏi trắc nghiệm thủ công), phương pháp đề xuất mang lại hiệu quả rõ rệt:

- Khâu chuẩn bị Prefix/TrueSuffix và trích xuất Lexical có chi phí tính toán rất thấp, chủ yếu đến từ việc gọi api mô hình mục tiêu.

- Việc sử dụng BERT để mã hóa Semantic có chi phí trung bình nhưng dễ dàng tối ưu bằng cơ chế xử lý song song (batch processing).

- Chi phí lớn nhất duy nhất là số lần gọi API tới mô hình hộp đen. Nhờ việc không cần kỹ sư câu lệnh (prompt engineering) phức tạp, hệ thống tự động hóa hoàn toàn luồng đánh giá với độ mở rộng rất cao trên các tập dữ liệu khổng lồ.

4.7. Hạn chế thực nghiệm

Dù đạt được kết quả khả quan, thực nghiệm vẫn tồn tại một số điểm cần cải thiện:

- Phạm vi đánh giá hiện tại chủ yếu tập trung vào họ mô hình OpenAI (GPT-3.5), cần mở rộng sang các kiến trúc khác như Claude, LLaMA hay Mistral.

- Với các văn bản tiếng Việt, việc sử dụng bert-base-uncased có thể làm giảm chất lượng biểu diễn ngữ nghĩa. Tương lai cần thay thế bằng PhoBERT hoặc các mô hình mã hóa đa ngữ.

- Nhãn "Member/Non-member" dựa trên thời gian (time cut-off) chỉ là nhãn đại diện (proxy), không đại diện chắc chắn 100% cho việc dữ liệu có thực sự được đưa vào file huấn luyện của nhà phát triển hay không.

Thực nghiệm trong Chương 4 đã chứng minh tính đúng đắn và hiệu quả vượt trội của phương pháp Behavioral Suffix Similarity MIA. Bằng cách từ bỏ việc phân tích văn bản thô hay truy cập xác suất token, phương pháp đã tận dụng thành công hành vi tự hồi quy của mô hình mục tiêu thông qua phép thử kích thích Prefix. Việc chuyển đổi độ tương đồng giữa văn bản sinh ra và văn bản tham chiếu thành một vector đặc trưng đa chiều (kết hợp từ vựng và ngữ nghĩa) đã cung cấp cho PrototypicalNet một không gian phân tách rõ ràng. Kết quả thực nghiệm không chỉ cho thấy tín hiệu phân biệt mạnh mẽ mà còn khẳng định đây là một giải pháp hộp đen tinh gọn, chi phí thấp, tính tự động hóa cao và hoàn toàn khả thi để ứng dụng trong thực tiễn kiểm định bản quyền LLM.

KẾT LUẬN

Đề án đã nghiên cứu bài toán tấn công suy luận thành viên đối với mô hình ngôn ngữ lớn, tập trung vào bối cảnh thực tế khi mô hình mục tiêu chỉ có thể được truy cập dưới dạng hộp đen thông qua API. Trong bối cảnh này, các thông tin nội bộ như trọng số, gradient, logit hay xác suất token thường không khả dụng, khiến nhiều phương pháp MIA truyền thống khó triển khai hoặc có chi phí cao. Từ hạn chế đó, đề án đề xuất hướng tiếp cận dựa trên hành vi sinh tiếp văn bản của mô hình.

Phương pháp đề xuất, Behavioral Suffix Similarity MIA, khai thác giả thuyết rằng nếu một mẫu văn bản từng xuất hiện trong dữ liệu huấn luyện, mô hình có xu hướng sinh ra phần tiếp nối gần giống với phần văn bản gốc hơn so với các mẫu chưa từng được học. Trên cơ sở đó, mỗi mẫu được tách thành prefix và true suffix; prefix được dùng để truy vấn mô hình mục tiêu, còn generated suffix được so sánh với true suffix bằng nhiều thước đo khác nhau. Các thước đo này được tổng hợp thành vector đặc trưng hành vi và đưa vào mô hình học máy để phân loại member/non-member.

