

ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA KHOA HỌC MÁY TÍNH

NGUYỄN HOÀNG PHÚC

KHÓA LUẬN TỐT NGHIỆP  
HỆ THỐNG TỰ ĐỘNG PHÁT HIỆN VÀ KIỂM CHỨNG TIN  
TỨC GIẢ MẠO TRONG CÁC CHIẾN DỊCH LỪA ĐẢO TRỰC  
TUYỂN BẰNG LLMs

AN AUTOMATED SYSTEM FOR DETECTING AND VERIFYING FAKE  
NEWS IN ONLINE FRAUD CAMPAIGNS USING LLMs

CỬ NHÂN NGÀNH TRÍ TUỆ NHÂN TẠO

TP. Hồ Chí Minh, 2025

**ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA KHOA HỌC MÁY TÍNH**

**NGUYỄN HOÀNG PHÚC - 22521129**

**KHÓA LUẬN TỐT NGHIỆP  
HỆ THỐNG TỰ ĐỘNG PHÁT HIỆN VÀ KIỂM CHỨNG TIN  
TỨC GIẢ MẠO TRONG CÁC CHIẾN DỊCH LỪA ĐẢO TRỰC  
TUYỂN BẰNG LLMs**

**AN AUTOMATED SYSTEM FOR DETECTING AND VERIFYING FAKE  
NEWS IN ONLINE FRAUD CAMPAIGNS USING LLMs**

**CỬ NHÂN NGÀNH TRÍ TUỆ NHÂN TẠO**

**GIẢNG VIÊN HƯỚNG DẪN:**

**TS. Phạm Văn Hậu**

**ThS. Phan Thế Duy**

**TP.Hồ Chí Minh, 2025**

## **LỜI CẢM ƠN**

Lời đầu tiên, em xin gửi lời cảm ơn chân thành nhất đến quý thầy cô trường Đại học Công nghệ Thông tin - ĐHQG TP.HCM, đặc biệt là các thầy cô khoa Khoa học máy tính, các thầy cô thuộc bộ môn Trí tuệ nhân tạo đã đồng hành suốt thời gian qua.

Nhóm xin được bày tỏ lòng biết ơn sâu sắc đến thầy Phan Thế Duy và thầy Phạm Văn Hậu vì hỗ trợ tận tình trong quá trình nghiên cứu và hoàn thành khóa luận. Cảm ơn thầy Phan Thế Duy đã dành nhiều thời gian tâm huyết để giúp đỡ và luôn đưa ra những góp ý, tạo động lực vượt qua các thách thức vượt ngoài khuôn khổ. Chính sự nhiệt tình và tận tâm của các thầy đã giúp em hoàn thành khóa luận một cách tốt nhất.

Bên cạnh đó, nhóm cũng chân thành cảm ơn gia đình và bạn bè, những người luôn đồng hành, động viên và chia sẻ khó khăn với nhóm trong suốt thời gian vừa qua.

Có thể nói, khóa luận này sẽ không thể hoàn thành được nếu thiếu sự hỗ trợ và giúp đỡ của tất cả mọi người. Một lần nữa, xin chân thành cảm ơn!

**Nguyễn Hoàng Phúc**

# MỤC LỤC

<b>LỜI CẢM ƠN</b>	i
<b>MỤC LỤC</b>	ii
<b>DANH MỤC CÁC CHỮ VIẾT TẮT</b>	vi
<b>DANH MỤC CÁC HÌNH VẼ</b>	vii
<b>DANH MỤC CÁC BẢNG BIỂU</b>	vii
<b>MỞ ĐẦU</b>	1
<b>CHƯƠNG 1. TỔNG QUAN</b>	<b>3</b>
1.1 Giới thiệu vấn đề	3
1.2 Giới thiệu các hướng tiếp cận phổ biến	4
1.3 Những thách thức	5
1.3.1 Giới hạn độ dài ngữ cảnh và duy trì liên kết thông tin	5
1.3.2 Đảm bảo tính chính xác và cập nhật nguồn dữ liệu	5
1.3.3 Phối hợp và đồng bộ trong hệ thống Multi-Agent	6
1.4 Câu hỏi nghiên cứu (Research Questions)	6
1.5 Mục tiêu, đối tượng, và phạm vi nghiên cứu	7
1.5.1 Mục tiêu nghiên cứu	7
1.5.2 Phạm vi nghiên cứu	7
1.5.3 Cấu trúc khóa luận tốt nghiệp	8
<b>CHƯƠNG 2. CÁC NGHIÊN CỨU LIÊN QUAN</b>	<b>10</b>
2.1 Giới thiệu tin giả và chiến dịch lừa đảo trực tuyến	10
2.1.1 Khái niệm Fake News	10
2.1.2 Các loại tin giả	11
2.1.3 Các dạng rối loạn thông tin	11

2.1.4	Chiến dịch lừa đảo trực tuyến . . . . .	12
2.2	Giới thiệu về Large Language Models . . . . .	14
2.2.1	Khái niệm . . . . .	14
2.2.2	Một số LLMs tiêu biểu . . . . .	14
2.2.3	Ứng dụng . . . . .	15
2.3	Học không mẫu (Zero-shot Prompting) . . . . .	16
2.4	Tinh chỉnh mô hình (Fine-tuning) . . . . .	18
2.5	Giới thiệu kỹ thuật truy xuất thông tin (RAG) . . . . .	19
2.5.1	Khái niệm . . . . .	19
2.5.2	Khả năng ứng dụng . . . . .	20
2.5.3	Kết hợp các kỹ thuật truy vấn . . . . .	20
2.5.4	Tìm kiếm dữ liệu mở với Google Search API . . . . .	25
2.6	Các công cụ hỗ trợ xác minh thông tin . . . . .	26
2.6.1	VirusTotal API . . . . .	26
2.6.2	Abstract Email API . . . . .	26
2.6.3	Abstract Phone API . . . . .	27
2.7	Multi-Agent AI trong hệ thống xác minh thông tin . . . . .	27
2.7.1	Khái niệm Multi AI Agent . . . . .	27
2.7.2	Giới thiệu về CrewAI Framework . . . . .	28
<b>CHƯƠNG 3.</b>	<b>PHƯƠNG PHÁP THỰC HIỆN</b>	<b>30</b>
3.1	Các mô hình sử dụng trong nghiên cứu . . . . .	30
3.1.1	Thử nghiệm các mô hình ban đầu . . . . .	30
3.1.2	Các mô hình được lựa chọn để Fine-tune . . . . .	31
3.2	Phương pháp Fine-tune (LoRA, Unsloth) . . . . .	32
3.2.1	LoRA (Low-Rank Adaptation) . . . . .	32
3.2.2	Unsloth – Framework tối ưu hoá Fine-tuning . . . . .	33
3.2.3	Quá trình huấn luyện . . . . .	34
3.3	Tổng quan kiến trúc hệ thống Multi-Agent đề xuất . . . . .	34

3.4	Các tác vụ của hệ thống Multi-Agent . . . . .	36
3.4.1	Tiếp nhận và phân tích đầu vào (Input Analyzer) . . . . .	36
3.4.2	Kiểm tra độ tin cậy của thực thể trích xuất (Entity Checker) . .	37
3.4.3	Phân tích tên miền và IP (Domain Analyst) . . . . .	38
3.4.4	Tìm kiếm thông tin mở trên internet (Web Researcher) . . . . .	38
3.4.5	Truy vấn thông tin từ cơ sở dữ liệu (Database Researcher) . . .	39
3.4.6	Tổng hợp và đánh giá cuối cùng (Final Verifier) . . . . .	40
3.5	Xây dựng cơ sở dữ liệu . . . . .	40
3.5.1	Bảng news_table - Tin tức đã xác minh . . . . .	41
3.5.2	Bảng ttp_table – Danh sách hành vi gian lận . . . . .	41
3.5.3	Bảng history_table – Lưu lịch sử tương tác . . . . .	42
3.6	Thiết kế phần mềm hệ thống . . . . .	43
3.6.1	Thiết kế hệ thống API phục vụ xác minh thông tin . . . . .	43
3.6.2	Các thành phần trong giao diện người dùng . . . . .	44
<b>CHƯƠNG 4.</b>	<b>THÍ NGHIỆM VÀ ĐÁNH GIÁ</b>	<b>47</b>
4.1	Thu thập và xây dựng bộ dữ liệu huấn luyện . . . . .	47
4.1.1	Nguồn dữ liệu và quá trình thu thập . . . . .	47
4.1.2	Đặc điểm của bộ dữ liệu . . . . .	48
4.1.3	Xử lý và gán nhãn dữ liệu . . . . .	49
4.1.4	Tăng cường dữ liệu . . . . .	50
4.2	Thiết lập huấn luyện mô hình . . . . .	50
4.3	Triển khai phần mềm ứng dụng . . . . .	53
4.3.1	API xử lý và phân tích thông tin . . . . .	53
4.3.2	Giao diện người dùng với Next.js . . . . .	54
4.3.3	Các thành phần trong giao diện người dùng . . . . .	55
4.4	Đánh giá và so sánh hiệu năng mô hình . . . . .	57
4.4.1	Các chỉ số đánh giá . . . . .	57
4.4.2	Phân tích, nhận xét kết quả . . . . .	57

<b>CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN</b>	<b>61</b>
5.1 Kết luận . . . . .	61
5.2 Hướng phát triển . . . . .	61
<b>TÀI LIỆU THAM KHẢO</b>	<b>62</b>

## **DANH MỤC CÁC CHỮ VIẾT TẮT**

API	Application Programming Interface
ATT	Adversarial Tactics, Techniques, and Common Knowledge
BM25	Best Matching 25
FAISS	Facebook AI Similarity Search
LLM	Large Language Model
LoRA	Low-Rank Adaptation
OSINT	Open Source Intelligence
PEFT	Parameter-Efficient Fine-Tuning
RAG	Retrieval-Augmented Generation
TTP	Tactics, Techniques, and Procedures
URC	Undecided Response Count



## DANH MỤC CÁC HÌNH VẼ

Hình 2.1	7 dạng nội dung tin giả phổ biến . . . . .	12
Hình 2.2	3 dạng rối loạn thông tin . . . . .	13
Hình 2.3	Hình ảnh minh họa chiến lược Học không mẫu . . . . .	17
Hình 2.4	Hình ảnh minh họa chiến lược Học không mẫu . . . . .	18
Hình 2.5	Kiến trúc Hybrid Search trong RAG . . . . .	21
Hình 2.6	Hình ảnh minh họa multi-agent system . . . . .	28
Hình 2.7	Hình ảnh minh họa quá trình CrewAI hoạt động . . . . .	29
Hình 3.1	Cơ chế cập nhật tham số trong LoRA . . . . .	32
Hình 3.2	Sơ đồ tổng quan hệ thống multi-agent . . . . .	35
Hình 3.3	Sơ đồ tổng quan phần mềm hệ thống hoạt động . . . . .	46
Hình 4.1	Giao diện hệ thống xác minh thông tin . . . . .	55
Hình 4.2	Khung nhập góp ý của người dùng về phản hồi xác minh . . . . .	56
Hình 4.3	Khung ánh xạ thông tin chiến dịch TTP liên quan . . . . .	56

## DANH MỤC CÁC BẢNG BIỂU

Bảng 3.1	Bảng news_table . . . . .	41
Bảng 3.2	Bảng ttp_table . . . . .	41
Bảng 3.3	Bảng history_table . . . . .	42
Bảng 3.4	Danh sách các API chính sử dụng trong hệ thống . . . . .	43
Bảng 4.1	Phân bố dữ liệu theo nhãn và từng tập huấn luyện . . . . .	50
Bảng 4.2	Các tham số huấn luyện chính trong quá trình fine-tune . . . . .	52
Bảng 4.3	So sánh hiệu suất các mô hình trước và sau tinh chỉnh . . . . .	58

## TÓM TẮT KHÓA LUẬN

Trong bối cảnh xã hội hiện đại với sự phát triển vượt bậc của công nghệ, đặc biệt là sự phổ biến và ứng dụng rộng rãi của trí tuệ nhân tạo (AI), người dùng ngày càng dễ dàng tiếp cận thông tin và tin tức một cách nhanh chóng và thuận tiện hơn bao giờ hết. Tuy nhiên, bên cạnh sự tiến bộ đó, các thủ đoạn lừa đảo trực tuyến cũng ngày càng tinh vi và phức tạp, khéo léo che giấu ranh giới giữa sự thật và giả dối. Công nghệ AI bị sử dụng để tự động sản sinh hàng loạt các thông tin mô phỏng gần như y hệt các tin tức thật, gây khó khăn lớn trong việc phân biệt thông tin chính xác và sai lệch. Điều này đặt ra thách thức nghiêm trọng đối với nhóm người dùng còn hạn chế kiến thức công nghệ và thiếu khả năng đánh giá thông tin, họ sẽ dễ dàng trở thành nạn nhân của các chiêu trò lừa đảo tinh vi.

Thực trạng này không chỉ đe dọa trực tiếp đến vấn đề bảo mật thông tin mà còn làm suy giảm nghiêm trọng niềm tin của xã hội vào môi trường truyền thông số, ảnh hưởng tiêu cực đến sự ổn định và phát triển bền vững của cộng đồng. Nhận thức được tính cấp thiết và tầm quan trọng của vấn đề, đề tài hướng tới việc phát triển một hệ thống hỗ trợ xác minh tin tức, dựa trên khả năng hiểu ngữ nghĩa một cách chuyên sâu và suy luận logic của các mô hình ngôn ngữ lớn (LLMs). Cho phép hệ thống tiếp cận và xử lý hiệu quả các nội dung đa dạng và phức tạp, qua đó nâng cao độ chính xác trong việc xác thực thông tin. Hệ thống được xây dựng theo kiến trúc đa tác tử AI (Multi-Agent Architecture), trong đó mỗi tác tử đảm nhận các nhiệm vụ chuyên biệt và phối hợp chặt chẽ nhằm tối ưu quy trình kiểm chứng thông tin, phân loại tin tức dựa trên các chiến lược lừa đảo phổ biến như APT (Advanced Persistent Threat) hay các kỹ thuật TTP (Tactics, Techniques, and Procedures).

Với phương pháp tiếp cận toàn diện và ứng dụng sâu rộng các tiên bộ của AI, đề tài kỳ vọng góp phần nâng cao hiệu quả và độ chính xác trong quy trình kiểm chứng

thông tin, đồng thời tăng cường tính minh bạch và khả năng giải trình của hệ thống, từ đó củng cố niềm tin của người dùng và thúc đẩy sự phát triển bền vững của môi trường truyền thông số.

**Từ khóa:** Fake News Detection, Multi AI Agent System, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Semantic Understanding, Online Fraud Detection, Zero-shot Learning, Fine-tuning

# CHƯƠNG 1. TỔNG QUAN

Chương này chúng tôi sẽ giới thiệu bối cảnh, lý do chọn đề tài, mục tiêu và phạm vi nghiên cứu. Đồng thời, trình bày các câu hỏi nghiên cứu, phương pháp tiếp cận và tóm tắt cấu trúc của toàn khóa luận.

## 1.1. Giới thiệu vấn đề

Trong kỷ nguyên số, sự bùng nổ thông tin trên Internet và mạng xã hội đã làm thay đổi hoàn toàn cách con người tiếp cận và chia sẻ thông tin. Tuy nhiên, đi cùng với sự phát triển đó là sự gia tăng nhanh chóng của các loại tin tức giả mạo, thông tin sai lệch và các hình thức lừa đảo trực tuyến ngày càng tinh vi. Những nội dung độc hại này lan truyền với tốc độ chóng mặt qua nhiều kênh như mạng xã hội, email, tin nhắn và diễn đàn trực tuyến, gây ảnh hưởng nghiêm trọng đến nhận thức, hành vi và cả tài sản của cộng đồng. Trong khi đó, việc phân biệt thật và giả ngày càng trở nên khó khăn hơn bao giờ hết bởi các kỹ thuật ngày một tinh vi. Ngày nay, các tin tức giả mạo được trình bày ngắn gọn, đánh vào cảm xúc, khó kiểm chứng và thiếu vắng các hệ thống kiểm tra thông tin tự động, hiệu quả và được cập nhật liên tục.

Theo khảo sát từ Statista năm 2022, có đến hơn 70% người dùng Internet toàn cầu từng tiếp xúc với thông tin sai lệch, chủ yếu từ mạng xã hội [15]. Đặc biệt, trong môi trường số đang phát triển như Việt Nam, tình trạng giả danh cơ quan nhà nước, ngân hàng, hoặc tuyển dụng ảo ngày càng phổ biến, khiến người dân dễ trở thành nạn nhân. Thực tế cho thấy, các vụ lừa đảo mạng hiện nay không chỉ đơn thuần là phát tán thông tin sai sự thật mà còn nhắm vào việc đánh cắp thông tin cá nhân, chiếm đoạt tài sản, gây mất niềm tin vào truyền thông chính thống. Trong khi đó, người dùng hiện nay vẫn thiếu các công cụ hiệu quả để tự động kiểm tra và xác minh thông tin, nhất là với

các dạng nội dung ngắn như tin nhắn, email, hay bài đăng mạng xã hội.

Trước thực trạng đó, đề tài "Hệ thống tự động phát hiện và kiểm chứng tin tức giả mạo trong các chiến dịch lừa đảo trực tuyến bằng LLMs" được lựa chọn với mong muốn đề xuất một giải pháp ứng dụng các công nghệ tiên tiến như mô hình ngôn ngữ lớn (LLMs) và Multi-Agent Systems để hỗ trợ người dùng phát hiện và xác minh thông tin một cách chính xác, minh bạch và theo thời gian thực. Đây là hướng đi cần thiết và cấp thiết nhằm nâng cao khả năng phòng chống tin giả, bảo vệ người dùng trong không gian mạng ngày càng phức tạp.

## **1.2. Giới thiệu các hướng tiếp cận phổ biến**

Chủ đề phát hiện và kiểm chứng tin giả nhận được sự quan tâm rộng rãi trong những năm gần đây, đặc biệt trong bối cảnh tin tức sai lệch và lừa đảo trực tuyến ngày càng gia tăng. Nhiều nghiên cứu học thuật đã tập trung xây dựng mô hình phân loại, xác minh thông tin, cũng như cơ chế lan truyền của tin giả trên mạng xã hội. Các bộ dữ liệu tiêu biểu như FakeNewsNet [13] và LIAR [17] đã được phát triển nhằm phục vụ cho việc huấn luyện và đánh giá các mô hình phát hiện tin tức sai lệch.

Một nhánh nghiên cứu phổ biến là ứng dụng học máy truyền thống (Machine Learning) để phân loại tin giả hoặc phát hiện email lừa đảo (phishing) dựa trên các đặc trưng [2], [14]. Tuy nhiên, hầu hết còn thiếu tính cập nhật, khả năng giải thích và hạn chế trong xử lý các loại nội dung ngắn.

Ở một hướng tiếp cận khác, một số nghiên cứu như MedOSINT đã ứng dụng phương pháp lập luận theo ví dụ (Case-Based Reasoning - CBR) [11]. Phương pháp này không đòi hỏi quá trình huấn luyện phức tạp, mà dựa vào so sánh nội dung mới với các trường hợp tương tự trong một kho tri thức (case base) đã được xây dựng [3]. CBR đặc biệt phù hợp trong các hệ thống yêu cầu giải thích kết luận một cách rõ ràng, giúp tăng mức độ tin cậy và minh bạch của hệ thống khi đưa ra kết luận.

Trong những năm gần đây, sự phát triển mạnh mẽ của mô hình ngôn ngữ lớn

(LLMs) đã mở ra nhiều khả năng mới trong việc hiểu và xử lý ngữ nghĩa của tin tức. Các mô hình có khả năng xử lý đa tác vụ, thậm chí không cần huấn luyện lại nhờ khả năng zero-shot learning. Một số hướng tiếp cận hiệu quả là kết hợp LLM với Retrieval-Augmented Generation (RAG) nhằm cung cấp dữ liệu thực tế từ các nguồn chính thống đáng tin cậy hoặc cơ sở dữ liệu OSINT [21].

Đề tài này kế thừa các hướng tiếp cận trên, nhằm phát triển một hệ thống kiểm chứng tin tức hiện đại, tổ chức LLM dưới dạng hệ thống AI đa tác tử (Multi-Agent System), trong đó mỗi AI Agent đảm nhiệm một chức năng cụ thể như tóm tắt yêu cầu, truy xuất dữ liệu, tìm kiếm thông tin, kiểm tra độ an toàn của liên kết hoặc đánh giá mức độ đáng tin cậy của nguồn phát tán. Cách tiếp cận này không chỉ nâng cao hiệu quả kiểm chứng mà còn giúp hệ thống có thể hoạt động tự động hóa, theo thời gian thực và có khả năng mở rộng trong các môi trường tiếng Việt và đa nền tảng.

### **1.3. Những thách thức**

#### ***1.3.1. Giới hạn độ dài ngữ cảnh và duy trì liên kết thông tin***

Các mô hình LLM hiện nay bị giới hạn về độ dài ngữ cảnh đầu vào, điều này hạn chế khả năng xử lý chuỗi thông tin dài và phức tạp trong các bài viết hoặc chuỗi hội thoại liên quan đến tin tức giả mạo. Khi kiểm chứng tin tức cần tổng hợp dữ liệu đa nguồn hoặc phân tích chuỗi sự kiện, mô hình dễ bị mất thông tin quan trọng do không thể duy trì liên tục bối cảnh và mạch lạc trong xử lý.

#### ***1.3.2. Đảm bảo tính chính xác và cập nhật nguồn dữ liệu***

Việc truy xuất thông tin từ các nguồn mở như Wikipedia, báo chí và cơ sở dữ liệu OSINT đòi hỏi nguồn dữ liệu phải chính xác, cập nhật và phù hợp với ngữ cảnh kiểm chứng. Đồng thời, nguồn dữ liệu về tin tức giả mạo và tin sai lệch dành cho tiếng Việt còn rất hạn chế về số lượng và chất lượng, ảnh hưởng đến khả năng xây dựng các mô hình nhận diện chính xác và toàn diện. Bên cạnh đó, các chiêu trò lừa đảo và tin

giả liên tục thay đổi về hình thức và chiến thuật, đòi hỏi hệ thống phải có khả năng linh hoạt thích ứng và cập nhật kịp thời nhằm duy trì hiệu quả kiểm chứng trong môi trường đa dạng và biến động.

### ***1.3.3. Phối hợp và đồng bộ trong hệ thống Multi-Agent***

Thiết kế và vận hành hệ thống đa tác tử với nhiều AI Agent đảm nhận các chức năng riêng biệt theo trình tự, trong đó mỗi agent thực hiện xong nhiệm vụ và chuyển giao thông tin đầu ra cho agent tiếp theo xử lý. Cách phối hợp này đòi hỏi một kiến trúc phức tạp nhưng hiệu quả. Bên cạnh đó, việc thực hiện nhiều tác vụ AI có thể gây tốn kém đáng kể về thời gian xử lý và chi phí tài nguyên, đặc biệt khi hệ thống xử lý lượng lớn dữ liệu hoặc hoạt động theo thời gian thực. Do vậy, cân bằng giữa độ chính xác, tốc độ phản hồi và chi phí vận hành là một thách thức lớn trong phát triển hệ thống.

## **1.4. Câu hỏi nghiên cứu (Research Questions)**

Dựa trên thực tiễn phát sinh ngày càng nhiều hình thức thông tin sai lệch và hành vi lừa đảo tinh vi trên môi trường số, cùng với yêu cầu phát triển một hệ thống xác minh tin tức tự động, khóa luận tập trung giải quyết các câu hỏi nghiên cứu sau:

1. Làm thế nào để xây dựng một tập dữ liệu đa dạng, đại diện cho nhiều loại thông tin sai lệch và hành vi lừa đảo để huấn luyện và đánh giá mô hình ngôn ngữ trong xác minh thông tin?
2. Làm thế nào để nâng cao hiệu suất và độ ổn định của mô hình ngôn ngữ lớn trong kiểm chứng tin tức tiếng Việt, đồng thời đảm bảo phản hồi đúng yêu cầu và tránh lỗi mơ hồ?
3. Làm thế nào để xây dựng hệ thống xác minh tự động, xử lý đa tác vụ với các công cụ xác thực tin cậy, cập nhật và giải quyết yêu cầu xác minh với tin tức mới?



4. Làm thế nào để cung cấp lời giải thích minh bạch về lý do đánh giá nội dung là đáng ngờ và đối chiếu với các chiến dịch lừa đảo phổ biến, tăng cường độ tin cậy của hệ thống?

## **1.5. Mục tiêu, đối tượng, và phạm vi nghiên cứu**

### ***1.5.1. Mục tiêu nghiên cứu***

Mục tiêu chính của đề tài là phát triển một hệ thống kiểm chứng tin tức tự động cho môi trường tiếng Việt, có khả năng phân tích và xác thực thông tin từ nhiều nguồn đầu vào như bài viết trên mạng xã hội, email, tin nhắn và các trang báo điện tử. Hệ thống sẽ ứng dụng nền tảng mô hình ngôn ngữ lớn (LLM), kết hợp kiến trúc đa tác tử (Multi-Agent System) và các công cụ kiểm chứng từ nguồn mở nhằm tăng cường tính chính xác, minh bạch và khả năng phản hồi thời gian thực trong quá trình xác minh.

Nghiên cứu sẽ đi sâu vào đánh giá hiệu suất của hệ thống dựa trên các tiêu chí trọng yếu như độ chính xác, tốc độ phản hồi và khả năng mở rộng nhằm đảm bảo tính khả thi và ứng dụng thực tế. Cuối cùng, xây dựng một giao diện trực quan nhằm hỗ trợ người dùng kiểm tra và xác thực tin tức nhanh chóng, tiện lợi, góp phần nâng cao nhận thức cộng đồng và giảm thiểu tác động tiêu cực của tin tức giả mạo trong xã hội hiện đại.

### ***1.5.2. Phạm vi nghiên cứu***

Đề tài tập trung vào việc xây dựng và đánh giá một hệ thống kiểm chứng tin tức tự động trong môi trường tiếng Việt, kết hợp sức mạnh của mô hình ngôn ngữ lớn (LLM), kiến trúc hệ thống đa tác tử (Multi-Agent) và các công cụ xác minh thông tin từ nguồn mở. Trọng tâm đầu tiên của nghiên cứu là thu thập và xây dựng một tập dữ liệu mới phục vụ bài toán phát hiện tin giả tiếng Việt. Tập dữ liệu được tổng hợp từ các nguồn báo chí chính thống, các bài viết giả mạo đã được xác minh, cùng những nội dung lừa đảo phổ biến trên mạng xã hội. Sau khi thu thập, dữ liệu được xử lý

thông qua các bước làm sạch, gán nhãn và tăng cường nhằm đảm bảo sự đa dạng về ngữ nghĩa và độ cân bằng giữa các lớp tin thật và tin giả.

Tiếp theo, nghiên cứu đánh giá khả năng áp dụng của một số mô hình ngôn ngữ lớn phổ biến như LLaMA, Qwen, Mistral và Gemma trong việc phát hiện tin giả tiếng Việt. Hai phương pháp được triển khai là zero-shot learning và fine-tuning để so sánh hiệu quả của các mô hình trên cùng bộ dữ liệu, từ đó xác định mức độ phù hợp và tiềm năng ứng dụng của các LLMs trong ngữ cảnh tiếng Việt.

Cuối cùng, đề tài triển khai một hệ thống kiểm chứng thông tin dựa trên mô hình GPT-4o kết hợp với kiến trúc đa tác tử. Mỗi agent trong hệ thống đảm nhiệm một vai trò cụ thể, chẳng hạn như phân tích nội dung đầu vào, nhận diện thực thể quan trọng (URL, email, số điện thoại), và xác minh rủi ro bằng cách tích hợp với các công cụ nguồn mở như VirusTotal, AbstractAPI và Google Search. Hệ thống được thiết kế để xử lý cả văn bản và hình ảnh, hỗ trợ kiểm chứng tin tức một cách tự động, minh bạch, và có khả năng mở rộng phù hợp với môi trường tiếng Việt.

### ***1.5.3. Cấu trúc khóa luận tốt nghiệp***

Khóa luận được tổ chức thành 5 chương chính và phần mở đầu, được trình bày theo tiến trình nghiên cứu và phát triển hệ thống, nhằm cung cấp cho người đọc cái nhìn toàn diện về vấn đề và giải pháp đề xuất:

- Chương 1 – Tổng quan: Giới thiệu về bối cảnh, lý do chọn đề tài, mục tiêu và phạm vi nghiên cứu. Chương này cung cấp cái nhìn tổng quan về tình trạng hiện tại, đồng thời đặt ra các câu hỏi nghiên cứu cần đi giải quyết và tóm tắt cấu trúc của toàn khóa luận.
- Chương 2 – Các nghiên cứu liên quan: Trình bày các kiến thức nền tảng phục vụ nghiên cứu, bao gồm khái niệm fake news và lừa đảo trực tuyến, LLMs, kỹ thuật truy xuất (RAG), fine-tuning, kiến trúc Multi-Agent và các công cụ hỗ trợ xác minh.

- Chương 3 – Phương pháp nghiên cứu: Mô tả chi tiết quy trình huấn luyện mô hình trên dữ liệu tiếng Việt, các kỹ thuật fine-tuning sử dụng LoRA và Unsloth, thiết kế kiến trúc hệ thống Multi-Agent và mô tả từng tác vụ riêng biệt của các AI Agent.
- Chương 4 – Thử nghiệm và đánh giá hệ thống: Trình bày cách xây dựng tập dữ liệu huấn luyện, các bước tiền xử lý, gán nhãn, tăng cường dữ liệu, và thiết lập huấn luyện mô hình. Sau đó là các thí nghiệm đánh giá hiệu năng, độ chính xác và hiệu suất của hệ thống trên các kịch bản thực tế.
- Chương 5 – Kết luận và hướng phát triển: Tổng kết những kết quả đạt được trong quá trình nghiên cứu và triển khai hệ thống. Chương này cũng nêu ra những hạn chế hiện tại và đề xuất các hướng cải tiến trong tương lai nhằm hoàn thiện hệ thống hơn nữa.

## CHƯƠNG 2. CÁC NGHIÊN CỨU LIÊN QUAN

Chương này trình bày các nền tảng lý thuyết và công nghệ liên quan phục vụ xây dựng hệ thống, bao gồm các khái niệm như tin giả, LLMs, kỹ thuật RAG, phương pháp fine-tuning, các công cụ xác minh thông tin, cùng kiến trúc hệ thống Multi-Agent AI.

### 2.1. Giới thiệu tin giả và chiến dịch lừa đảo trực tuyến

#### 2.1.1. Khái niệm Fake News

Trong bối cảnh dòng chảy thông tin không ngừng của kỷ nguyên số, khái niệm “tin tức giả mạo” (Fake News) đã nổi lên như một vấn đề cấp bách. Đây không chỉ đơn thuần là thông tin sai lệch mà còn là những nội dung được xây dựng và phát tán có chủ đích nhằm mục tiêu gây hiểu lầm nghiêm trọng, thao túng nhận thức công chúng, hoặc phục vụ các động cơ cá nhân, chính trị, hay thương mại. Theo Ireton & Posetti, tin giả được định nghĩa là những “thông tin được tạo ra, trình bày dưới dạng một bản tin, nhưng không phản ánh thực tế, với mục đích đánh lừa người đọc.” [8].

Khái niệm này thường bị lẫn lộn với các thuật ngữ tương cận như misinformation (thông tin sai lệch không có chủ ý) và disinformation (thông tin sai lệch có chủ đích). Tuy nhiên, bất kể hình thái nào, tính chất nguy hiểm của tin giả vẫn luôn hiện hữu. Điều đó càng được lộ rõ hơn khi mà truyền thông xã hội và trí tuệ nhân tạo đang trong đà phát triển mạnh mẽ, tin tức giả mạo không còn chỉ giới hạn ở dạng văn bản thuần túy. Chúng còn biến đổi một cách tinh vi thành hình ảnh đã qua chỉnh sửa, hoặc thậm chí là các nội dung được AI tổng hợp hoàn toàn. Chính sự đa dạng về mặt hình thức và tính chân thực một cách giả tạo này đã biến việc phát hiện và kiểm chứng tin giả thành một thách thức ngày càng gia tăng về mức độ phức tạp, đòi hỏi sự phát triển của các giải pháp và công nghệ tiên tiến [9].

### **2.1.2. Các loại tin giả**

Tin giả hiện nay được phân loại thành nhiều hình thức khác nhau dựa trên mục đích, phương thức tạo lập và mức độ gây hại. Theo phân loại của Wardle & Derakhshan [18], có ít nhất 7 loại tin giả phổ biến:

**Nội dung gây hiểu nhầm (Misleading content):** Thông tin đúng nhưng bị diễn giải sai lệch, làm sai lệch bản chất vấn đề

**Nội dung nguy tạo (Fabricated Content):** Tin tức hoàn toàn bịa đặt, không có bất kỳ căn cứ thực tế nào.

**Nội dung sai ngữ cảnh (False Context):** Thông tin thật nhưng được đưa vào bối cảnh sai để đánh lừa người đọc.

**Nội dung sai nguồn (False Connection):** Tiêu đề, hình ảnh hoặc đoạn trích không liên quan đến nội dung thực tế.

**Nội dung châm biếm (Satire/Parody):** Nội dung hài hước không có mục đích lừa đảo nhưng dễ bị hiểu lầm là tin thật.

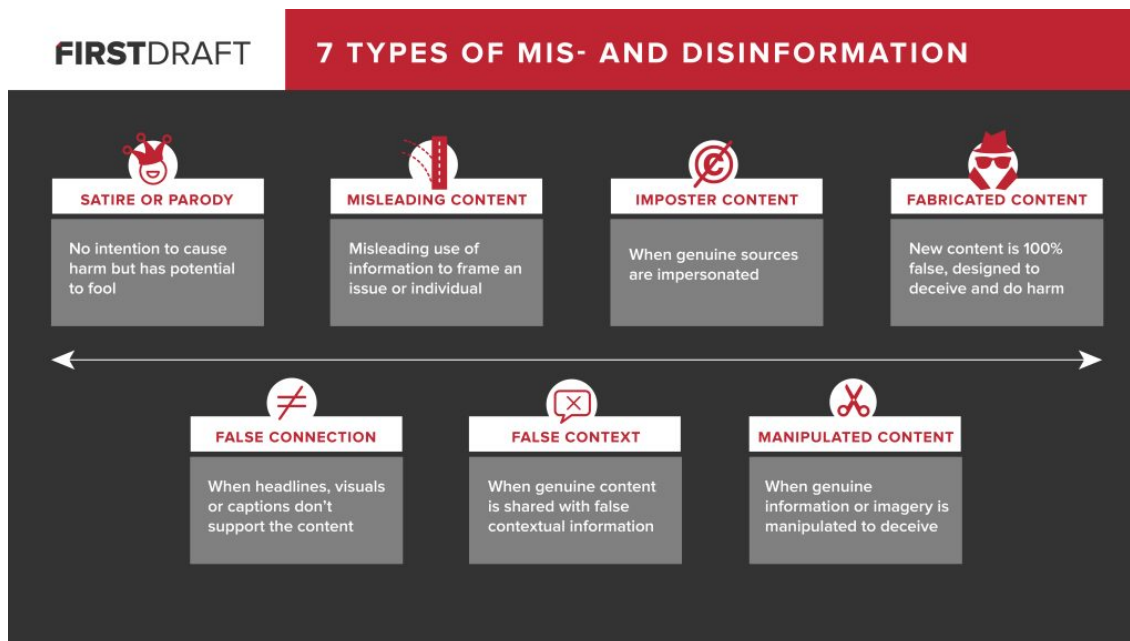
**Nội dung bị chỉnh sửa (Manipulated Content):** Hình ảnh hoặc tài liệu đã bị thay đổi nhằm bóp méo sự thật.

**Nội dung giả dạng danh tính (Imposter Content):** Giả mạo thông tin từ các nguồn uy tín như báo chí, tổ chức nhà nước hoặc người nổi tiếng.

Trên các nền tảng mạng xã hội như Facebook, Zalo, Telegram, một số dạng tin giả còn xuất hiện dưới dạng liên kết giả mạo (phishing), tin tuyển dụng ảo, “công ty ma” hoặc các ứng dụng đầu tư trực tuyến lừa đảo. Chúng thường sử dụng tiêu đề giật gân, hình ảnh gây sốc và ngôn ngữ cảm xúc để thu hút người đọc và tăng tốc độ lan truyền.

### **2.1.3. Các dạng rối loạn thông tin**

Trong bối cảnh số, khái niệm "tin giả" thường bị sử dụng một cách mơ hồ, bao quát cả những nội dung sai lệch, xuyên tạc và thậm chí là thông tin thật nhưng bị khai thác sai mục đích. Theo Wardle & Derakhshan, để hiểu đúng bản chất của các hành vi thao



**Hình 2.1:** 7 dạng nội dung tin giả phổ biến

túng thông tin, cần phân biệt rõ ba loại rối loạn thông tin (information disorder), dựa trên hai yếu tố chính: tính chính xác và mức độ gây hại.

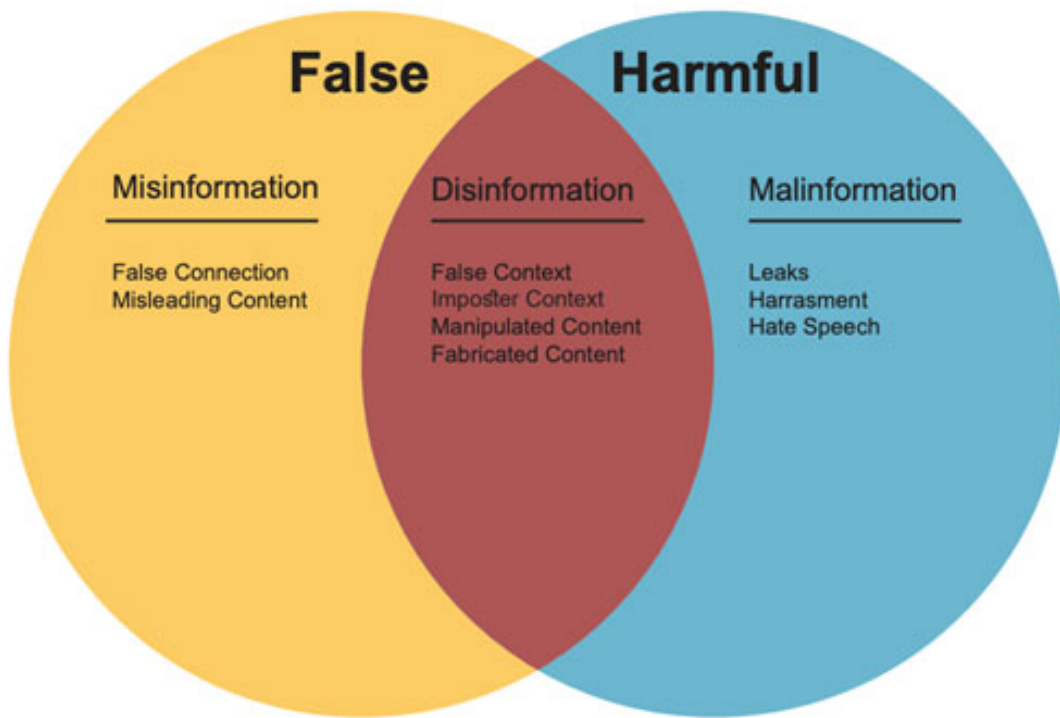
**Misinformation:** thông tin sai sự thật, nhưng không được tạo ra với mục đích gây hại. Nó thường xuất phát từ việc người dùng chia sẻ nội dung mà họ tin là đúng, hoặc không kiểm chứng trước khi lan truyền.