Về mặt kỹ thuật, đề án đã xây dựng được một quy trình thực nghiệm tương đối hoàn chỉnh, bao gồm chuẩn bị dữ liệu member/non-member, sinh bằng chứng hành vi từ mô hình mục tiêu, tính toán các đặc trưng tương đồng, huấn luyện bộ phân loại và đánh giá bằng các chỉ số Accuracy, AUC và TPR@5%FPR. Kết quả tạm thời cho thấy hướng tiếp cận này có khả năng phân biệt member và non-member với tín hiệu khá rõ ràng, đặc biệt khi sử dụng các đặc trưng hành vi đa chiều thay vì phụ thuộc vào một thước đo đơn lẻ.

Đóng góp chính của đề án nằm ở việc chuyển bài toán MIA trong bối cảnh LLM hộp đen từ phân tích xác suất nội bộ sang phân tích hành vi đầu ra. Cách tiếp cận này phù hợp hơn với các mô hình thương mại hoặc mô hình đóng, giảm yêu cầu truy cập nội bộ và hạn chế chi phí so với các phương pháp cần huấn luyện shadow models quy mô lớn. Đồng thời, việc kết hợp nhiều đặc trưng tương đồng giúp giảm rủi ro nhận diện sai do các văn bản phổ biến hoặc có cấu trúc quen thuộc.

Tuy vậy, đề án vẫn còn một số hạn chế. Quy mô thực nghiệm hiện còn phụ thuộc vào chi phí truy vấn mô hình mục tiêu, số lượng miền dữ liệu chưa đủ rộng để khẳng định tính tổng quát tuyệt đối, và kết quả có thể bị ảnh hưởng bởi cấu hình sinh như độ dài prefix/suffix, temperature, số lần sinh và cách thiết kế prompt. Ngoài ra, việc gán nhãn member/non-member trong bối cảnh LLM thực tế vẫn là một thách thức, do không thể biết chính xác toàn bộ dữ liệu huấn luyện của mô hình mục tiêu.

Trong hướng phát triển tiếp theo, đề án có thể mở rộng thực nghiệm trên nhiều miền dữ liệu hơn, bao gồm văn bản tiếng Anh, tiếng Việt, văn bản pháp luật, học thuật và dữ liệu web. Bên cạnh đó, cần đánh giá sâu hơn ảnh hưởng của các cấu hình sinh, so sánh nhiều bộ phân loại khác nhau trong điều kiện dữ liệu gán nhãn hạn chế, cũng như kiểm tra độ ổn định của phương pháp trên nhiều mô hình mục tiêu. Một hướng mở rộng khác là kết hợp thêm các đặc trưng ngữ nghĩa hoặc cơ chế phát hiện trùng lặp xấp xỉ để tăng khả năng phát hiện ghi nhớ không nguyên văn.

Tổng kết lại, đề án cho thấy việc khai thác hành vi sinh tiếp văn bản là một hướng khả thi cho bài toán suy luận thành viên đối với mô hình ngôn ngữ lớn trong kịch bản hộp đen. Phương pháp đề xuất không yêu cầu truy cập nội bộ mô hình, có khả năng tận dụng các tín hiệu hành vi đa chiều và tạo nền tảng cho các nghiên cứu tiếp theo về kiểm định quyền riêng tư, bản quyền dữ liệu và mức độ ghi nhớ của các mô hình ngôn ngữ lớn.

TÀI LIỆU THAM KHẢO

[1] M. M. Grynbaum và R. Mac, "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work," The New York Times, 2023. [Online]. Available: https://www.nytimes.com/2026/06/25/technology/times-lawsuit-openai-microsoft.html - truy cập ngày 20/05/2026

[2] A. Guadamuz, "Artists file class action lawsuit against Stability AI, DeviantArt and Midjourney," TechnoLlama, Jan. 2023. [Online]. Available: https://www.technollama.co.uk/artists-file-class-action-lawsuit-against-stability-ai-deviantart-and-midjourney - truy cập ngày 20/05/2026

[3] C. Yang et al., "Detecting Training Data For Large Language Models: A Survey," ACM Comput. Surv., vol. 58, no. 9, Art. no. 221, Feb. 2026, doi: 10.1145/3779430.

[4] M. Nasr et al., "Scalable extraction of training data from (production) language models," arXiv preprint arXiv:2311.17035, 2023. Available: https://arxiv.org/abs/2311.17035

[5] N. Carlini et al., "Extracting Training Data from Large Language Models," in Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 2633–2650.