**Disinformation:** thông tin sai sự thật và được tạo ra có chủ đích để gây tổn hại cho cá nhân, nhóm xã hội, tổ chức hoặc quốc gia. Hình thức này thường được sử dụng trong các chiến dịch thao túng dư luận, phá hoại chính trị hoặc gieo rắc hoang mang.

**Malinformation:** thông tin có thật, nhưng bị khai thác sai bối cảnh hoặc mục đích nhằm gây hại. Đây có thể là những thông tin riêng tư, phát ngôn cũ bị trích lại sai thời điểm, hoặc dữ liệu bị rò rỉ có chủ đích.

#### 2.1.4. Chiến dịch lừa đảo trực tuyến

Các chiến dịch lừa đảo trực tuyến (Online Fraud Campaigns) ngày nay không còn đơn lẻ mà thường được tổ chức theo cấu trúc bài bản, có hệ thống hạ tầng riêng, kịch



**Hình 2.2:** 3 dạng rối loạn thông tin

bản lừa đảo được chuẩn hóa và chiến thuật phát tán rõ ràng. Chúng khai thác yếu tố tâm lý và hành vi người dùng qua các nền tảng như mạng xã hội, SMS, email, hoặc ứng dụng di động.

Khác với phương pháp phát hiện phishing truyền thống vốn chỉ dựa vào URL hay giao diện giả, các chiến dịch hiện đại thường sử dụng nội dung có tính kể chuyện, xen lẫn thông tin giả mạo để dẫn dụ người dùng. Việc này đòi hỏi hệ thống phải hiểu được ngữ cảnh thông tin và ý đồ thao túng, thay vì chỉ nhận diện tín hiệu kỹ thuật.

Một hướng tiếp cận mới là ánh xạ các hành vi lừa đảo về các Tactic, Technique, and Procedure (TTP), từ đó nhận diện và phân loại các chiến dịch giả mạo theo cấu trúc tương tự như khung MITRE ATT&CK. Cách làm này không chỉ phục vụ cho cảnh báo sớm mà còn hỗ trợ phân tích mối đe dọa ở cấp độ chiến dịch. Trong đề tài này, sau bước xác minh thông tin, hệ thống sẽ tiếp tục gắn nhãn chiến dịch theo tập Fraud-TTP nhằm tăng khả năng phân loại và phản ứng sớm với các hình thức lừa đảo có tổ chức.

## 2.2. Giới thiệu về Large Language Models

### 2.2.1. Khái niệm

Mô hình ngôn ngữ lớn (LLM) là các mô hình học sâu với số lượng tham số rất lớn, được huấn luyện trên tập dữ liệu văn bản khổng lồ với hàng tỷ, thậm chí hàng nghìn tỷ tham số, nhằm học cách sinh và hiểu ngôn ngữ tự nhiên. Các mô hình này được xây dựng chủ yếu dựa trên kiến trúc Transformer, vốn nổi bật nhờ khả năng xử lý song song và ghi nhớ quan hệ ngữ cảnh dài hạn trong văn bản.

Cơ chế hoạt động cơ bản của LLMs xoay quanh việc học cách dự đoán từ tiếp theo trong một chuỗi văn bản dựa trên ngữ cảnh đã biết, hoặc điền vào chỗ trống trong câu. Quá trình này giúp mô hình không chỉ nắm bắt được cú pháp mà còn cả ngữ nghĩa và ý đồ giao tiếp của ngôn ngữ tự nhiên.

Khả năng nổi bật của LLMs là tổng quát hóa kiến thức từ dữ liệu huấn luyện, cho phép chúng thực hiện đa dạng tác vụ ngôn ngữ như sinh văn bản, dịch máy, tóm tắt, trả lời câu hỏi, nhận diện thực thể và thậm chí lập luận có điều kiện – mà không cần huấn luyện riêng biệt cho từng tác vụ. Chính vì vậy, LLMs ngày càng trở thành nền tảng quan trọng cho các ứng dụng xử lý ngôn ngữ tự nhiên (NLP) hiện đại.

### 2.2.2. Một số LLMs tiêu biểu

Trong những năm gần đây, nhiều mô hình ngôn ngữ lớn (LLMs) đã được giới thiệu và đóng vai trò quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên. Các mô hình này khác biệt về quy mô, kiến trúc, dữ liệu huấn luyện và khả năng thích ứng với từng tác vụ cụ thể.

Một trong những mô hình tiên phong là GPT (Generative Pre-trained Transformer) do OpenAI phát triển [12]. Với các phiên bản từ GPT-2 đến GPT-4, mô hình này đã chứng minh khả năng sinh ngôn ngữ tự nhiên mạch lạc, linh hoạt và phù hợp với nhiều nhiệm vụ đa dạng như viết lại văn bản, trả lời câu hỏi và thậm chí lập luận logic. Có



thể nói, GPT-4 đang được đánh giá cao nhờ khả năng hiểu ngữ cảnh rộng, hỗ trợ xử lý đa phương thức (multimodal) và có khả năng truy xuất thông tin ngoài mô hình thông qua API.

Bên cạnh đó, mô hình BERT (Bidirectional Encoder Representations from Transformers) [5] của Google lại hướng tới việc hiểu sâu ngữ cảnh bằng cách mã hóa văn bản theo cả hai chiều. BERT đã đạt thành tích vượt trội trong các bài toán như phân loại văn bản, nhận diện thực thể có tên (NER), và đánh giá mức độ liên quan giữa các câu. Các biến thể như RoBERTa (Facebook AI), ALBERT (Google), và DeBERTa (Microsoft) đã cải thiện hiệu quả huấn luyện, độ chính xác và khả năng tổng quát.

Không dừng lại ở đó, Google tiếp tục phát triển mô hình T5 (Text-to-Text Transfer Transformer) và đánh dấu một hướng tiếp cận mới khi biểu diễn mọi tác vụ NLP dưới dạng chuyển đổi văn bản – từ đầu vào đến đầu ra đều là chuỗi ngôn ngữ tự nhiên. Điều này không chỉ đơn giản hóa quy trình xây dựng hệ thống mà còn nâng cao khả năng tùy biến và mở rộng trong các ứng dụng phức tạp.

Ngoài ra, nhiều mô hình mã nguồn mở mạnh mẽ cũng đã được phát hành nhằm hỗ trợ nghiên cứu và triển khai thực tế. Những mô hình này có kích thước đa dạng, từ nhỏ gọn phù hợp với thiết bị giới hạn tài nguyên đến các phiên bản lớn đạt hiệu suất cao trên các benchmark chuẩn. Các thư viện và nền tảng như Hugging Face Transformers, LangChain, LlamaIndex và OpenLLM đóng vai trò quan trọng trong việc triển khai, tích hợp và điều phối các LLMs, giúp rút ngắn thời gian phát triển ứng dụng và tối ưu hiệu quả sử dụng mô hình.

### **2.2.3. Ứng dụng**

Sự bùng nổ dữ liệu trên không gian mạng đã tạo điều kiện thuận lợi cho việc lan truyền nhanh chóng các tin tức giả và thông tin sai lệch. Trong bối cảnh đó, phương pháp kiểm chứng thông tin thủ công dần trở nên bất khả thi, đòi hỏi những giải pháp tự động hóa hiệu quả và có khả năng mở rộng. Trong bối cảnh đó, các mô hình ngôn ngữ lớn (LLMs) đã trở thành công cụ nền tảng, giúp nâng cao hiệu quả phân tích và

xác minh tin tức trong môi trường số.

Các mô hình này có thể đảm nhiệm nhiều tác vụ cốt lõi trong quy trình phát hiện và kiểm chứng tin giả. Cụ thể, LLMs có thể nhận diện thực thể như tên người, địa điểm, tổ chức, cũng như tóm tắt nội dung quan trọng. Nhờ đó, giúp người dùng nắm bắt được ý chính của văn bản dài và hỗ trợ so sánh thông tin với các nguồn dữ liệu đáng tin cậy để phát hiện mâu thuẫn, dấu hiệu thao túng ngôn ngữ hoặc cảm xúc cực đoan – những đặc điểm thường thấy trong tin tức giả mạo. Ngoài ra, LLMs còn có thể tổng hợp dữ liệu từ nhiều nguồn và thực hiện tác vụ hỏi – đáp, đưa ra lập luận logic và minh bạch nhằm hỗ trợ đánh giá tính xác thực của thông tin một cách khách quan và hiệu quả.

Hai nhóm mô hình ngôn ngữ lớn phổ biến được ứng dụng trong đề tài:

**Các mô hình mã nguồn mở từ Hugging Face:** bao gồm LLaMA, Mistral, Qwen và Gemma. Những mô hình này có ưu điểm lớn về khả năng tùy biến, cho phép tinh chỉnh (fine-tune) trên tập dữ liệu tiếng Việt nhằm thích ứng với bối cảnh cụ thể. Nhờ tính linh hoạt cao và chi phí vận hành thấp, nhóm mô hình này phù hợp cho các nghiên cứu học thuật hoặc ứng dụng chuyên biệt đòi hỏi kiểm soát toàn bộ pipeline xử lý.

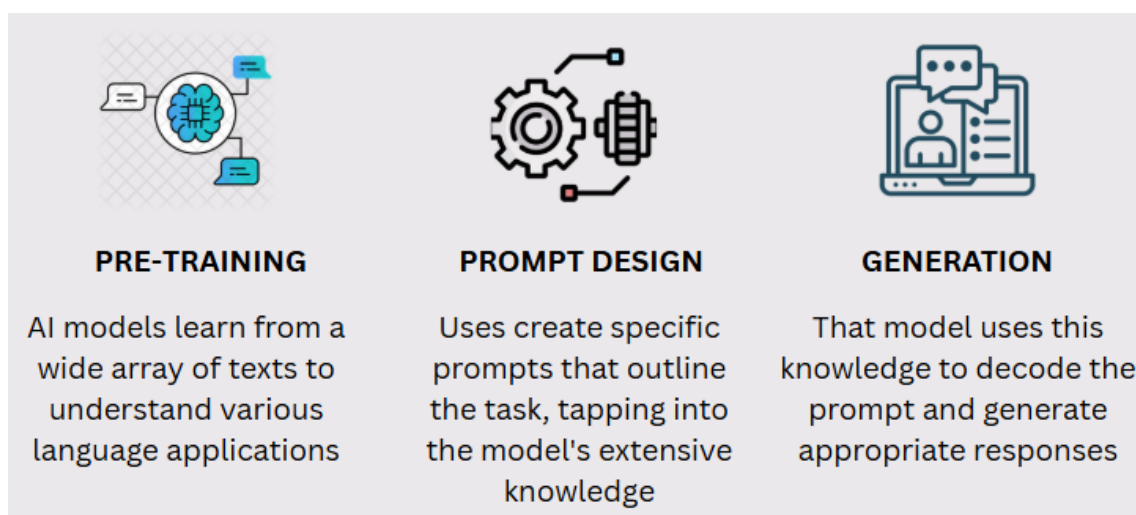
**Mô hình GPT của OpenAI:** nổi bật với độ chính xác cao, khả năng hiểu ngôn ngữ sâu và xử lý đa phương thức (multimodal), hỗ trợ cả văn bản và hình ảnh. Với khả năng truy xuất kiến thức mạnh mẽ thông qua API, GPT đặc biệt hiệu quả trong các tác vụ yêu cầu tổng hợp thông tin, lập luận logic và đánh giá nội dung trên quy mô lớn. Tuy nhiên, điểm hạn chế là mô hình không thể tinh chỉnh trực tiếp và thường đòi hỏi chi phí sử dụng cao.

## 2.3. Học không mẫu (Zero-shot Prompting)

Trong bối cảnh sử dụng mô hình ngôn ngữ lớn (LLMs), chiến lược zero-shot prompting – một biến thể ứng dụng của zero-shot learning – cho phép mô hình thực hiện các tác vụ mà không cần được huấn luyện trước trên các ví dụ cụ thể của tác vụ

đó. Thay vì học từ dữ liệu được gán nhãn như trong học có giám sát truyền thống, phương pháp này dựa vào kiến thức đã được tích lũy trong quá trình tiền huấn luyện, và từ đó suy luận câu trả lời chỉ bằng cách sử dụng một câu lệnh hướng dẫn (prompt) được viết bằng ngôn ngữ tự nhiên.

Phương pháp này đặc biệt phù hợp với các mô hình ngôn ngữ lớn (LLMs) đã được tiền huấn luyện trên một kho dữ liệu văn bản khổng lồ, cho phép chúng nắm bắt cấu trúc ngôn ngữ, quan hệ ngữ nghĩa và tri thức phổ quát. Khi áp dụng vào một nhiệm vụ cụ thể, người dùng chỉ cần thiết kế một prompt – tức một câu lệnh hoặc mô tả ngắn bằng ngôn ngữ tự nhiên – để hướng dẫn mô hình hiểu và thực hiện tác vụ được yêu cầu (Hình 2.3). Quá trình này không cần huấn luyện thêm và mô hình sẽ sinh ra kết quả đầu ra trực tiếp dựa trên kiến thức đã học trước đó.



**Hình 2.3:** Hình ảnh minh họa chiến lược Học không mẫu

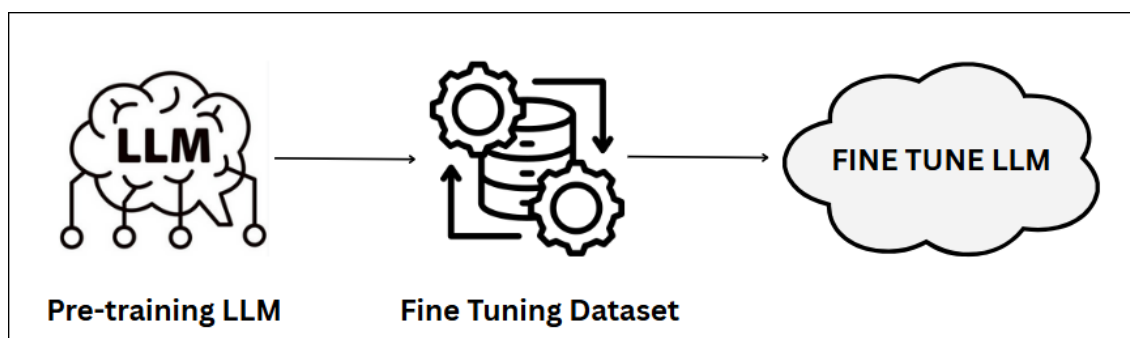
Trong bối cảnh kiểm chứng và phát hiện tin giả, zero-shot mang lại nhiều lợi thế thực tiễn. Việc không phụ thuộc vào dữ liệu huấn luyện giúp hệ thống phản hồi nhanh với các nội dung mới xuất hiện trên không gian mạng – nơi mà thông tin sai lệch thường thay đổi liên tục và khó lường. Hơn nữa, khả năng tổng quát hóa cao của LLMs giúp hệ thống dễ dàng xử lý nhiều kiểu văn bản khác nhau, từ bài viết chính thống đến các bình luận phi cấu trúc trên mạng xã hội.

Tuy nhiên, phương pháp này cũng tồn tại một số hạn chế nhất định. Việc phụ thuộc

hoàn toàn vào prompt khiến hiệu quả suy luận của mô hình dễ bị ảnh hưởng bởi cách diễn đạt và độ rõ ràng của hướng dẫn đầu vào. Ngoài ra, trong các ngữ cảnh ngôn ngữ đặc thù như tiếng Việt, mô hình tiền huấn luyện đa ngữ có thể chưa hiểu rõ sắc thái biểu đạt, ngôn ngữ mạng hay từ vựng địa phương, dẫn đến kết quả thiếu chính xác. Những hạn chế này cho thấy, để nâng cao hiệu quả của hệ thống kiểm chứng trong môi trường ngôn ngữ cụ thể và nội dung nhạy cảm, cần phải kết hợp thêm kỹ thuật tinh chỉnh mô hình (fine-tuning) để tối ưu hóa hiệu năng.

## 2.4. Tinh chỉnh mô hình (Fine-tuning)

Tinh chỉnh mô hình (fine-tuning) là quá trình điều chỉnh lại một mô hình đã được tiền huấn luyện, bằng cách tiếp tục huấn luyện nó trên một tập dữ liệu có gắn nhãn phù hợp với nhiệm vụ cụ thể. Khác với zero-shot – vốn chỉ dựa vào prompt và tri thức có sẵn – fine-tuning cho phép mô hình học sâu hơn về các đặc điểm ngôn ngữ, cấu trúc và sắc thái biểu đạt đặc thù trong ngữ liệu mục tiêu, từ đó nâng cao độ chính xác và tính phù hợp với môi trường ứng dụng thực tế.



**Hình 2.4:** Hình ảnh minh họa chiến lược Học không mẫu

Trong bài toán kiểm chứng tin giả, tinh chỉnh mô hình trên tập dữ liệu tiếng Việt hoặc các nội dung mạng xã hội địa phương là bước quan trọng để mô hình có thể nhận diện chính xác các dấu hiệu sai lệch, ngôn ngữ lừa đảo, cũng như các mối liên hệ tiềm ẩn giữa tuyên bố và nguồn phát tán. Quy trình tinh chỉnh thường bao gồm các bước như chuẩn bị dữ liệu có gắn nhãn, thiết kế cấu trúc mô hình phù hợp, và huấn luyện lại

mô hình với tốc độ học thấp để tránh làm mất tri thức đã học trước đó (Hình 2.4). Tùy vào khối lượng dữ liệu và độ phức tạp của nhiệm vụ, quá trình tinh chỉnh có thể được thực hiện toàn phần (full fine-tuning) hoặc một phần (parameter-efficient tuning).

Ưu điểm phương pháp này là khả năng tùy biến mô hình cho từng ngữ cảnh cụ thể, đặc biệt hiệu quả với các ngôn ngữ ít tài nguyên như tiếng Việt – nơi các mô hình ngôn ngữ lớn đa ngữ có thể chưa được huấn luyện đầy đủ. Bằng cách học từ dữ liệu nội địa, mô hình có thể hiểu rõ hơn về đặc điểm ngôn ngữ, từ ngữ địa phương, cấu trúc câu không chuẩn hay cách diễn đạt thường thấy trong tin giả. Kết quả từ nhiều nghiên cứu cũng chỉ ra rằng, mô hình sau khi được tinh chỉnh trên tập dữ liệu chuyên biệt thường cho độ chính xác và độ tin cậy cao hơn đáng kể so với khi sử dụng ở chế độ zero-shot.

Việc triển khai cả hai phương pháp trong cùng một tập nhiệm vụ cho phép đánh giá mức độ cải thiện mà fine-tuning mang lại so với khả năng zero-shot nguyên bản. Tuy nhiên, fine-tuning cũng có một số yêu cầu nhất định về mặt kỹ thuật như cần GPU để huấn luyện hay chuẩn bị dữ liệu có nhãn chất lượng. Dù vậy, trong môi trường ứng dụng đòi hỏi độ chính xác cao như kiểm chứng tin tức, việc đầu tư vào tinh chỉnh mô hình là hoàn toàn cần thiết và có giá trị lâu dài, đặc biệt khi hệ thống phải xử lý dữ liệu nhạy cảm hoặc hoạt động ở môi trường đa ngôn ngữ.

## **2.5. Giới thiệu kỹ thuật truy xuất thông tin (RAG)**

### **2.5.1. Khái niệm**

Retrieval-Augmented Generation (RAG) là một phương pháp kết hợp giữa khả năng sinh văn bản của các mô hình ngôn ngữ lớn (LLMs) và kỹ thuật truy xuất thông tin từ các kho dữ liệu bên ngoài nhằm nâng cao độ chính xác và tính minh bạch của hệ thống. Thay vì chỉ dựa vào kiến thức cố định được học trong giai đoạn tiền huấn luyện, RAG cho phép mô hình truy vấn dữ liệu thực tế từ các kho thông tin, sau đó sử dụng các kết quả này làm ngữ cảnh để tạo ra phản hồi phù hợp.

Mô hình RAG thường bao gồm hai thành phần chính: bộ truy vấn (retriever) và bộ sinh (generator). Bộ truy vấn sử dụng các thuật toán tìm kiếm hiệu quả để lựa chọn ra những đoạn văn bản hoặc tài liệu phù hợp nhất với truy vấn đầu vào từ kho dữ liệu lớn, có thể là các bộ sưu tập tài liệu chuyên biệt hoặc dữ liệu mở (open-domain). Bộ sinh sau đó dùng các đoạn dữ liệu được truy xuất làm ngữ cảnh để tạo ra phản hồi, giúp cải thiện tính chính xác, cập nhật và khả năng giải thích của kết quả.

### ***2.5.2. Khả năng ứng dụng***

RAG ngày càng được ứng dụng rộng rãi trong các hệ thống hỏi đáp tự động, trợ lý ảo, cũng như trong các hệ thống kiểm chứng tin tức tự động với các tác vụ bổ sung thông tin liên quan hoặc yêu cầu cập nhật kiến thức liên tục và xử lý dữ liệu mở không giới hạn.

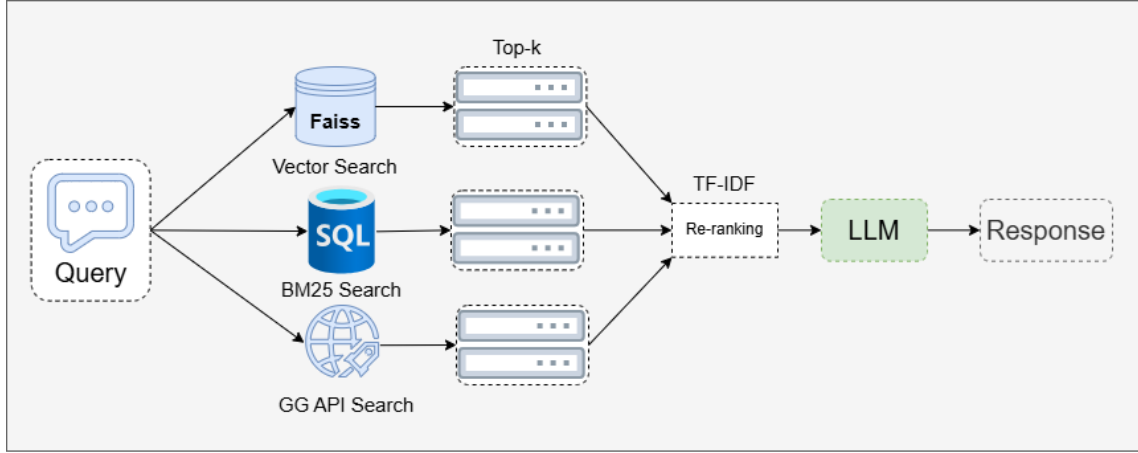
Trong bài toán phát hiện tin giả, bằng cách kết hợp khả năng truy xuất thông tin từ kho dữ liệu bên ngoài với khả năng sinh văn bản của mô hình ngôn ngữ lớn, RAG đóng vai trò quan trọng khi giúp mô hình giảm thiểu hiện tượng “hallucination” – việc tạo ra thông tin sai lệch không có căn cứ thường gặp ở các mô hình sinh tự do. Chính vì vậy, RAG được xem là một kỹ thuật cốt lõi và được áp dụng rộng rãi trong các hệ thống kiểm chứng tin tức hiện đại nhằm đảm bảo tính minh bạch và tin cậy trong việc đánh giá thông tin [10].

### ***2.5.3. Kết hợp các kỹ thuật truy vấn***

Truy vấn thông tin - Retrieval-Augmented Generation (RAG), đóng vai trò thiết yếu để cung cấp ngữ cảnh chính xác cho mô hình sinh văn bản. Trong hệ thống kiểm chứng thông tin đề xuất, khả năng truy xuất chính xác và toàn diện đóng vai trò cốt lõi nhằm cung cấp ngữ cảnh cho mô hình ngôn ngữ lớn (LLM). Để đáp ứng yêu cầu đó, nghiên cứu triển khai chiến lược truy vấn kết hợp – **Hybrid Search**, kết hợp giữa hai hướng tiếp cận:

- **Tìm kiếm từ vựng (lexical search):** Dựa trên kỹ thuật BM25.
- **Tìm kiếm ngữ nghĩa (semantic search):** Dựa trên biểu diễn vector của văn bản trong không gian ngữ nghĩa.

Hình 2.5 minh họa luồng xử lý tổng thể của kiến trúc Hybrid Search.



**Hình 2.5:** Kiến trúc Hybrid Search trong RAG

#### 2.5.3.1. BM25 – Best Matching 25

Là một thuật toán xếp hạng thuộc họ mô hình xác suất, được thiết kế để đánh giá mức độ liên quan giữa truy vấn và tài liệu trong một kho dữ liệu văn bản. Thuật toán dựa trên tần suất xuất hiện từ khóa (term frequency - TF), độ dài tài liệu, và tần suất nghịch đảo tài liệu (inverse document frequency - IDF) để đưa ra điểm số cho từng tài liệu. Công thức tính điểm BM25 như sau:

$$\text{BM25}(q, d) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \quad (2.1)$$

- $f(q_i, d)$ : số lần từ  $q_i$  xuất hiện trong tài liệu  $d$ .
- $|d|$ : độ dài của tài liệu.

- avgdl: độ dài trung bình của các tài liệu trong tập.
- $k_1$  và  $b$ : tham số điều chỉnh (thường  $k_1 = 1.2$  đến  $2.0$ ,  $b = 0.75$ ).

BM25 được sử dụng như bộ lọc sơ cấp, đánh giá độ liên quan giữa truy vấn đầu vào và các tài liệu trong kho dữ liệu dựa trên tần suất xuất hiện và độ dài văn bản. Đây là bước truy vấn chính giúp rút gọn tập dữ liệu cần xử lý.

#### 2.5.3.2. *Embedding Search – Truy vấn ngữ nghĩa*

Trái ngược với phương pháp BM25 vốn dựa vào tần suất từ khóa và độ khớp từ vựng, **Embedding Search** sử dụng biểu diễn ngữ nghĩa của văn bản dưới dạng vector trong không gian nhiều chiều để đánh giá mức độ tương đồng giữa truy vấn và tài liệu. Đây là hướng tiếp cận hiện đại nhằm khắc phục các hạn chế của truy vấn từ vựng, chẳng hạn như hiện tượng “trùng từ nhưng khác nghĩa” hoặc “khác từ nhưng cùng nghĩa”.

Trong hệ thống đề xuất, các mô hình embedding nhẹ như MiniLM được sử dụng để mã hóa cả truy vấn đầu vào và các văn bản trong kho dữ liệu thành các vector ngữ nghĩa. Các vector này sau đó được so sánh bằng độ đo **cosine similarity** để xác định mức độ tương đồng nội dung giữa truy vấn và tài liệu. Công thức tính như sau:

$$\text{cosine\_similarity}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \cdot \|\vec{d}\|} \quad (2.2)$$

- $\vec{q}$ : vector biểu diễn truy vấn đầu vào.
- $\vec{d}$ : vector biểu diễn một tài liệu trong kho dữ liệu.
- $\vec{q} \cdot \vec{d}$ : tích vô hướng giữa hai vector.
- $\|\vec{q}\|$  và  $\|\vec{d}\|$ : độ dài (chuẩn Euclid) của các vector tương ứng.



Dựa trên chỉ số tương đồng này, hệ thống lựa chọn các tài liệu có độ gần ngữ nghĩa cao nhất để đưa vào danh sách kết quả. Kỹ thuật này đặc biệt hữu ích trong các trường hợp truy vấn và tài liệu sử dụng từ ngữ khác nhau nhưng cùng biểu đạt một nội dung. Ví dụ:

- **Truy vấn:** “Sập sàn đầu tư tài chính đa cấp”
- **Tài liệu liên quan:** “Nhiều người mất trắng vì mô hình huy động vốn kiểu Ponzi trá hình”

Mặc dù không có từ khóa trùng khớp rõ ràng, nội dung hai câu trên mang cùng một thông điệp. Embedding Search có khả năng phát hiện mối liên hệ này thông qua biểu diễn ngữ nghĩa, từ đó gia tăng độ chính xác và độ bao phủ của hệ thống.

Cuối cùng, các kết quả từ Embedding Search sẽ được kết hợp với các kết quả từ BM25 và đưa vào bước tái xếp hạng bằng TF-IDF để lựa chọn ra tập tài liệu phù hợp nhất cung cấp cho mô hình sinh phản hồi (LLM).

#### 2.5.3.3. *TF-IDF – Term Frequency-Inverse Document Frequency*

Là một chỉ số thống kê cổ điển trong lĩnh vực tìm kiếm thông tin, kết hợp tần suất xuất hiện của từ trong tài liệu (TF) và độ hiếm của từ trong tập văn bản (IDF). Chỉ số TF-IDF của từ  $t$  trong tài liệu  $d$  được tính như sau:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \log \left( \frac{N}{\text{DF}(t)} \right) \quad (2.3)$$

- $\text{TF}(t, d)$ : số lần từ  $t$  xuất hiện trong tài liệu  $d$ .
- $N$ : tổng số tài liệu trong tập dữ liệu.
- $\text{DF}(t)$ : số tài liệu chứa từ  $t$ .

Sau khi truy xuất được danh sách tài liệu liên quan, TF-IDF tiếp tục được dùng để đánh giá và tái xếp hạng (re-rank) các tài liệu đã được truy xuất bởi hai phương pháp trên, giúp chọn ra những tài liệu phù hợp nhất để cung cấp cho mô hình sinh văn bản.

#### 2.5.3.4. Hybrid Search - Quy trình tìm kiếm kết hợp

Quy trình hoạt động của Hybrid Search gồm ba giai đoạn:

**(1) Truy vấn song song:** Truy vấn đầu vào được gửi đồng thời đến hai chỉ mục:

- **BM25 Index (SQL):** Tìm kiếm từ khóa gần truy vấn nhất và trả về  $S_{BM25s}$  gồm top- $k$  tài liệu.
- **Vector Index (FAISS):** Tìm kiếm ngữ nghĩa bằng cách ánh xạ truy vấn và văn bản sang vector nhúng (embedding), sau đó tính độ tương đồng bằng *cosine similarity*. Kết quả trả về là  $S_{vec}$ .

**(2) Tính điểm kết hợp:** Với mỗi tài liệu  $d$ , hệ thống tính điểm kết hợp giữa hai phương pháp:

$$\text{Score}_{\text{hybrid}}(q, d) = \alpha \cdot \text{score}_{\text{vec}}(q, d) + (1 - \alpha) \cdot \text{score}_{\text{BM25}}(q, d) \quad (2.4)$$

Trong đó:

- $\text{score}_{\text{BM25}}(q, d)$ : điểm BM25 chuẩn hóa,
- $\text{score}_{\text{vec}}(q, d)$ : cosine similarity của vector embedding,
- $\alpha \in [0, 1]$ : hệ số điều chỉnh giữa hai phương pháp

(nghiên cứu sử dụng  $\alpha = 0.5$  để cân bằng giữa tìm kiếm từ khóa và vector).

**(3) Tái xếp hạng với TF-IDF:** Sau khi chọn ra các tài liệu có điểm tổng hợp cao nhất, hệ thống tiến hành bước tái xếp hạng bằng chỉ số TF-IDF nhằm tăng độ đặc trưng và loại bỏ nhiễu.

Bước này giúp sàng lọc những tài liệu vừa có tính liên kết cao, vừa phù hợp ngữ cảnh để cung cấp cho mô hình sinh văn bản.

#### Thuật toán Truy vấn Kết hợp (Hybrid Search)

**Input:** Truy vấn người dùng  $q$ , chỉ mục BM25  $I_{BM25}$ , chỉ mục Vector  $I_{vec}$ , hệ số kết hợp  $\alpha \in [0, 1]$

**Output:** Danh sách tài liệu được xếp hạng cuối cùng  $D_{ranked}$

1. Mã hoá truy vấn:  $\vec{q} \leftarrow \text{Embed}(q)$
2. Truy xuất từ BM25:  $S_{BM25} \leftarrow \text{Retrieve}(q, I_{BM25})$
3. Truy xuất từ vector index:  $S_{vec} \leftarrow \text{Retrieve}(\vec{q}, I_{vec})$
4. Khởi tạo dictionary  $D_{score} \leftarrow \emptyset$
5. **For** mỗi tài liệu  $d \in S_{BM25} \cup S_{vec}$ :
  - $s_{BM25} \leftarrow \text{BM25Score}(d)$
  - $s_{vec} \leftarrow \text{VectorScore}(d)$
  - $D_{score}[d] \leftarrow \alpha \cdot s_{vec} + (1 - \alpha) \cdot s_{BM25}$
6. Sắp xếp  $D_{score}$  theo thứ tự giảm dần điểm
7. **Return** danh sách tài liệu đã được tái xếp hạng  $D_{ranked}$

#### 2.5.4. Tìm kiếm dữ liệu mở với Google Search API

Bên cạnh các kho dữ liệu tĩnh, hệ thống còn tích hợp Google Search API như một công cụ hỗ trợ mở rộng phạm vi truy vấn ra Internet. API này cho phép thu thập dữ liệu từ các nguồn tin cậy, báo chí chính thống và trang web uy tín. Các kết quả sẽ được thu thập, xử lý và đưa vào hệ thống truy xuất nội bộ. Mặc dù không phải là thành phần

trực tiếp trong kiến trúc RAG truyền thống, việc sử dụng Google Search API đóng vai trò như một công cụ thu thập dữ liệu hỗ trợ truy xuất, góp phần nâng cao độ bao phủ và độ mới của kho dữ liệu.

## **2.6. Các công cụ hỗ trợ xác minh thông tin**

Trong bối cảnh các chiến dịch lừa đảo trực tuyến ngày càng tinh vi, việc xác thực tính hợp lệ của các đầu mối thông tin như đường dẫn (URL), địa chỉ thư điện tử (email), và số điện thoại là một bước thiết yếu trong quy trình phát hiện tin tức giả mạo. Để nâng cao độ chính xác và tự động hóa quá trình này, hệ thống nghiên cứu đã tích hợp các công cụ phân tích và kiểm chứng dựa trên API đáng tin cậy từ các nhà cung cấp uy tín. Các công cụ này hỗ trợ phân tích hành vi, phát hiện tín hiệu rủi ro, và đưa ra kết luận sơ bộ trước khi chuyển sang bước phân tích sâu bằng mô hình ngôn ngữ lớn (LLM).

### **2.6.1. VirusTotal API**

VirusTotal là nền tảng được sở hữu bởi Google, chuyên cung cấp dịch vụ phân tích bảo mật cho tệp tin và liên kết URL bằng cách tổng hợp kết quả từ hàng chục công cụ chống mã độc và trình quét bảo mật.

Trong hệ thống nghiên cứu, URL do người dùng cung cấp sẽ được mã hóa và gửi truy vấn tới VirusTotal thông qua API. Kết quả trả về cung cấp thông tin về mức độ “malicious”, “suspicious” hay “harmless” của từng liên kết. Từ đó, hệ thống sẽ suy luận mức độ rủi ro tổng quát để cảnh báo người dùng hoặc điều phối các bước phân tích tiếp theo.

### **2.6.2. Abstract Email API**

Trong các chiến dịch phát tán thông tin sai lệch, email thường là kênh liên lạc phổ biến và dễ bị giả mạo. Abstract Email API được sử dụng để xác nhận sự tồn tại

của máy chủ thư và phân loại email theo các đặc tính như miễn phí (free), tạm thời (disposable), đại diện tổ chức (role-based) hay tên miền cá nhân. Dựa trên tổ hợp các chỉ số này, hệ thống có thể phân loại email là hợp lệ, tiềm ẩn rủi ro, hay không khả dụng, từ đó nâng cao hiệu quả sàng lọc đầu vào.

### **2.6.3. Abstract Phone API**

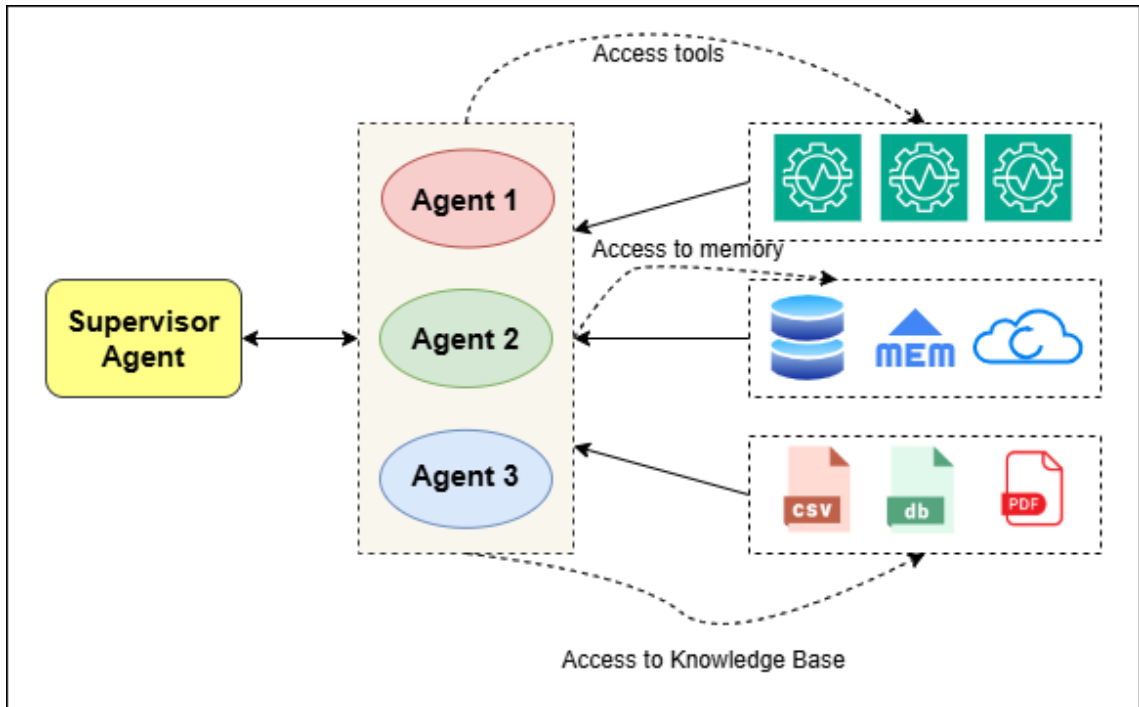
Trong bối cảnh gia tăng các cuộc gọi lừa đảo xuyên quốc gia, giả mạo từ cơ quan chức năng hoặc tổ chức quốc tế nhằm phát tán thông tin sai lệch và đánh cắp dữ liệu cá nhân. Abstract Phone API hỗ trợ xác minh số điện thoại bằng cách chuẩn hóa định dạng quốc tế và kiểm tra các thuộc tính liên quan như quốc gia, nhà mạng, loại số, đồng thời cảnh báo nếu số đó thuộc các quốc gia thường xuất hiện trong các chiến dịch lừa đảo. Nhờ khả năng nhận diện các số từ quốc gia có rủi ro cao hoặc từ ngoài lãnh thổ Việt Nam, API này giúp hệ thống nhanh chóng phân loại các đầu mối liên lạc khả nghi, góp phần nâng cao hiệu quả trong việc kiểm soát và xác minh thông tin.

## **2.7. Multi-Agent AI trong hệ thống xác minh thông tin**

### **2.7.1. Khái niệm Multi AI Agent**

Hệ thống Multi-Agent AI (Multi AI Agent System) là một tập hợp nhiều tác tử trí tuệ nhân tạo (AI Agent) hoạt động đồng thời hoặc phối hợp nhằm giải quyết các tác vụ phức tạp hơn mà một tác tử đơn lẻ không thể đảm nhiệm hiệu quả. Trong đó, AI Agent là một phần mềm ứng dụng các mô hình chuyên sâu phù hợp như LLM, được thiết kế để thực hiện một nhiệm vụ độc lập như truy vấn dữ liệu, phân tích ngôn ngữ, hoặc tự động đưa ra hành động nhằm đạt được mục tiêu cụ thể.