[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 3–18.

[7] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: Phi-1.5 technical report," arXiv preprint arXiv:2309.05463, 2023

[8] Z. Liu, T. Zhu, C. Tan, H. Lu, B. Liu, and W. Chen, "Probing language models for pre-training data detection," arXiv preprint arXiv:2406.01333, 2024

[9] B. Zhou et al., "DPDLLM: A black-box framework for detecting pre-training data from large language models," in Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 644–653

[10] N. Carlini et al., "Quantifying memorization across neural language models," arXiv preprint arXiv:2202.07646, 2022. [Online]. Available: https://arxiv.org/abs/2202.07646

[11] J. Mattern, F. Mireshghallah, Z. Jin, B. Schölkopf, M. Sachan, and T. Berg-Kirkpatrick, "Membership inference attacks against language models via neighbourhood comparison," arXiv preprint arXiv:2305.18462, 2023.

[12] W. Shi et al., "Detecting pretraining data from large language models," arXiv preprint arXiv:2310.16789, 2023.

[13] Z. Liu, T. Zhu, C. Tan, H. Lu, B. Liu, and W. Chen, "Probing language models for pre-training data detection," arXiv preprint arXiv:2406.01333, 2024.

[14] A. V. Duarte, X. Zhao, A. L. Oliveira, and L. Li, "De-cop: Detecting copyrighted content in language models training data," arXiv preprint arXiv:2402.09910, 2024.

Minh Đam Vũ

đề_án_final vuminhdam.docx

Đề án tốt nghiệp khóa 2024 đợt 1 kỹ thuật

Khoa Đào tạo Sau đại học

Posts and Telecommunications Institute of Technology

Thông tin chi tiết về tài liệu

ID bài nộp
trn:oid:::1:3619130333

Ngày nộp
11:43 GMT+7 29 thg 7, 2026

Ngày tải xuống
11:47 GMT+7 29 thg 7, 2026

Tên tệp
đề_án_final_vuminhdam.docx

Kích cỡ tệp
4.2 MB

48 Trang

18.743 Từ

70.717 Ký tự

turnitin Trang 1/51 - Trang bìa ID bài nộp: trn:oid:::1:3619130333

turnitin Trang 2/51 - Tổng quan về tính toàn vẹn ID bài nộp: trn:oid:::1:3619130333

19% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm các nguồn trùng lặp, cho mỗi n...

Đã lọc khỏi Báo cáo

- Mục lục tham khảo
- Văn bản được trích dẫn
- Văn bản được trích dẫn
- Kết quả trùng khớp nhỏ (ít hơn 8 từ)

Nguồn hàng đầu

17% Nguồn Internet

7% Ấn bản

9% Bài tập được nộp (bài của học sinh)

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. DƯƠNG TRẦN ĐỨC

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Vũ Minh Đam

Chuyên ngành: HTTT

Khóa: 2024 đợt 2

Tên đề tài: Nghiên cứu phương pháp phát hiện dữ liệu riêng tư được sử dụng trong huấn luyện mô hình ngôn ngữ lớn.

Người hướng dẫn khoa học: TS. Dương Trần Đức

Ngày bảo vệ: 21/07/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Bố cục lại nội dung đề án	Học viên đã rà soát và bố cục lại nội dung đề án theo hướng logic, thống nhất hơn; rút gọn phần mở đầu, bổ sung và hoàn thiện Chương 1.
2	Làm rõ phương pháp, kết quả thực nghiệm và kết quả nghiên cứu của đề án	Học viên đã bổ sung và làm rõ quy trình thu thập, gán nhãn, xử lý và phân chia dữ liệu.
3	Chỉnh sửa lại các lỗi sai sót trong đề án	Học viên đã rà soát, chỉnh sửa các lỗi soạn thảo, các lỗi ngữ pháp, và nhận xét tại bảng dữ liệu.