Mô hình này kế thừa ý tưởng từ lĩnh vực hệ thống đa tác tử trong trí tuệ nhân tạo, nơi các agent có khả năng giao tiếp, phối hợp, chia sẻ mục tiêu và xử lý phân tán. Theo Wooldridge (2009), hệ thống đa tác tử có ưu thế trong việc phân chia nhiệm vụ, mở rộng quy mô linh hoạt và tăng độ tin cậy nhờ tính tự chủ của từng agent [20].



**Hình 2.6:** Hình ảnh minh họa multi-agent system

Việc triển khai Multi AI Agent trong nghiên cứu này cũng hỗ trợ giảm tải tính toán cho từng thành phần, tăng tính linh hoạt trong cập nhật và bảo trì hệ thống, đồng thời cho phép mở rộng quy mô khi cần thiết để xử lý lượng lớn thông tin đa dạng và liên tục thay đổi trong môi trường mạng hiện nay.

### **2.7.2. Giới thiệu về CrewAI Framework**

CrewAI là một framework hiện đại hỗ trợ xây dựng và điều phối các hệ thống Multi-Agent AI một cách hiệu quả, an toàn và có khả năng mở rộng linh hoạt [4]. Nền tảng này cho phép người dùng dễ dàng thiết kế và vận hành hệ thống gồm nhiều tác tử AI đảm nhiệm các vai trò khác nhau.

Được thiết kế dựa trên kiến trúc microservices, cho phép mỗi tác tử (agent) hoạt động như một cá nhân độc lập với các giao diện riêng biệt. Điều này không chỉ giúp tăng khả năng tái sử dụng và dễ dàng cập nhật hoặc thay thế từng agent mà không ảnh hưởng đến toàn bộ hệ thống – một lợi thế quan trọng cho các ứng dụng yêu cầu cập

nhật liên tục – mà còn đảm bảo tính ổn định và khả năng phối hợp hiệu quả giữa các tác tử thông qua cơ chế giao tiếp và quản lý nội bộ, từ đó giảm thiểu lỗi và tối ưu hóa hiệu suất tổng thể.

```
Crew Execution Started

Crew Execution Started
Name: crew
ID: 56e565bd-c5e0-4f0a-a47c-2a104154047f

Crew: crew
└─ Task: 1ea06e2f-6b96-40a5-a788-8bf3cee64584
   Status: Executing Task...

Crew: crew
└─ Task: 1ea06e2f-6b96-40a5-a788-8bf3cee64584
   Status: Executing Task...
   └─ Agent: Final Verifier
      Status: In Progress

# Agent: Final Verifier
## Task: Phân tích dữ liệu bên dưới và đưa ra đánh giá rõ ràng theo cấu trúc:
1. Kết luận (✅ An toàn / ⚠️ Đáng ngờ / ❌ Lừa đảo)
2. Giải thích lý do
3. Gợi ý hành động
```

*Hình 2.7: Hình ảnh minh họa quá trình CrewAI hoạt động*

CrewAI hoạt động bằng cách khởi tạo một "Crew" để quản lý các tác vụ. Trên terminal, có thể thấy nhiệm vụ đang được giao cho một tác tử chuyên biệt, "Agent: Final Verifier", với trạng thái "In Progress". Vai trò của tác tử này được định nghĩa rõ ràng và Agent nào đang được giao nhiệm vụ. Quá trình này giúp người dùng theo dõi từng bước hoạt động của Crew và các Agent trong quá trình thực thi các tác vụ phức tạp.

## CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN

Chương này trình bày phương pháp nghiên cứu được áp dụng để xây dựng hệ thống xác minh thông tin, bao gồm lựa chọn và tinh chỉnh mô hình LLM, thiết kế kiến trúc hệ thống Multi-Agent, xây dựng cơ sở dữ liệu và giới thiệu các thành phần chính của hệ thống.

### 3.1. Các mô hình sử dụng trong nghiên cứu

Trong những năm gần đây, các mô hình ngôn ngữ lớn (LLMs) như LLaMA, Qwen, Mistral, hay Gemma đã thu hút sự quan tâm lớn từ cộng đồng nghiên cứu nhờ khả năng hiểu và phản hồi chính xác theo ngữ cảnh người dùng. Không chỉ sở hữu hiệu năng mạnh mẽ trong nhiều tác vụ xử lý ngôn ngữ tự nhiên, các mô hình này còn được phát hành dưới dạng mã nguồn mở và miễn phí sử dụng, cho phép các nhóm nghiên cứu dễ dàng tiếp cận, khai thác và tinh chỉnh (fine-tune) mà không cần chi trả chi phí như khi sử dụng các API thương mại.

Một đặc điểm nổi bật của các LLM hiện đại là khả năng zero-shot và few-shot learning, nghĩa là chúng có thể thực hiện tốt các nhiệm vụ mới mà không cần huấn luyện lại hoặc chỉ cần một lượng dữ liệu huấn luyện rất nhỏ [22]. Ví dụ, nghiên cứu của Bucher và cộng sự (2024) cho thấy các mô hình như ChatGPT có thể đưa ra dự đoán đáng tin cậy về tin giả chỉ dựa trên lời dẫn đầu vào (prompt), mà không cần tinh chỉnh thêm [19].

#### 3.1.1. Thử nghiệm các mô hình ban đầu

Trong nghiên cứu này, chúng tôi tiến hành đánh giá ban đầu các mô hình mở gồm **Mistral-7B**, **Llama3.2-3B**, **Qwen2.5-3B**, **Meta-Llama3.1-8B** và **Gemma2-9B** theo



phương pháp zero-shot inference – tức là mô hình sẽ được yêu cầu phân loại tin tức mà không trải qua quá trình huấn luyện lại. Những mô hình này được chọn dựa trên khả năng xử lý ngôn ngữ tự nhiên cao và được nhắc đến trong các công trình mới nhất, nhằm kiểm tra khả năng tổng quát hóa (generalization) vốn có của các LLM, nhờ vào lượng kiến thức đã được tích lũy trong quá trình tiền huấn luyện.

Các mô hình sẽ được thử nghiệm trực tiếp trên tập dữ liệu tiếng Việt không qua huấn luyện, nhằm xác định những mô hình có tiềm năng cao để đưa vào giai đoạn fine-tuning nhằm cải thiện độ chính xác trong phân loại tin thật và tin giả. Nhóm sẽ tiến hành thu thập và phân tích số lượng lỗi phát hiện (Error\_Detect) – đại diện cho các trường hợp mô hình không thể phân loại hoặc trả lời sai hoàn toàn. Kết quả từ giai đoạn này là cơ sở để nhận diện những mô hình có hiệu năng tiềm năng, từ đó lựa chọn cho bước fine-tuning tiếp theo.

### ***3.1.2. Các mô hình được lựa chọn để Fine-tune***

Dựa trên kết quả từ đánh giá zero-shot, chúng tôi quyết định loại bỏ Mistral-7B khỏi danh sách fine-tune khi hiệu suất đánh giá thấp nhất với Accuracy chỉ 45.5% cùng với tỉ lệ phản hồi sai vượt quá 20% (cụ thể ở bảng 4.3). Bốn mô hình còn lại bao gồm Llama3.2-3B, Qwen2.5-3B, Meta-Llama3.1-8B và Gemma2-9B được lựa chọn để đưa vào giai đoạn fine-tuning nhờ vào độ ổn định và chất lượng phản hồi cao hơn.

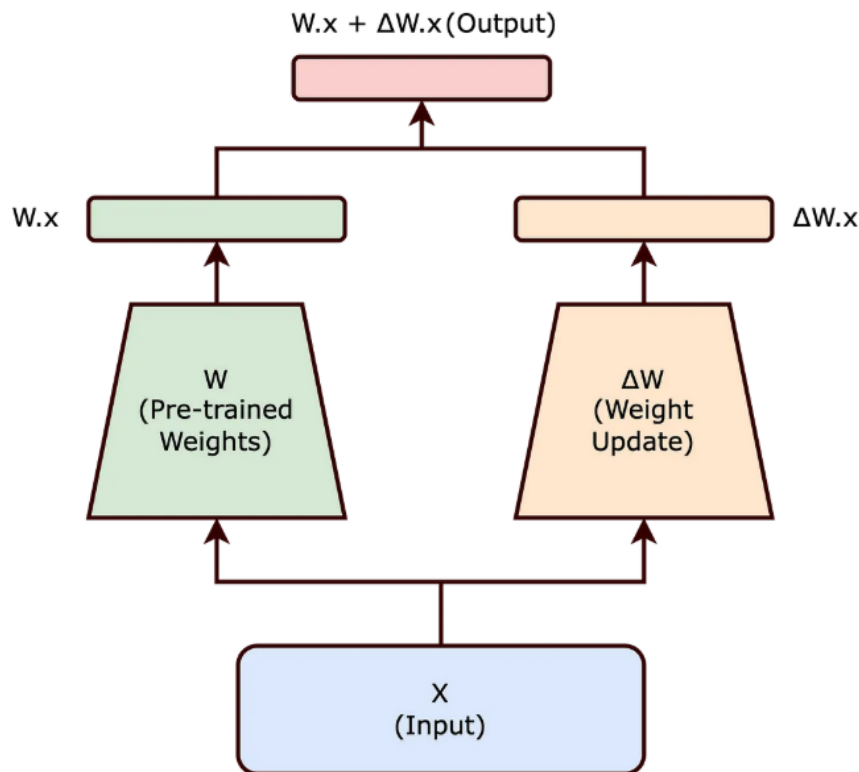
Các mô hình được tinh chỉnh độc lập trên tập dữ liệu tiếng Việt đã được gán nhãn, với mục tiêu tối ưu khả năng phân loại giữa hai nhãn "real" và "fake". Kỹ thuật LoRA (Low-Rank Adaptation) được áp dụng nhằm tăng hiệu quả tinh chỉnh mà vẫn giảm thiểu tài nguyên tính toán. Bước fine-tuning này đóng vai trò quan trọng trong việc thích nghi mô hình với ngữ cảnh và đặc điểm ngôn ngữ cụ thể của tiếng Việt, giúp cải thiện đáng kể độ chính xác và độ tin cậy trong nhận diện tin giả.

## 3.2. Phương pháp Fine-tune (LoRA, Unsloth)

Để tinh chỉnh các mô hình ngôn ngữ lớn (LLM) một cách hiệu quả và tiết kiệm tài nguyên, chúng tôi áp dụng phương pháp Low-Rank Adaptation (LoRA) – một kỹ thuật thuộc nhóm huấn luyện tham số hiệu quả (PEFT – Parameter-Efficient Fine-Tuning) kết hợp với Unsloth, một framework fine-tune chuyên biệt.

### 3.2.1. LoRA (Low-Rank Adaptation)

LoRA là một phương pháp fine-tuning hiệu quả thuộc nhóm huấn luyện tham số tiết kiệm (PEFT – Parameter-Efficient Fine-Tuning). LoRA giữ nguyên ma trận trọng số gốc  $\mathbf{W} \in \mathbb{R}^{d \times k}$  và chỉ thêm hai ma trận phụ có hạng thấp là  $\mathbf{A} \in \mathbb{R}^{d \times r}$  và  $\mathbf{B} \in \mathbb{R}^{r \times k}$ .



**Hình 3.1:** Cơ chế cập nhật tham số trong LoRA

Khi đó, ma trận trọng số mới được điều chỉnh như sau:

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + \mathbf{AB} \quad (3.1)$$

Trong đó:

- **W**: Ma trận trọng số ban đầu (giữ nguyên trong quá trình huấn luyện).
- **A, B**: Các ma trận low-rank được thêm vào và huấn luyện.
- $r \ll \min(d, k)$ : Hạng thấp của ma trận giúp giảm chi phí tính toán và bộ nhớ.

Thay vì cập nhật toàn bộ trọng số của mô hình, LoRA đóng băng (freeze) tất cả các trọng số gốc (original weights), tức là giữ nguyên không cho thay đổi, và chỉ học thêm một tập nhỏ các ma trận có dạng phân hạng thấp (low-rank matrices). Kỹ thuật này giúp giảm đáng kể số lượng tham số cần cập nhật, từ đó giảm chi phí bộ nhớ và thời gian huấn luyện, nhưng vẫn giữ được hoặc cải thiện hiệu quả mô hình [6].

Tuy nhiên, quá trình fine-tuning yêu cầu GPU có tối thiểu 12GB VRAM (như T4 hoặc RTX 3060) để xử lý ổn định các mô hình 3B–7B; với mô hình lớn hơn như Gemma2-9B, nên sử dụng GPU từ 16–24GB VRAM để đạt hiệu năng tối ưu.

### 3.2.2. *Unsloth – Framework tối ưu hoá Fine-tuning*

Unsloth là một thư viện huấn luyện chuyên dụng được thiết kế để triển khai các kỹ thuật PEFT như LoRA một cách tối ưu [19]. Ưu điểm của Unsloth bao gồm tự động đóng băng toàn bộ trọng số gốc và chỉ thêm adapter vào các lớp tuyến tính cần thiết. Hỗ trợ chuẩn dữ liệu đầu vào định dạng JSONL, thuận tiện cho quá trình huấn luyện. Dễ dàng tích hợp sử dụng các mô hình từ HuggingFace và ổn định với tài nguyên phần cứng thấp.

### 3.2.3. Quá trình huấn luyện

Trong giai đoạn fine-tuning, các mô hình được huấn luyện độc lập trên tập dữ liệu tin tức đã được gán nhãn “fake” hoặc “real”. Để đảm bảo tính nhất quán, một lời nhắc (prompt) thống nhất được thiết kế cho tất cả mô hình, theo hướng biến bài toán thành nhiệm vụ phân loại nhị phân có giám sát.

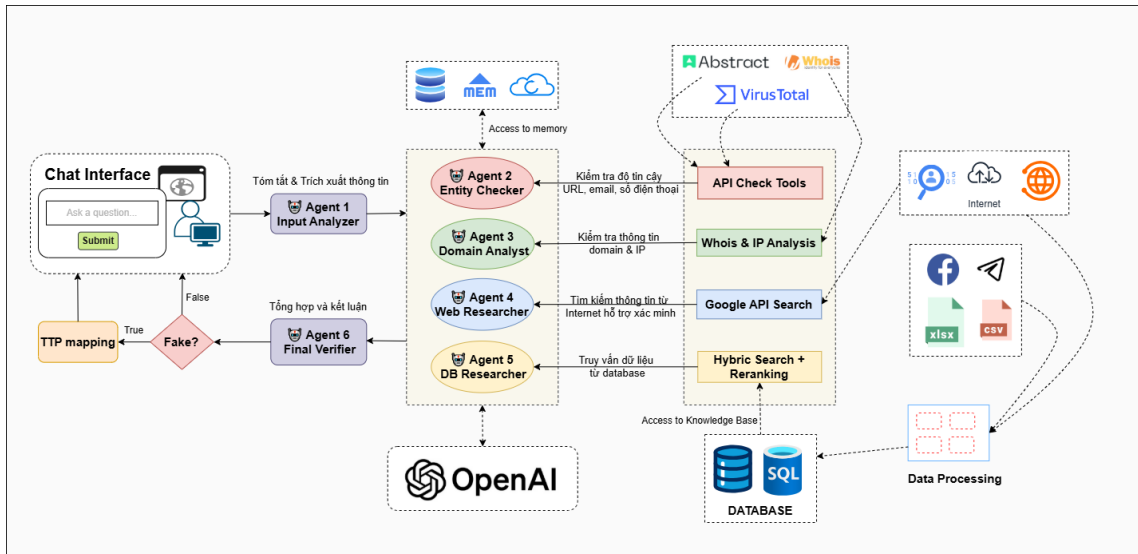
Quá trình huấn luyện được tiến hành theo chiến lược đa epoch, sử dụng bộ tối ưu AdamW và hàm mất mát tiêu chuẩn (Cross-entropy). Việc huấn luyện qua nhiều epoch cho phép mô hình tự động cải thiện độ phù hợp với dữ liệu, “chiến lược đa epoch cho phép các mô hình cải thiện lặp đi lặp lại, dẫn đến dự đoán chính xác hơn” [1]. Kết quả cuối cùng là mỗi mô hình thu được một tập tham số LoRA chuyên biệt, tối ưu hóa cho nhiệm vụ phát hiện tin giả tiếng Việt, và thể hiện sự cải thiện rõ rệt so với trạng thái ban đầu (zero-shot).

## 3.3. Tổng quan kiến trúc hệ thống Multi-Agent đề xuất

Sau khi xác định được bài toán và các yêu cầu chức năng, kiến trúc hệ thống được thiết kế theo hướng đa tác tử (multi-agent), nhằm tận dụng tối đa khả năng chuyên môn hóa của từng thành phần xử lý. Việc chia nhỏ quy trình xác minh thông tin thành các tác vụ riêng biệt, do từng tác tử đảm nhiệm, không chỉ giúp hệ thống xử lý hiệu quả hơn mà còn nâng cao khả năng giám sát, mở rộng và bảo trì. Thay vì triển khai một mô hình đơn lẻ cố gắng đảm đương toàn bộ quy trình từ nhận diện, đối chiếu, truy xuất đến đánh giá, kiến trúc phân tán theo vai trò này cho phép từng tác tử chuyên sâu vào một khía cạnh cụ thể, từ đó tạo ra một hệ sinh thái kiểm chứng thông tin rõ ràng và có tổ chức.

Sơ đồ tổng quan hệ thống được trình bày ở Hình 3.2, gồm sáu agent chính, với mỗi agent phụ trách một nhiệm vụ cụ thể trong quy trình xác minh.

Cấu trúc hệ thống được tổ chức theo dạng pipeline tuần tự, trong đó mỗi agent đảm nhiệm một vai trò chức năng cụ thể trong chuỗi xử lý, đồng thời có thể hoạt



**Hình 3.2:** Sơ đồ tổng quan hệ thống multi-agent

động song song ở những bước không phụ thuộc nhau. Các tác tử đầu vào bao gồm Input Analyzer, Entity Checker và Domain Analyst sẽ đóng vai trò xử lý tiền kiểm tra, giúp chuẩn hóa và làm rõ dữ liệu đầu vào. Các tác tử truy xuất và phân tích như Web Researcher và Database Researcher chịu trách nhiệm thu thập và đánh giá thông tin từ nhiều nguồn khác nhau – bao gồm cả dữ liệu mở trên Internet và dữ liệu có cấu trúc trong nội bộ hệ thống. Cuối cùng, kết quả được tổng hợp bởi Final Verifier để đưa ra đánh giá xác thực, đồng thời nếu cần, tiến hành ánh xạ sang các chiến lược lừa đảo phổ biến theo mô hình TTP.

Để đảm bảo hệ thống hoạt động hiệu quả và cập nhật, các tác tử được tích hợp với nhiều công nghệ hiện đại. Cụ thể, mô hình ngôn ngữ lớn GPT-4o được sử dụng làm nền tảng cho tất cả các agent, cung cấp khả năng hiểu ngữ nghĩa sâu sắc và phản hồi ngôn ngữ tự nhiên. Bên cạnh đó, hệ thống còn tích hợp các công cụ chuyên biệt như VirusTotal API, Abstract API để kiểm tra mức độ uy tín của các thực thể như địa chỉ email, số điện thoại và đường dẫn URL. Với các tác vụ tìm kiếm mở, hệ thống sử dụng Google Search API để truy xuất dữ liệu mới nhất từ Internet, đồng thời kết hợp kỹ thuật RAG để truy vấn dữ liệu nội bộ và tái xếp hàng nhằm cung cấp thông tin bổ sung cho mô hình xử lý.

Đáng chú ý, kiến trúc tác tử cũng tạo điều kiện thuận lợi cho việc kiểm soát từng bước xử lý và lý giải kết quả đầu ra cho người dùng. Mỗi tác tử đều có khả năng cung cấp phản hồi trung gian rõ ràng, giúp hệ thống không chỉ dừng lại ở kết luận đúng/sai mà còn giải thích được vì sao một thông tin được đánh giá là đáng tin cậy hay không. Điều này là vô cùng quan trọng trong bối cảnh người dùng ngày càng cần minh bạch trong các hệ thống trí tuệ nhân tạo.

Tóm lại, kiến trúc hệ thống được đề xuất không chỉ phù hợp với yêu cầu chức năng của bài toán xác minh thông tin mà còn thể hiện tính linh hoạt, mở rộng và khả năng tích hợp mạnh mẽ với các công nghệ tiên tiến hiện nay. Các phần tiếp theo của chương này sẽ trình bày chi tiết về từng tác tử, vai trò, công nghệ hỗ trợ cũng như quy trình phối hợp giữa các thành phần trong toàn hệ thống.

### **3.4. Các tác vụ của hệ thống Multi-Agent**

Hệ thống AI đa tác tử được thiết kế để hoạt động theo chuỗi xử lý nhiệm vụ, trong đó mỗi tác tử (agent) đảm nhận một vai trò chức năng rõ ràng và độc lập. Các tác vụ này phản ánh toàn bộ quy trình xác minh thông tin, từ giai đoạn tiếp nhận đầu vào đến khi đưa ra đánh giá và phản hồi cho người dùng. Việc phân tách nhiệm vụ thành từng tác tử chuyên biệt giúp đảm bảo hiệu quả xử lý, dễ dàng mở rộng, và tăng tính minh bạch trong quá trình suy luận của hệ thống.

#### ***3.4.1. Tiếp nhận và phân tích đầu vào (Input Analyzer)***

Tác tử đầu tiên của hệ thống (Input Analyzer) có nhiệm vụ tiếp nhận yêu cầu từ người dùng dưới dạng văn bản hoặc hình ảnh. Sau đó, nội dung sẽ được tóm tắt và các thực thể quan trọng như URL, email, số điện thoại và từ khóa sẽ được trích xuất, tạo nền tảng cho các bước xác minh tiếp theo.

Đây là bước đặc biệt quan trọng vì nó quyết định chất lượng thông tin đầu vào mà các tác tử tiếp theo sẽ dựa vào để xử lý. Mục tiêu của tác tử này là rút gọn yêu cầu

chính như yêu cầu kiểm chứng, nội dung nghi ngờ, và giữ lại các từ khóa cốt lõi để giảm nhiễu do ngữ cảnh dài dòng. Tác tử sử dụng kết hợp giữa khả năng xử lý của LLM và hàm trích xuất truyền thống để phát hiện các thực thể như đường dẫn URL, địa chỉ email và số điện thoại.

Để giảm thiểu rủi ro xảy ra lỗi chồng lỗi trong hệ thống đa tác tử, tức là khi tác tử đầu tiên xử lý sai dẫn đến việc các tác tử tiếp theo tiếp tục sử dụng thông tin sai lệch. Input Analyzer được thiết kế với cơ chế kiểm tra chéo giữa kết quả của mô hình LLM và kết quả của các hàm trích xuất. Nếu phát hiện mâu thuẫn hoặc thiếu sót trong thông tin trích xuất, tác tử sẽ thực hiện lại bước phân tích thay vì chuyển tiếp ngay cho các tác tử khác. Điều này tạo một lớp lọc đầu vào có khả năng tự hiệu chỉnh.

#### **3.4.2. Kiểm tra độ tin cậy của thực thể trích xuất (Entity Checker)**

Sau khi các thực thể quan trọng như địa chỉ email, số điện thoại và URL được trích xuất từ thông tin đầu vào, tác tử tiếp theo là Entity Checker sẽ thực hiện đánh giá mức độ đáng tin cậy của từng thực thể bằng cách tích hợp các công cụ xác minh chuyên dụng. Mục tiêu của bước này là phát hiện nhanh chóng và hiệu quả các dấu hiệu bất thường, nguy cơ lừa đảo hoặc rủi ro bảo mật tiềm ẩn, cung cấp cơ sở vững chắc để hỗ trợ quy trình xác minh thông tin một cách toàn diện.

**Kiểm tra URL với VirusTotal API:** Hệ thống tích hợp VirusTotal API – nền tảng bảo mật uy tín tổng hợp từ hơn 70 công cụ chống mã độc – để phân tích độ an toàn của URL, phát hiện sớm các liên kết nguy hiểm. VirusTotal cung cấp thông tin chi tiết về mức độ độc hại, loại hình mối đe dọa (phishing, malware, spam), lịch sử quét, thời điểm đăng ký tên miền, giúp hệ thống nhanh chóng cảnh báo người dùng về rủi ro tiềm ẩn.

**Xác minh email và số điện thoại với Abstract API:** Hệ thống sử dụng Abstract API để kiểm tra tính hợp lệ, độ tin cậy của địa chỉ email và số điện thoại. Đối với email, Abstract API xác minh sự tồn tại thực tế, đánh giá mức độ uy tín và kiểm tra lịch sử spam hoặc lừa đảo. Với số điện thoại, API kiểm tra tính hợp lệ, quốc gia đăng

ký, nhà mạng cung cấp, loại hình dịch vụ, và cảnh báo nếu xuất hiện trong các cơ sở dữ liệu spam, lừa đảo. Khả năng này đặc biệt hữu ích trong việc phát hiện sớm các nghi vấn thường được sử dụng trong các chiến dịch lừa đảo trực tuyến qua tin nhắn SMS hoặc cuộc gọi giả mạo.

Việc sử dụng kết hợp các công cụ xác minh uy tín và mạnh mẽ như VirusTotal và Abstract API không chỉ nâng cao đáng kể độ chính xác và khả năng phát hiện rủi ro của hệ thống, mà còn giúp tự động hóa quy trình xử lý một cách nhanh chóng và hiệu quả. Điều này đặc biệt quan trọng trong các tình huống kiểm chứng tin tức trực tuyến, nơi mà tính kịp thời và độ tin cậy của thông tin luôn là ưu tiên hàng đầu.

#### ***3.4.3. Phân tích tên miền và IP (Domain Analyst)***

Tác tử Domain Analyst thực hiện phân tích sâu các thông tin kỹ thuật liên quan tới tên miền và địa chỉ IP thông qua truy vấn WHOIS và các nguồn dữ liệu IP uy tín. Quá trình này bao gồm kiểm tra ngày đăng ký tên miền, trạng thái thông tin sở hữu (ẩn danh hay công khai), tổ chức quản lý tên miền, và nguồn gốc IP. Việc phân tích chi tiết này nhằm phát hiện sớm những tên miền có dấu hiệu nguy hiểm hoặc liên quan đến các chiến dịch lừa đảo phổ biến, hỗ trợ việc xác minh và cảnh báo kịp thời cho người dùng.

#### ***3.4.4. Tìm kiếm thông tin mở trên internet (Web Researcher)***

Tác tử Web Researcher sử dụng nội dung đã được tóm tắt từ Input Analyzer để làm truy vấn đầu vào, kết hợp với Google Search API để tìm kiếm thông tin bổ sung từ các nguồn công khai trên internet. Tác tử này đặc biệt ưu tiên các nguồn tin chính thống, bài viết từ các trang báo uy tín, cơ quan truyền thông đáng tin cậy hoặc thông tin chính thức từ các tổ chức và cơ quan nhà nước.

Sau khi nhận kết quả tìm kiếm, tiến hành truy cập, trích xuất và làm sạch nội dung bằng cách loại bỏ hoàn toàn các thành phần dư thừa như thẻ HTML, quảng cáo, thông tin điều hướng và các dữ liệu không liên quan trực tiếp đến nội dung chính. Chỉ những



nội dung thật sự có giá trị và liên quan đến yêu cầu xác minh mới được giữ lại để sử dụng làm dữ liệu bổ sung.

Trong trường hợp không có kết quả nào phù hợp, tác tử này sẽ chủ động bỏ qua việc thu thập dữ liệu từ internet, đảm bảo rằng thông tin đưa vào quá trình xác minh cuối cùng luôn có độ chính xác và mức độ liên quan cao nhất.

#### **3.4.5. Truy vấn thông tin từ cơ sở dữ liệu (Database Researcher)**

Tác tử **Database Researcher** chịu trách nhiệm truy vấn dữ liệu từ **cơ sở dữ liệu nội bộ** – nơi lưu trữ các bài viết, tin tức đã được chọn lọc và xử lý từ những nguồn chính thống như báo chí nhà nước, cổng thông tin điện tử của các tổ chức uy tín, hoặc cơ quan chính phủ. Trước khi được lưu vào cơ sở dữ liệu, các tài liệu này đã trải qua quy trình làm sạch nội dung, loại bỏ HTML, chuẩn hóa định dạng và loại trừ dữ liệu nhiễu, đảm bảo độ tin cậy và sẵn sàng phục vụ cho các tác vụ xác minh tự động.

Trong quá trình truy vấn, hệ thống áp dụng chiến lược **Hybrid Search**, kết hợp đồng thời hai kỹ thuật, với **BM25** được sử dụng để truy xuất sơ bộ các tài liệu có mức độ trùng khớp từ khóa cao với truy vấn đầu vào, **Semantic Embedding Search** sử dụng mô hình embedding MiniLM để mã hóa truy vấn và tài liệu thành vector, sau đó đo độ tương đồng ngữ nghĩa bằng *cosine similarity*.

Danh sách kết quả từ cả hai phương pháp trên sau đó được **tái xếp hạng (re-ranking)** bằng thuật toán **TF-IDF kết hợp độ tương tự cosine**, nhằm chọn lọc những tài liệu thực sự phù hợp cả về từ vựng lẫn ngữ nghĩa với truy vấn xác minh.

Sau khi có danh sách các văn bản tiềm năng, tác tử sẽ tiếp tục thực hiện bước **đọc hiểu ngữ cảnh**, phân tích nội dung chi tiết và loại bỏ các kết quả nhiễu, không liên quan hoặc trùng lặp. Chỉ những tài liệu có tính xác thực và đóng góp trực tiếp cho quá trình đánh giá nội dung nghi vấn mới được giữ lại, làm **ngữ cảnh hỗ trợ đầu vào** cho bước xác minh tổng hợp tiếp theo trong hệ thống RAG của kiến trúc Multi-Agent.

### **3.4.6. Tổng hợp và đánh giá cuối cùng (Final Verifier)**

Tác tử cuối cùng – Final Verifier – tổng hợp và phân tích toàn bộ kết quả thu được từ các tác tử trước đó (bao gồm thông tin thực thể, phân tích domain/IP, dữ liệu mở trên internet và dữ liệu nội bộ). Dựa trên những dữ liệu đã xác minh này, tác tử đưa ra đánh giá tổng thể và kết luận cuối cùng về mức độ đáng tin cậy của thông tin đầu vào, đồng thời cung cấp giải thích rõ ràng về căn cứ của kết luận này.

Sau khi tác tử Final Verifier đưa ra kết luận, nếu kết quả đánh giá cho thấy thông tin có dấu hiệu gian lận hoặc nguy hiểm, hệ thống sẽ tiến hành một bước xử lý bổ sung, không thuộc phạm vi tác vụ của các AI Agent, gọi là bước ánh xạ hành vi (TTP Mapping). Ở bước này, hệ thống sẽ đối chiếu kết luận của tác tử cuối cùng với các mẫu chiến thuật, kỹ thuật và thủ tục (TTP – Tactics, Techniques, Procedures) lừa đảo đã được lưu trữ trong cơ sở dữ liệu. Quá trình này giúp xác định chính xác và chi tiết loại chiến dịch hoặc mô hình tấn công lừa đảo cụ thể mà thông tin đầu vào có thể đang sử dụng, qua đó hỗ trợ công tác phân tích, theo dõi và ngăn chặn các mối đe dọa tiềm tàng trong tương lai một cách chủ động và hiệu quả.

## **3.5. Xây dựng cơ sở dữ liệu**

Để hỗ trợ cho quá trình truy vấn và xác minh thông tin một cách hiệu quả, hệ thống đề xuất xây dựng một cơ sở dữ liệu nội bộ bao gồm các thông tin có tính xác thực cao, được thu thập từ các nguồn chính thống và được xử lý trước nhằm đảm bảo độ sạch và nhất quán của dữ liệu. Cơ sở dữ liệu này đóng vai trò làm kho tri thức hỗ trợ cho các tác tử như Database Researcher, đồng thời hỗ trợ lưu trữ thông tin phục vụ phân tích hành vi và cải tiến hệ thống trong các phiên bản sau.

Hệ thống cơ sở dữ liệu được thiết kế dưới dạng quan hệ, bao gồm ba bảng chính:

### 3.5.1. Bảng *news\_table* - Tin tức đã xác minh

STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả ngắn gọn
1	news_id	INTEGER (PK)	Mã định danh duy nhất cho bài viết
2	title	TEXT	Tiêu đề của bài viết
3	publish_date	DATE	Ngày xuất bản bài viết
4	source	TEXT	Nguồn báo chí đăng tải
5	content	TEXT	Nội dung chi tiết của bài viết

**Bảng 3.1:** Bảng *news\_table*

### 3.5.2. Bảng *ttp\_table* – Danh sách hành vi gian lận

STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả ngắn gọn
1	ttp_id	INTEGER (PK)	Mã định danh hành vi TTP
2	pattern	TEXT	Mẫu phát hiện
3	fraud_type	TEXT	Loại hành vi gian lận
4	technique	TEXT	Mô tả kỹ thuật gian lận cụ thể
5	source	TEXT	Nguồn nghiên cứu ghi nhận

**Bảng 3.2:** Bảng *ttp\_table*

### 3.5.3. Bảng *history\_table* – Lưu lịch sử tương tác

STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả ngắn gọn
1	case_id	INTEGER (PK)	Mã định danh của phiên tương tác
2	user_input	TEXT	Yêu cầu người dùng gửi đến
3	response	TEXT	Phản hồi cuối cùng hệ thống trả về
4	agent1_res	TEXT	Kết quả tóm tắt và trích xuất thông tin
5	agent2_res	TEXT	Kết quả kiểm tra thực thể
6	agent3_res	TEXT	Phản hồi kiểm tra thông tin miền, IP
7	agent4_res	TEXT	Kết quả tìm kiếm web
8	agent5_res	TEXT	Kết quả truy vấn nội bộ
9	agent6_res	TEXT	Nhận định xác thực cuối cùng
10	ttp_id	INTEGER (FK)	Mã ttp ánh xạ đến bảng ttp_table
11	feedback	TEXT	Góp ý từ người dùng
12	timestamp	DATETIME	Thời điểm bắt đầu diễn ra tương tác

**Bảng 3.3:** Bảng *history\_table*

Việc lưu trữ dữ liệu được xử lý và chuẩn hóa vào cơ sở dữ liệu SQLite nhằm đảm bảo tính nhẹ, dễ triển khai và phù hợp cho các hệ thống AI cục bộ hoặc dạng nguyên mẫu (prototype).

Toàn bộ cơ sở dữ liệu được thiết kế nhằm đáp ứng ba mục tiêu chính: (1) cung cấp nguồn thông tin tham chiếu có độ tin cậy cao để đối chiếu trong quá trình xác minh, (2) hỗ trợ phát hiện và phân loại các chiến dịch gian lận qua ánh xạ TTP, và (3) lưu vết toàn bộ lịch sử truy vấn – phản hồi nhằm phục vụ mục tiêu cải tiến hệ thống liên tục.

## 3.6. Thiết kế phần mềm hệ thống

### 3.6.1. Thiết kế hệ thống API phục vụ xác minh thông tin

Để hiện thực hóa hệ thống xác minh thông tin một cách hiệu quả, nhóm nghiên cứu sẽ tiến hành thiết kế một tập hợp các API RESTful đóng vai trò như xương sống cho các tác vụ liên quan đến xử lý đầu vào, truy vấn dữ liệu, đánh giá độ tin cậy và hỗ trợ người dùng.

**Bảng 3.4:** Danh sách các API chính sử dụng trong hệ thống

Phương thức	Endpoint	Chức năng
POST	/add_news	Thêm tin tức vào hệ thống
POST	/add_ttps	Thêm dữ liệu TTP
POST	/add_ttps_form_file	Nhập TTP từ file CSV/Excel
POST	/ttp_embeddings	Embedding và lưu FAISS cho TTP
DELETE	/delete_news	Xoá tin tức dựa trên ID
POST	/pipeline_crawl_news	Pipeline thu thập dữ liệu
POST	/verify_input	Xác minh thông tin đa tác vụ (văn bản hoặc hình ảnh)
GET	/get_news	Lấy danh sách tin tức đã lưu
GET	/get_ttps	Lấy danh sách TTP đã lưu
GET	/get_history	Lấy lịch sử hỗ trợ người dùng
POST	/feedback	Góp ý, đánh giá của người dùng

Các API được thiết kế module hóa, dễ mở rộng và tích hợp với giao diện người dùng. Trong đó, `verify_input` là API chính của hệ thống xác minh, được tích hợp trực tiếp với kiến trúc multi-agent. Khi người dùng gửi nội dung cần kiểm chứng, API này sẽ phân phối tác vụ cho các tác tử chuyên biệt cùng thực hiện nhiệm vụ xác minh tin tức.

### **3.6.2. Các thành phần trong giao diện người dùng**

Nhằm mang lại trải nghiệm tốt nhất, hệ thống được thiết kế với giao diện trực quan, tối giản và dễ thao tác. Mỗi thành phần trong giao diện đều có vai trò riêng, được sắp xếp hợp lý để người dùng có thể dễ dàng làm quen và sử dụng. Mục tiêu là tạo ra một công cụ không chỉ hiệu quả mà còn thân thiện, phù hợp với mọi đối tượng người dùng.

Hình 3.3 trình bày các chức năng chính trong giao diện hệ thống. Cụ thể:

#### **1. Khung nhập liệu thông tin:**

Phần nhập liệu là nơi hệ thống tiếp nhận thông tin từ người dùng để bắt đầu quá trình xác minh. Tại đây, người dùng có thể lựa chọn giữa việc nhập văn bản trực tiếp hoặc tải lên hình ảnh có chứa nội dung cần kiểm chứng. Phần này được thiết kế tối giản và dễ sử dụng, giúp đảm bảo thao tác diễn ra dễ dàng, phù hợp với cả người dùng phổ thông lẫn chuyên môn. Đây là bước mở đầu quan trọng, giúp quy trình xử lý và phân tích thông tin một cách hiệu quả.

#### **2. Nút "Xác thực thông tin":**

Khi hoàn tất việc nhập liệu, người dùng chỉ cần nhấn nút xác thực để gửi yêu cầu đến hệ thống. Thông tin ngay lập tức được chuyển đến API `verify_input`, nơi quy trình phân tích và kiểm chứng sẽ được khởi động, với sự phối hợp của nhiều tác tử AI nhằm đánh giá độ tin cậy của nội dung được cung cấp.

#### **3. Thành phần phản hồi kết quả phân tích:**

Phần hiển thị kết quả xác minh là nơi người dùng theo dõi phản hồi xác minh từ hệ thống. Tại đây, hệ thống sẽ trình bày rõ ràng đánh giá về tính xác thực của thông tin (“An toàn”, “Đáng ngờ” hoặc “Lừa đảo”), kèm theo lời giải thích và các căn cứ được sử dụng để đưa ra kết luận. Nếu có thông tin được xác minh đi kèm, nội dung liên quan cũng sẽ được hiển thị trực tiếp để người dùng tham khảo. Ngoài ra, phần này còn đưa ra các gợi ý hành động phù hợp, giúp người dùng cẩn trọng khi chia sẻ hoặc báo cáo nội dung nghi vấn.