Hà Nội, ngày 29 tháng 07 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
TS. Nguyễn Duy Phương	TS. Đào Thị Thúy Quỳnh	TS. Dương Trần Đức	Vũ Minh Đam

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1949/QĐ-HV ngày 07 tháng 07 năm 2026 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi....9..giờ..30..phút, ngày..21 tháng 07 năm 2026 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Vũ Minh Đam**

Tên đề án tốt nghiệp: **Nghiên cứu phương pháp phát hiện dữ liệu riêng tư được sử dụng trong huấn luyện mô hình ngôn ngữ lớn**

Chuyên ngành: **Hệ thống thông tin** Mã số: **8480104**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ....5.../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**TS. Nguyễn Duy Phương**	Chủ tịch	
2	**TS. Đào Thị Thúy Quỳnh**	Thư ký	
3	**TS. Đào Ngọc Phong**	Phản biện 1	
4	**TS. Trần Minh Tân**	Phản biện 2	
5	**PGS.TS. Đặng Văn Đức**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

1) Học viên sử dụng chiến lược gán nhãn trong thái thành viên (Member - Non member) dựa vào mốc thời gian xuất bản DL và thời điểm hạn thu thập dữ liệu. Học viên làm rõ nội dung này ?

2) Học viên đã thực hiện đánh giá chi phí tính toán chưa ?

7. Trả lời của học viên:

1) Hiện tại cái pp luận này xây dựng thủ công còn phương pháp đề xuất của em được thực hiện hoàn toàn tự động
2) Đánh giá chi phí bảo trì phụ thuộc toàn bộ vào thời gian gọi API

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo)*.

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: TS. Nguyễn Duy Phương
  2. Uỷ viên Ban kiểm phiếu: TS. Đào Ngọc Phong
  3. Uỷ viên Ban kiểm phiếu: TS. Đào Thị Thúy Quỳnh
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 8.5

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

- Bổ sung lại nội dung đề án
- Làm rõ phương pháp, kết quả thực nghiệm, kết quả nghiên cứu của đề án
- Chỉnh sửa lại các lỗi sai xót trong đề án

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Đề nghị công nhận

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS..... Không.....

Buổi làm việc kết thúc vào..... cùng ngày.

Chủ tịch

Phương

Thư ký

Đào Thị Thúy Quỳnh

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

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(*Dùng cho người phản biện*)

Tên đề tài đề án tốt nghiệp: Nghiên cứu phương pháp phát hiện dữ liệu riêng tư được sử dụng trong huấn luyện mô hình ngôn ngữ lớn.

Chuyên ngành: Hệ thống thông tin.

Mã chuyên ngành: 8.48.01.04.

Họ và tên học viên: Vũ Minh Đam.

Họ và tên người nhận xét: Trần Minh Tân.

Học hàm, học vị: Tiến sỹ.

Chuyên ngành: Công nghệ thông tin.

Cơ quan công tác: Học viện Chiến lược Khoa học và Công nghệ.

Số điện thoại: 0913275577; E-mail: tantm@mst.gov.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Đề tài nghiên cứu bài toán phát hiện dữ liệu riêng tư đã được sử dụng trong quá trình huấn luyện mô hình ngôn ngữ lớn (LLM) là một chủ đề có tính thời sự trong bối cảnh AI tạo sinh được ứng dụng rộng rãi. Đề án đã nêu được bối cảnh, động lực nghiên cứu và những thách thức của các phương pháp Membership Inference Attack truyền thống khi áp dụng cho LLM, hệ thống cơ sở lý thuyết khá đầy đủ về LLM, các phương pháp MIA, phương pháp tham chiếu và không tham chiếu, đồng thời phân tích các hạn chế của các hướng tiếp cận hiện nay để làm cơ sở đề xuất phương pháp mới.

Học viên đã xác định được khoảng trống nghiên cứu đối với kịch bản hộp đen và dữ liệu tiếng Việt, đây là hướng nghiên cứu còn ít công trình công bố. Vì vậy đề tài có cơ sở khoa học và tính thực tiễn cao.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Đề án được bố cục hợp lý gồm bốn chương: cơ sở lý thuyết; tổng quan các phương pháp hiện tại và điểm nghẽn; phương pháp đề xuất; thực nghiệm và đánh giá. Trình tự nghiên cứu logic, phù hợp với mục tiêu đề ra.