#### **4. Thành phần nhập góp ý về phản hồi:**

Đây là nơi người dùng phản hồi lại hệ thống về chất lượng của kết quả xác minh. Phản hồi này có thể là lời nhận xét, bổ sung nguồn tin hoặc chỉ ra sai sót. Các phản hồi sẽ được hệ thống lưu trữ và phân tích định kỳ để cải thiện thuật toán xác minh và hiệu quả xử lý thông tin. Điều này góp phần xây dựng một môi trường xác minh mang tính cộng đồng và không ngừng học hỏi.

#### **5. Khung hiển thị ánh xạ TTP:**

Đây là một chức năng nâng cao giúp người dùng hiểu rõ hơn về các mối đe dọa tiềm ẩn liên quan đến nội dung đang xác minh. Khi hệ thống phát hiện dấu hiệu khả nghi trong thông tin đầu vào, nó sẽ tự động kích hoạt quá trình dò tìm và ánh xạ TTP.

Cơ chế hoạt động như sau: nội dung văn bản người dùng cung cấp sẽ được đối chiếu với cơ sở dữ liệu TTP đã được biên soạn sẵn. Việc đối chiếu này được thực hiện bằng các phương pháp tìm kiếm kết hợp, bao gồm so khớp theo biểu thức chính quy và tìm kiếm tương tự theo embedding (sử dụng FAISS). Nếu phát hiện có sự tương đồng cao với một mẫu TTP nào đó, hệ thống sẽ hiển thị thông tin chi tiết cho người dùng.

Các mẫu TTP trong hệ thống bao gồm nhiều dạng hành vi phổ biến của tội phạm mạng, chẳng hạn như:

1. Giả mạo cơ quan nhà nước (ví dụ: mạo danh công an, toà án để đe dọa).
2. Lừa đảo ngân hàng (giả mạo OTP, khoá tài khoản, yêu cầu xác thực).
3. Tuyển dụng việc nhẹ lương cao (dẫn dụ vào các đường link xấu).
4. Lừa đảo qua tin nhắn người thân (giả vờ con gặp nạn, nhờ chuyển tiền).
5. Chiêu trò phát tán mã độc, dụ tải app lạ hoặc quét mã QR độc hại.

Kết quả ánh xạ sẽ bao gồm: tên dạng tấn công (TTP), nhóm hành vi, đường link dẫn tới nguồn tham chiếu (thường là trang MITRE ATT&CK), và mô tả cụ thể. Ví dụ: nếu người dùng nhập nội dung như “mẹ ơi con đang gặp nạn, con cần tiền gấp”, hệ thống sẽ nhận diện đây là một hành vi thuộc TTP “Giả mạo người thân – T1204”

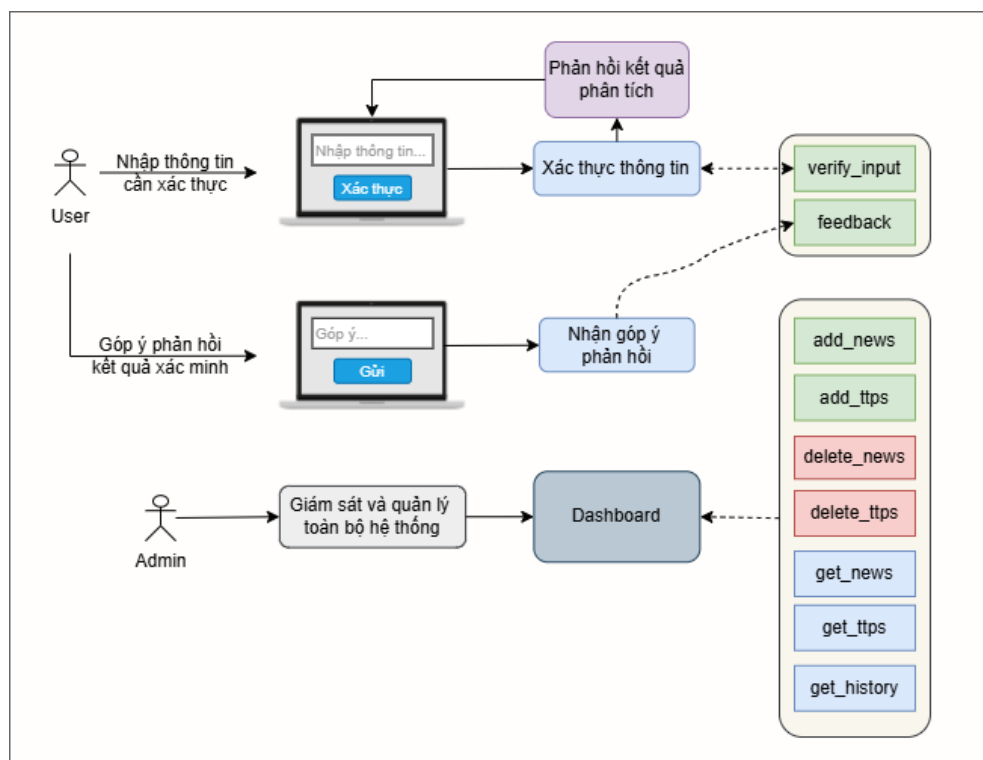
và cảnh báo người dùng tương ứng.

Thông qua tính năng này, hệ thống không chỉ đưa ra cảnh báo mà còn cung cấp ngữ cảnh chuyên sâu để người dùng hiểu rõ lý do tại sao thông tin bị đánh giá là rủi ro. Điều này góp phần nâng cao nhận thức, giúp người dùng không chỉ phát hiện mà còn tự bảo vệ mình trước các chiến thuật lừa đảo ngày càng tinh vi.

## 6. Bảng điều khiển (Dashboard quản trị):

Dashboard là nơi tổng hợp mọi hoạt động của hệ thống, cung cấp cho chúng tôi cái nhìn toàn cảnh về kết quả xác minh từ hệ thống và phản hồi và đánh giá của người dùng qua bản ghi lịch sử.

Ngoài ra còn hỗ trợ tương tác trực tiếp với cơ sở dữ liệu như xóa, sửa hay thêm mới thông tin. Đây là công cụ quan trọng để duy trì độ tin cậy, cập nhật các nguồn tin mới nhất, và kịp thời điều chỉnh hệ thống khi có sai sót xảy ra.



**Hình 3.3:** Sơ đồ tổng quan phần mềm hệ thống hoạt động



## CHƯƠNG 4. THÍ NGHIỆM VÀ ĐÁNH GIÁ

Chương này trình bày quá trình thử nghiệm và đánh giá hệ thống, bao gồm xây dựng bộ dữ liệu huấn luyện, thiết lập mô hình, các tiêu chí đánh giá, cũng như phân tích kết quả nhằm kiểm chứng hiệu quả của hệ thống trong thực tế.

### 4.1. Thu thập và xây dựng bộ dữ liệu huấn luyện

#### 4.1.1. Nguồn dữ liệu và quá trình thu thập

Dữ liệu trong nghiên cứu được xây dựng từ hai nguồn chính:

**Nguồn dữ liệu có sẵn:** Nghiên cứu này kế thừa các tập dữ liệu đã được gán nhãn từ hai bài báo “Utilizing Transformer Models to Detect Vietnamese Fake News on Social Media Platforms” [7] và “Detecting Vietnamese Fake News” [16]. Cả hai nguồn đều cung cấp dữ liệu văn bản tiếng Việt, đã được phân loại thành hai nhãn real và fake, phản ánh đa dạng các chủ đề từ xã hội, chính trị đến tin tức giả mạo phổ biến trên mạng xã hội. Đây là nguồn dữ liệu nền tảng, giúp hình thành bộ tập huấn luyện ban đầu cho mô hình trong nghiên cứu.

**Nguồn dữ liệu thu thập bổ sung:** Nhằm tăng cường tính đa dạng và đảm bảo tính cập nhật cho bộ dữ liệu, nghiên cứu đã xây dựng một pipeline thu thập tự động từ các nguồn báo chí chính thống và có độ tin cậy cao tại Việt Nam. Hệ thống được hiện thực hóa bằng thư viện BeautifulSoup, triển khai dưới dạng API để nhận đầu vào là danh sách các URL từ các trang như VnExpress, Dân Trí, Thanh Niên, Nhân Dân, Công An, VTV,... Từ mỗi bài viết, hệ thống tự động trích xuất tiêu đề, nội dung, ngày đăng và liên kết bài viết, sau đó lưu trữ vào cơ sở dữ liệu phục vụ cho quá trình huấn luyện và kiểm thử mô hình.

Ngoài ra, chúng tôi còn thu thập bổ sung các nội dung liên quan đến hành vi lừa đảo và giả mạo được lan truyền trên các nền tảng phổ biến như SMS, Telegram, Facebook Messenger, và các cảnh báo từ cơ quan chức năng (Bộ TT&TT, Cục An toàn thông tin). Các trường hợp được thu thập bao gồm hành vi giả mạo cơ quan nhà nước, lừa đảo tuyển dụng, mời gọi đầu tư tài chính không minh bạch, và giả danh ngân hàng để đánh cắp dữ liệu cá nhân — đây đều là các hình thức phát tán thông tin sai lệch phổ biến với mức độ nguy hại cao trong bối cảnh hiện nay.

#### ***4.1.2. Đặc điểm của bộ dữ liệu***

Bộ dữ liệu được xây dựng với sự đa dạng về thể loại tin tức và về nội dung, phản ánh thực tiễn phức tạp của các loại hình tin tức và thông tin sai lệch tại Việt Nam. Cụ thể, dữ liệu bao gồm các nhóm đặc điểm sau:

**Tin tức chính thống:** Đây là các văn bản tin tức được thu thập từ các cơ quan báo chí uy tín, phản ánh các vấn đề về chính trị, kinh tế, xã hội, pháp luật và thời sự trong nước. Nhóm dữ liệu này có đặc điểm ngôn ngữ chuẩn mực, cấu trúc logic, có nguồn trích dẫn rõ ràng, thể hiện phong cách hành chính – công vụ phổ biến trong truyền thông báo chí chính thống.

**Tin giả và tin sai lệch:** Bao gồm các tin xuyên tạc, bóp méo thông tin sự kiện, tạo dựng hoang tin nhằm thu hút sự chú ý, thao túng dư luận hoặc trục lợi tài chính.. Một số tin giả mang tính giật gân, ngôn từ cảm tính, thiếu kiểm chứng, và thường được chia sẻ trên các nền tảng mạng xã hội với tốc độ lan truyền nhanh.

**Nội dung lừa đảo và giả mạo:** Bao gồm các thông tin giả mạo cơ quan nhà nước, ngân hàng, công ty tuyển dụng hoặc cá nhân, nhằm chiếm đoạt thông tin hoặc tài sản của người dùng. Những nội dung này thường xuất hiện ở dạng tin nhắn từ đa nền tảng mạng xã hội và có xu hướng ngôn ngữ giả lập tính tin cậy nhằm đánh lừa người đọc.

Đặc điểm này giúp mô hình học được các dấu hiệu ngôn ngữ đặc trưng, từ đó tăng cường khả năng phát hiện tin giả và các hành vi lừa đảo phức tạp trong môi trường tiếng Việt.

### **4.1.3. Xử lý và gán nhãn dữ liệu**

Sau quá trình thu thập, toàn bộ dữ liệu được đưa vào quy trình tiền xử lý nhằm đảm bảo chất lượng và tính nhất quán cho việc huấn luyện mô hình. Việc tiền xử lý là bước thiết yếu trong quy trình, đặc biệt đối với dữ liệu thu thập từ nhiều nguồn không kiểm soát. Do hai bộ dữ liệu được tác giả sử dụng được thu thập tự động từ các nguồn báo chí và mạng xã hội, nên có khả năng một bản tin có thể xuất hiện lặp lại từ nhiều nguồn khác nhau. Điều này dẫn đến tình trạng dữ liệu trùng lặp, khiến mô hình có nguy cơ học quá mức từ một số mẫu nhất định (overfitting), làm suy giảm khả năng tổng quát hóa đối với dữ liệu mới.

Ngoài ra, trong quá trình thu thập, có những bản tin mang tính mơ hồ hoặc không đủ thông tin để đưa ra nhãn chính xác. Nếu giữ lại các trường hợp này trong tập huấn luyện sẽ gây nhiễu, khiến mô hình khó phân biệt ranh giới giữa tin thật và tin giả một cách rõ ràng. Vì vậy, việc lọc bỏ những tin tức trùng lặp và không rõ ràng không chỉ giúp tăng độ tin cậy của tập dữ liệu mà còn góp phần giảm chi phí tính toán, rút ngắn thời gian huấn luyện và nâng cao hiệu quả học của mô hình bằng cách loại bỏ các thông tin không mang giá trị.

Quy trình xử lý bao gồm loại bỏ các thẻ HTML, định dạng markdown và các yếu tố kỹ thuật nhằm giữ lại nội dung văn bản cần thiết. Chuẩn hóa nội dung văn bản về chữ thường, chuẩn hóa dấu câu, xử lý lỗi chính tả phổ biến, thống nhất định dạng ngày tháng. Lọc dữ liệu không hợp lệ, trùng lặp, tin rác, hoặc không mang giá trị ngữ nghĩa.

Quá trình gán nhãn được thực hiện tự động dựa trên nguồn dữ liệu và đặc điểm nội dung. Cụ thể, các bài viết thu thập từ báo chí chính thống được gán nhãn real (tin thật), trong khi các nội dung được xác minh là sai sự thật, có yếu tố giả mạo hoặc lừa đảo, thực hiện gán nhãn fake (tin giả). Sau đó được kiểm tra và hiệu chỉnh lại thủ công để đảm bảo độ chính xác cao.

#### 4.1.4. Tăng cường dữ liệu

Sử dụng thay thế bằng từ đồng nghĩa có kiểm soát với WordNet cùng hoán vị cấu trúc câu nhằm tăng cường khả năng tổng quát hoá của mô hình. Nội dung ngữ nghĩa cốt lõi được giữ nguyên cùng nhãn gán. Bước này giúp mở rộng tập dữ liệu hiện có mà không cần thu thập thêm dữ liệu thô, từ đó cải thiện hiệu suất của mô hình khi gặp các biểu hiện ngữ nghĩa mới trong thực tế.

Bộ dữ liệu cuối cùng thu được 4.221 dữ liệu, bao gồm 2.299 tin giả và 1.922 tin thật. Tiến hành chia bộ dữ liệu thành tập huấn luyện, tập xác thực và tập kiểm thử. Trong đó, tập huấn luyện dùng để đào tạo mô hình, tập xác thực phục vụ điều chỉnh siêu tham số, và tập kiểm thử được sử dụng để đánh giá khách quan hiệu suất của mô hình.

**Bảng 4.1:** Phân bố dữ liệu theo nhãn và từng tập huấn luyện

Label	Total Numbers	Train	Val	Test
Fake	2,299	1,391	454	454
Real	1,922	1,143	389	390
Total	4,221	2,534	843	844

## 4.2. Thiết lập huấn luyện mô hình

Sau khi hoàn tất quá trình xử lý và chuẩn bị dữ liệu, chúng tôi tiến hành thiết lập cấu hình huấn luyện nhằm tối ưu hiệu suất trong điều kiện tài nguyên tính toán có giới hạn. Việc lựa chọn tham số được cân nhắc kỹ lưỡng để đảm bảo mô hình hội tụ ổn định, tránh quá khớp (overfitting), đồng thời duy trì hiệu quả trong việc học biểu diễn ngữ nghĩa.

Một trong những tham số được sử dụng là `target_modules` – quy định các lớp trọng yếu trong kiến trúc Transformer. Bảy module được lựa chọn bao gồm các

lớp thuộc khối Attention như `q_proj`, `k_proj`, `v_proj`, `o_proj`, và các lớp trong mạng Feed-Forward như `gate_proj`, `up_proj`, và `down_proj`. Việc chèn các adapter LoRA vào những vị trí này giúp mô hình học được các điều chỉnh nhỏ nhưng quan trọng, từ đó tăng khả năng tùy biến mà không cần cập nhật toàn bộ tham số.

Ngoài ra, các tham số khác cũng được thiết lập dựa trên đặc thù của mô hình và phần cứng sử dụng. Những lựa chọn này giúp cân bằng giữa tốc độ huấn luyện và chất lượng mô hình đầu ra.

Bảng 4.2 trình bày các tham số cốt lõi được áp dụng trong toàn bộ quá trình tinh chỉnh. Các mô hình đều được huấn luyện theo cùng một cấu hình nhằm đảm bảo tính nhất quán và công bằng trong quá trình đánh giá, từ đó cho phép phân tích rõ ràng hiệu năng của từng kiến trúc trên cùng tập dữ liệu.

**Bảng 4.2:** Các tham số huấn luyện chính trong quá trình fine-tune

Tham số	Giá trị	Mô tả
max_seq_length	2048	Độ dài chuỗi tối đa, cho phép mô hình xử lý văn bản dài, phù hợp với ngữ cảnh phức tạp.
lora_r	16	Rank của ma trận low-rank, cân bằng giữa hiệu suất và chi phí bộ nhớ.
lora_alpha	16	Hệ số điều chỉnh ảnh hưởng của adapter lên đầu ra mô hình.
target_modules	7 module	Các lớp trọng yếu được áp dụng LoRA để tăng hiệu quả học biểu diễn.
batch_size	2	Batch nhỏ do giới hạn GPU, dùng tích lũy gradient để bù.
grad_steps	4	Giúp mô phỏng batch lớn hơn, tăng ổn định khi cập nhật trọng số.
epochs	3	Số vòng lặp trên toàn bộ dữ liệu huấn luyện.
learning_rate	2e-4	Tốc độ học tối ưu cho fine-tune từ mô hình pre-trained.
weight_decay	0.01	Giảm overfitting bằng cách điều chuẩn trọng số.
optimizer	AdamW (8-bit)	Giảm bộ nhớ và tăng tốc độ huấn luyện.
lr_scheduler	Linear	Giảm dần learning rate theo thời gian huấn luyện.

### 4.3. Triển khai phần mềm ứng dụng

Mục tiêu của phần này là xây dựng giao diện phần mềm phục vụ cho người dùng xác minh thông tin một cách trực quan và dễ sử dụng. Giao diện phần mềm được thiết kế để tối ưu hóa trải nghiệm người dùng, cho phép họ dễ dàng tương tác với hệ thống và nhận kết quả xác minh một cách nhanh chóng và chính xác.

#### 4.3.1. API xử lý và phân tích thông tin

Hệ thống API đóng vai trò then chốt trong việc xử lý và phân tích các yêu cầu của người dùng, từ thu thập dữ liệu đến việc xác minh tính xác thực của thông tin. Các API này được xây dựng bằng FastAPI, một framework nhanh chóng và dễ sử dụng để xây dựng API RESTful. FastAPI cung cấp tính năng tự động tạo tài liệu API thông qua Swagger, giúp cho việc tích hợp và kiểm thử trở nên dễ dàng hơn. Các API này phục vụ nhiều tác vụ, bao gồm thu thập dữ liệu từ các nguồn báo điện tử, xác minh thông tin, tìm kiếm và quản lý cơ sở dữ liệu.

API trong hệ thống đảm nhiệm các nhiệm vụ thu thập thông tin từ các nguồn báo điện tử, lưu trữ dữ liệu vào cơ sở dữ liệu, và thực hiện các tác vụ xác minh tin tức dựa trên văn bản hoặc hình ảnh. Các phương thức HTTP như GET, POST, và DELETE được sử dụng để xử lý các yêu cầu từ người dùng, giúp hệ thống hoạt động hiệu quả. Việc sử dụng FastAPI giúp hệ thống API trở nên mạnh mẽ và dễ bảo trì, đồng thời tối ưu hóa quá trình xử lý và phân tích dữ liệu.

Một trong những đặc điểm quan trọng của hệ thống API là khả năng mở rộng và linh hoạt. Các API được thiết kế module hóa, có thể dễ dàng tích hợp với các hệ thống khác hoặc mở rộng thêm các tính năng mới mà không làm ảnh hưởng đến toàn bộ hệ thống.