Học viên đã thực hiện các nội dung chính:

- Tổng hợp và phân tích các phương pháp phát hiện dữ liệu huấn luyện của LLM;

- Đề xuất phương pháp phát hiện dựa trên hành vi sinh văn bản trong môi trường hộp đen;

- Kết hợp nhiều đặc trưng từ vựng và ngữ nghĩa (BERT embeddings, ROUGE-L, Token F1, Cosine Similarity...) để xây dựng vector đặc trưng;

- Sử dụng mô hình học máy để phân loại thành viên/không thành viên;

- Xây dựng các thực nghiệm đánh giá ảnh hưởng của đặc trưng, bộ phân loại và so sánh với các phương pháp hiện có.

Với các nội dung trên, Đề tài có ý nghĩa cả về nghiên cứu và ứng dụng trong kiểm toán dữ liệu huấn luyện, bảo vệ dữ liệu cá nhân và đánh giá mức độ minh bạch của các mô hình ngôn ngữ lớn.

Tuy nhiên, còn một số điểm cần làm rõ thêm, cụ thể:

1. Mức độ mới của phương pháp. Cần làm rõ hơn đóng góp mới nằm ở đâu: ở quy trình kết hợp đặc trưng, cách lựa chọn đặc trưng hay mô hình phân loại (hiện tại Đề án mới kết hợp các đặc trưng đã được biết đến như ROUGE-L, Cosine Similarity, BERT embeddings,... và mô hình phân loại học máy).

2. Thiết kế thực nghiệm. Cần trình bày rõ hơn quy mô tập dữ liệu tiếng Việt, cách xây dựng tập member/non-member, tiêu chí gán nhãn và khả năng tái lập thực nghiệm để đánh giá được tính chính xác của kết quả thực nghiệm.

3. Khả năng tổng quát hóa. Phần thực nghiệm cần phân tích rõ hơn khả năng áp dụng phương pháp với nhiều LLM khác nhau (GPT, Llama, Qwen...) thay vì chỉ trên một hoặc một vài mô hình.

4. Đánh giá chi phí tính toán. Mặc dù đề án có mục đánh giá chi phí và khả năng mở rộng, cần làm rõ hơn thời gian xử lý, số lần truy vấn API khi triển khai trong thực tế.

III/ Những vấn đề cần giải thích thêm:

1. Đóng góp mới quan trọng nhất của phương pháp đề xuất so với các phương pháp Membership Inference Attack hiện nay là gì?

2. Vì sao học viên lựa chọn kết hợp đặc trưng từ vựng và đặc trưng ngữ nghĩa thay vì chỉ sử dụng các chỉ số như Perplexity hoặc Min-K% Prob?

3. Nếu áp dụng phương pháp cho các mô hình LLM khác (như GPT-4, Llama, Qwen,...) thì có cần hiệu chỉnh mô hình phân loại hay không?

IV/ Kết luận:

Học viên đã hoàn thành các nội dung nghiên cứu theo đề cương đặt ra. Báo cáo Đề án đã nghiên cứu, lựa chọn một bài toán có ý nghĩa khoa học và thực tiễn cao trong lĩnh vực an toàn dữ liệu và trí tuệ nhân tạo, đã tiến hành nghiên cứu có hệ thống các phương pháp Membership Inference Attack, phân tích các hạn chế của các hướng tiếp cận hiện có và đề xuất một phương pháp phát hiện dựa trên hành vi sinh văn bản của mô hình ngôn ngữ lớn trong bối cảnh hộp đen. Tổng thể chung, kết quả báo cáo đạt yêu cầu.

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp./.

Ngày 20 tháng 7 năm 2026
NGƯỜI NHẬN XÉT

Trần Minh Tân

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án: Nghiên cứu phương pháp phát hiện dữ liệu riêng tư được sử dụng trong huấn luyện mô hình ngôn ngữ lớn.

Chuyên ngành: Hệ thống thông tin.

Mã số chuyên ngành: 8.48.01.04.

Học viên thực hiện: Vũ Minh Đam

Họ và tên người nhận xét: Đào Ngọc Phong

Học hàm, học vị: Tiến sĩ

Chuyên ngành:Hệ thống thông tin

Cơ quan công tác: Khoa Công nghệ thông tin 1

Số điện thoại: 0912239696 E-mail: phongdn@ptit.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Sự phát triển mạnh mẽ của các Mô hình Ngôn ngữ Lớn (Large Language Models - LLMs) đã mang lại hiệu suất vượt trội trong hàng loạt tác vụ xử lý ngôn ngữ tự nhiên (NLP) như trả lời câu hỏi, dịch máy và tóm tắt văn bản. Sức mạnh của các mô hình này chủ yếu đến từ việc sử dụng một lượng dữ liệu khổng lồ trong quá trình huấn luyện.

Tuy nhiên, việc mở rộng quy mô dữ liệu không được kiểm soát chặt chẽ đã dẫn đến việc các tập dữ liệu huấn luyện (training datasets) thường bao hàm cả các thông tin nhạy cảm như dữ liệu cá nhân, sách, báo chí và các tài liệu có bản quyền. Hậu quả là, LLM có nguy cơ rò rỉ quyền riêng tư hoặc vi phạm bản quyền nghiêm trọng nếu mô hình sinh ra các đoạn văn bản giống hệt hoặc có độ tương đồng cao với dữ liệu gốc.

Mục đích nghiên cứu của đề án là nghiên cứu bài toán phát hiện dữ liệu riêng tư được sử dụng trong quá trình huấn luyện hoặc tiền huấn luyện mô hình ngôn ngữ lớn, từ đó đánh giá khả năng xác định một văn bản cụ thể có từng xuất hiện trong dữ liệu huấn luyện của mô hình hay không. Trên cơ sở đó, đề án hướng tới việc xây dựng và thử nghiệm một phương pháp phát hiện phù hợp với bối cảnh mô hình hộp đen, đặc biệt muốn đóng góp cho thử nghiệm, nghiên cứu phương pháp phát hiện dữ liệu liên

quan đến dữ liệu tiếng Việt - một hướng nghiên cứu hiện còn thiếu các công trình chuyên sâu và hệ thống.

Do đó, đề án có ý nghĩa khoa học và thực tế áp dụng.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Đề án đã nghiên cứu bài toán tấn công suy luận thành viên đối với mô hình ngôn ngữ lớn, tập trung vào bối cảnh thực tế khi mô hình mục tiêu chỉ có thể được truy cập dưới dạng hộp đen. Từ hạn chế đó, đề án đề xuất hướng tiếp cận dựa trên hành vi sinh tiếp văn bản của mô hình.

Đề án đề xuất phương pahps Suy luận thành viên dựa trên độ tương đồng của hậu tố sinh ra (Behavioral Suffix Similarity Membership Inference Attack - BSS-MIA). Mục tiêu cốt lõi của phương pháp là giải quyết bài toán Tấn công Suy luận Thành viên (Membership Inference Attack) thông qua việc phân loại nhị phân: xác định xem một mẫu văn bản cụ thể có thuộc tập dữ liệu huấn luyện của mô hình mục tiêu hay không

Về mặt kỹ thuật, đề án đã xây dựng được một quy trình thực nghiệm tương đối hoàn chỉnh, bao gồm chuẩn bị dữ liệu member/non-member, sinh bằng chứng hành vi từ mô hình mục tiêu, tính toán các đặc trưng tương đồng, huấn luyện bộ phân loại và đánh giá bằng các chỉ số.

III/ Những vấn đề cần giải thích thêm:

Một số góp ý:

- Xem lại bố cục đề án, phần mở đầu không cần viết cụ thể. Chương 1 quá ngắn (5-6 trang),...
- Xem lại trang 38, "Kết quả ở Bảng 4.5 cho thấy PrototypicalNet đạt AUC cao nhất trên toàn bộ năm bộ dữ liệu", tuy nhiên đối chiếu với bảng 4.5 thì không đúng.
- Đề án còn một số lỗi trình bày, lỗi chính tả.

Một số câu hỏi:

- Học viên sử dụng chiến lược gán nhãn trạng thái thành viên (Member/Non-member) dựa vào mốc thời gian xuất bản dữ liệu và thời điểm giới hạn thu thập dữ liệu. Học viên làm rõ việc thực hiện nội dung này khi chuẩn bị dữ liệu thực nghiệm (trang 31,32).
- Với các văn bản tiếng Việt, việc sử dụng bert-base-uncased có thể làm giảm chất lượng biểu diễn ngữ nghĩa. Do đó, học viện làm rõ thêm về kết quả đánh giá với tập dữ liệu tiếng Việt.

V/ Kết luận:

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày 20 tháng 07 năm 2026

NGƯỜI NHẬN XÉT

Đào Ngọc Phong