#### 4.3.2. *Giao diện người dùng với Next.js*

Giao diện người dùng của hệ thống xác minh thông tin được phát triển bằng Next.js, một framework hiện đại giúp xây dựng các ứng dụng website động với khả năng tối ưu hóa hiệu suất. Next.js không chỉ hỗ trợ server-side rendering (SSR) mà còn có khả năng static site generation (SSG), cho phép trang web được tải nhanh hơn nhờ vào việc pre-render nội dung phía server trước khi gửi tới người dùng. Điều này giúp tăng tốc độ tải trang và cung cấp trải nghiệm người dùng mượt mà, đặc biệt trong các ứng dụng yêu cầu hiệu suất cao.

Ngoài việc tối ưu hiệu suất, Next.js còn hỗ trợ các cấu trúc trang động thông qua các tính năng như API routes và dynamic imports, giúp xây dựng giao diện động mà không cần tải lại trang, tạo ra trải nghiệm người dùng liên tục và không gián đoạn. Điều này rất quan trọng khi người dùng tương tác với hệ thống, vì họ sẽ nhận được phản hồi ngay lập tức mà không phải chờ đợi lâu, đặc biệt là khi xác minh các tin tức có độ phức tạp cao.

Quan trọng hơn hết, Next.js cũng rất mạnh mẽ trong việc tích hợp với các API được xây dựng bằng FastAPI, vì nó dễ dàng hỗ trợ các kết nối HTTP, cho phép giao diện người dùng tương tác với backend mà không gặp vấn đề về hiệu suất hoặc độ trễ. Khi người dùng gửi yêu cầu xác minh thông tin, hệ thống giao diện sẽ gửi các yêu cầu đến các API RESTful, nhận kết quả phân tích và hiển thị chúng một cách trực quan.

Những điểm mạnh khác của Next.js là tính năng Automatic Static Optimization, giúp tự động tối ưu hóa các trang tĩnh, giảm thiểu thời gian tải lại và tăng hiệu suất, đồng thời hỗ trợ SEO để hệ thống dễ dàng tiếp cận người dùng qua các công cụ tìm kiếm. Ngoài ra, tính năng Incremental Static Regeneration (ISR) cho phép cập nhật các trang tĩnh sau một khoảng thời gian mà không cần tái biên dịch toàn bộ, rất hữu ích trong các hệ thống yêu cầu cập nhật thông tin theo thời gian thực, như việc xác minh tin tức từ các nguồn báo điện tử.




#### 4.3.3. Các thành phần trong giao diện người dùng


**FENSE - Hệ thống xác thực thông tin**

Nhập văn bản hoặc hình ảnh cần kiểm chứng vào bên dưới để hệ thống xử lý.

Tôi nhận được tin trúng thưởng sau, có thật sự vậy không:  
Bạn vừa trúng thưởng 1 chiếc điện thoại Iphone16, hãy nhấn vào link sau để nhận thưởng: [trungthuong.com.vn](http://trungthuong.com.vn)



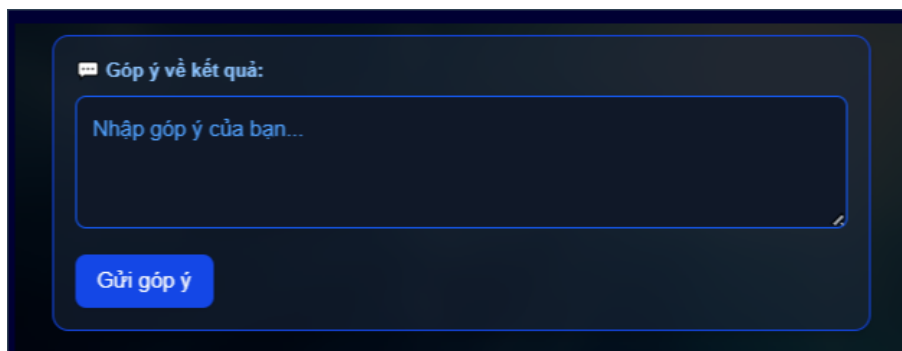
**Xác thực thông tin**

 **Phân tích AI:**

1. **Kết luận:** ❌ Lừa đảo
2. **Giải thích lý do:** Không có thông tin xác thực nào từ các nguồn đáng tin cậy như Apple hoặc các tổ chức uy tín khác xác nhận về chương trình trúng thưởng iPhone 1
6. Các thông tin từ web results không liên quan đến chương trình trúng thưởng này. Thông thường, các thông báo trúng thưởng không rõ nguồn gốc thường là dấu hiệu của lừa đảo.:
3. **Gợi ý hành động:** Người dùng nên bỏ qua thông báo này và không cung cấp bất kỳ thông tin cá nhân nào. Nếu có nghi ngờ, hãy liên hệ trực tiếp với các tổ chức chính thức để xác minh thông tin.

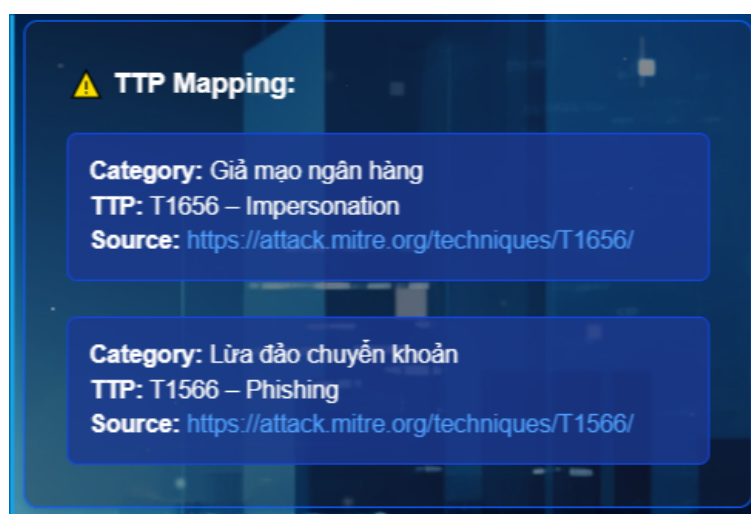
**Hình 4.1:** Giao diện hệ thống xác minh thông tin

Hình 4.1 hiển thị giao diện tổng quan của hệ thống xác minh thông tin. Đây là phần giao diện chính, nơi người dùng nhập thông tin cần xác minh (văn bản hoặc hình ảnh). Các thành phần quan trọng bao gồm khung nhập liệu, nơi người dùng cung cấp dữ liệu đầu vào, và ấn "Xác thực thông tin", kích hoạt quá trình xác minh. Sau khi hệ thống hoàn tất xác minh, kết quả phân tích sẽ được hiển thị rõ ràng, giúp người dùng đánh giá mức độ tin cậy của thông tin đã nhập.



**Hình 4.2:** Khung nhập góp ý của người dùng về phản hồi xác minh

Hình 4.2 minh họa khung nhập góp ý của người dùng. Thành phần này cho phép người dùng gửi phản hồi về kết quả xác minh thông tin, giúp hệ thống nhận diện các vấn đề và cải thiện khả năng phân tích trong tương lai.



**Hình 4.3:** Khung ánh xạ thông tin chiến dịch TTP liên quan

Hình 4.3 minh họa khung ánh xạ TTP, nơi hiển thị các mối đe dọa có liên quan đến nội dung đang được xác minh. Khi hệ thống phát hiện dấu hiệu bất thường, khung này sẽ liệt kê các kỹ thuật tấn công tương ứng đã được ghi nhận trước đó. Người dùng có thể nhấn vào các liên kết để xem chi tiết từng kỹ thuật, qua đó hiểu rõ hơn về bản chất và mức độ nguy hiểm của hành vi lừa đảo liên quan. Tính năng này góp phần tăng cường tính minh bạch và cung cấp kiến thức nền giúp người dùng nhận diện các chiến dịch lừa đảo hiệu quả hơn.

## 4.4. Đánh giá và so sánh hiệu năng mô hình

### 4.4.1. Các chỉ số đánh giá

Trong báo cáo này, Các tiêu chí đánh giá được sử dụng là Accuracy, F1-Score, Precision và Recall. Độ chính xác phản ánh tỷ lệ dự đoán đúng trên tổng số mẫu, trong khi F1-Score là trung bình điều hòa giữa độ chính xác (precision) và độ hồi tưởng (recall), giúp đánh giá hiệu năng cân bằng giữa khả năng phát hiện đúng tin giả và tránh dự đoán sai.

Ngoài các chỉ số truyền thống, nhóm còn bổ sung một chỉ số quan trọng khác là **URC (Undecided Response Count)**, đại diện cho số lượng mẫu mà mô hình không đưa ra được một phản hồi phân loại rõ ràng – tức không xác định được nhãn "real" hoặc "fake" như yêu cầu.

URC được đưa vào nhằm phản ánh khả năng mô hình hiểu đúng yêu cầu nhiệm vụ và tuân thủ định dạng đầu ra trong các hệ thống xác minh thông tin. Cung cấp một góc nhìn thực tiễn hơn về hiệu năng trong môi trường ứng dụng thật – nơi việc phản hồi dứt khoát, đúng yêu cầu và đúng định dạng là yếu tố bắt buộc. Mô hình có URC thấp không chỉ thể hiện hiệu quả về mặt phân loại, mà còn cho thấy mức độ ổn định, đáng tin cậy và khả năng tương tác tốt với hệ thống hoặc người dùng đầu cuối.

### 4.4.2. Phân tích, nhận xét kết quả

Dưới đây là bảng trình bày kết quả so sánh hiệu năng trước và sau fine-tuned

**Bảng 4.3:** So sánh hiệu suất các mô hình trước và sau tinh chỉnh

Model	Accuracy	F1-Score	Precision	Recall	URC
<b>Zero-shot</b>					
Mistral-7B	0.455	0.4684	0.7081	0.455	194
Llama3.2-3B	0.6315	0.6718	0.7275	0.6315	114
Qwen2.5-3B	0.6647	0.6371	0.7623	0.6647	33
Meta-Llama3.1-8B	0.609	0.5678	0.7099	0.609	44
Gemma2-9B	0.7737	0.7809	0.7929	0.7737	20
<b>Fine-Tuned</b>					
Llama3.2-3B	0.8021	0.8155	0.8335	0.8021	26
Qwen2.5-3B	0.8199	0.8229	0.8448	0.8199	14
Meta-Llama3.1-8B	0.7654	0.7733	0.8394	0.7654	30
<b>Gemma2-9B</b>	<b>0.9159</b>	<b>0.9160</b>	<b>0.9193</b>	<b>0.9159</b>	<b>0</b>

Bảng 4.3 trình bày kết quả thực nghiệm của các mô hình trong hai giai đoạn đánh giá: **zero-shot** (chưa qua huấn luyện bổ sung) và **fine-tuned** (sau khi tinh chỉnh với tập dữ liệu chuyên biệt tiếng Việt). Việc cải thiện đồng thời các chỉ số đánh giá sau quá trình fine-tuning, bao gồm Accuracy, F1-Score, Precision, Recall và URC (Undecided Response Count), chứng tỏ rằng mô hình không chỉ nâng cao hiệu suất cùng độ ổn định mà còn cải thiện khả năng hiểu ngôn ngữ và đảm bảo phản hồi đúng yêu cầu. Điều này góp phần nâng cao hiệu quả của mô hình trong ngữ cảnh tiếng Việt phức tạp.

#### **Hiệu suất của các mô hình trong giai đoạn Zero-shot.**

Trong giai đoạn zero-shot (trước khi thực hiện huấn luyện bổ sung), mô hình

Gemma2-9B cho thấy hiệu suất vượt trội so với các mô hình còn lại, với Accuracy đạt 77.37% và F1-Score đạt 78.09%. Điều này cho thấy khả năng tổng quát hóa tốt của Gemma2-9B và sự thích ứng mạnh mẽ với ngữ cảnh tiếng Việt ngay cả khi chưa qua quá trình fine-tuning.

Ngược lại, mô hình Mistral-7B thể hiện kết quả thấp nhất với Accuracy chỉ đạt 45.5%, cùng với Recall thấp và URC lên đến 194 (hơn 20% mẫu đầu vào không được không phản hồi đúng yêu cầu). Những chỉ số này cho thấy Mistral-7B không đáp ứng đủ yêu cầu phân loại trong ngữ cảnh tiếng Việt, với tỷ lệ phản hồi mơ hồ hoặc không rõ ràng về mặt phân loại (“real” hay “fake”). Phản ánh việc mô hình này không đáp ứng các yêu cầu hiệu suất cơ bản. Do đó, Mistral-7B đã được loại khỏi quá trình fine-tuning và không được đưa vào so sánh trong các giai đoạn tiếp theo.

Các mô hình còn lại như LLaMA3.2-3B, Qwen2.5-3B, và Meta-LLaMA3.1-8B có hiệu suất trung bình, nhưng URC vẫn còn khá cao, cho thấy mô hình chưa hoàn toàn ổn định và cần cải thiện khả năng tuân thủ yêu cầu phân loại.

### **Ảnh hưởng của quá trình Fine-tuning đến hiệu suất mô hình.**

Kết quả fine-tuning với kỹ thuật LoRA mang lại sự cải thiện rõ rệt về hiệu suất của các mô hình. Sau khi tinh chỉnh, tất cả các mô hình đều ghi nhận sự gia tăng đáng kể về Accuracy và F1-Score, với sự cải thiện ít nhất 10 điểm phần trăm so với giai đoạn zero-shot. Đặc biệt, mô hình **Gemma2-9B (Fine-Tuned)** đạt kết quả xuất sắc với Accuracy là 91.59%, F1-Score là 91.60%, Precision đạt 91.93%, và Recall đạt 91.59%.

Điều quan trọng là Gemma2-9B không chỉ cải thiện về độ chính xác mà còn đạt URC = 0, tức là không có mẫu nào bị phân loại sai hay phản hồi mơ hồ. Điều này chứng tỏ rằng sau quá trình fine-tuning, mô hình đã học và hiểu các yêu cầu phân loại trong ngữ cảnh tiếng Việt, đồng thời giảm thiểu các lỗi phản hồi không phù hợp.

### **Ý nghĩa của chỉ số URC đối với hiệu quả và độ tin cậy của mô hình.**

URC (Undecided Response Count) là một chỉ số quan trọng phản ánh khả năng của mô hình trong việc đáp ứng yêu cầu phân loại chính xác. Một mô hình có Accuracy

cao nhưng URC lớn có thể dẫn đến những sai sót nghiêm trọng trong thực tế, chẳng hạn như không đưa ra nhãn phân loại rõ ràng hoặc trả lời một cách mơ hồ, không đáp ứng đúng yêu cầu của hệ thống.

Việc URC giảm xuống 0 sau khi fine-tuning, đặc biệt là đối với Gemma2-9B, cho thấy mức độ tiến bộ rõ rệt trong việc giúp mô hình tuân thủ yêu cầu đầu ra một cách chính xác và nhất quán. Điều này có ý nghĩa quan trọng trong việc đảm bảo tính ổn định của mô hình khi được triển khai trong các ứng dụng thực tế.

### **Nhận xét.**

Kết quả thực nghiệm cho thấy rằng fine-tuning với kỹ thuật LoRA đã mang lại hiệu quả rõ rệt trong việc nâng cao hiệu suất và độ ổn định của mô hình.

Việc áp dụng LoRA không chỉ cải thiện các chỉ số truyền thống như Accuracy và F1-Score mà còn giúp giảm thiểu các lỗi phản hồi sai lệch, nâng cao khả năng tuân thủ yêu cầu phân loại. Mô hình Gemma2-9B fine-tuned nổi bật với độ chính xác cao và khả năng phản hồi đúng yêu cầu, làm gương mẫu cho các hệ thống kiểm chứng tin tức tiếng Việt.

Quá trình fine-tuning không chỉ giúp cải thiện hiệu suất mà còn giúp mô hình học được các đặc trưng ngữ nghĩa cần thiết, đảm bảo khả năng triển khai thực tế hiệu quả và tiết kiệm chi phí, góp phần quan trọng trong việc xây dựng các hệ thống xác minh tin tức tự động.

## **CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

### **5.1. Kết luận**

Nghiên cứu này tập trung vào việc phát triển một hệ thống hỗ trợ xác minh thông tin ứng dụng các công nghệ hiện đại. Xuất phát từ thực trạng thông tin sai lệch lan truyền ngày càng nhiều trên mạng xã hội và các nền tảng số, vốn có thể gây tác động tiêu cực đến nhận thức và hành vi của người dùng. Hệ thống được xây dựng với mục tiêu cung cấp một công cụ hỗ trợ kiểm chứng hiệu quả, minh bạch và dễ sử dụng.

Sau quá trình nghiên cứu, thiết kế và triển khai thực nghiệm, khóa luận đã giải quyết các câu hỏi nghiên cứu bằng cách xây dựng bộ dữ liệu đa dạng, phản ánh phong phú các loại tin tức trong môi trường tiếng Việt, giúp mô hình học tốt các đặc trưng ngôn ngữ để phân loại chính xác thông tin. Việc tinh chỉnh mô hình đã góp phần giảm thiểu lỗi đánh giá và nâng cao hiệu quả phát hiện tin giả. Đồng thời, triển khai hệ thống với kiến trúc Multi-Agent, tích hợp đầy đủ công cụ kiểm chứng, không chỉ đưa ra kết luận mà còn cung cấp lời giải thích rõ ràng giúp người dùng phân biệt được hành vi và chiến thuật đứng sau các tin giả, tin lừa đảo.

### **5.2. Hướng phát triển**

Trong tương lai, hệ thống có thể nâng cao khả năng kiểm chứng và ứng dụng thực tế bằng cách mở rộng để hỗ trợ xác minh đa phương tiện như video, âm thanh và phát hiện nội dung giả mạo qua giọng nói, hình ảnh, nhằm đáp ứng nhu cầu xác minh thông tin trong các tình huống ngày càng phức tạp và đa dạng.

## TÀI LIỆU THAM KHẢO

- [1] S. Ahmad, M. Khan, and S. Kumari, “Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models,” *Procedia Computer Science*, vol. 218, pp. 2761–2770, 2023.
- [2] M. Alenezi and A. Alkadi, “Multilingual email phishing attacks detection using osint and machine learning,” in *Proceedings of the 2023 International Conference on Cybersecurity and Artificial Intelligence*, [Online]. Available: <https://arxiv.org/abs/2501.08723>, 2023.
- [3] M. E. Almandouh, M. F. Alrahmawy, M. Eisa, M. Elhoseny, and A. S. Tolba, “Ensemble based high performance deep learning models for fake news detection,” *Scientific Reports*, 2024.
- [4] CrewAI, *Crewai: Build agentic workflows with multi-agent collaboration*, CrewAI Documentation, 2024.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019.
- [6] E. Hu, Y. Shen, P. Wallis, *et al.*, *Lora: Low-rank adaptation of large language models*, arXiv preprint, arXiv:2106.09685, 2021.
- [7] A.-T. Huynh and P. Tran, “Utilizing transformer models to detect vietnamese fake news on social media platforms,” *KSII Transactions on Internet and Information Systems*, vol. 19, no. 2, pp. 1234–1249, 2025. [Online]. Available: <https://itiis.org/digital-library/102085>.



- [8] C. Ireton and J. Posetti, *Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training*. UNESCO Publishing, 2018.
- [9] E. C. T. Jr, Z. W. Lim, and R. Ling, “Defining “fake news”: A typology of scholarly definitions,” *Digital Journalism*, 2018.
- [10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: <https://arxiv.org/abs/2005.11401>.
- [11] S. M. M. Monterrubio, A. Noain-Sánchez, E. V. Pérez, and R. G. Crespo, *Coronavirus fake news detection via medosint check in health care official bulletins with cbr explanation: The way to find the real information source through osint, the verifier tool for official journals*, Universidad Internacional de La Rioja - UNIR, 2023.
- [12] OpenAI, *Gpt-4 technical report*, 2023. [Online]. Available: <https://openai.com/research/gpt-4>.
- [13] K. Shu, *Fakenewsnet*, 2021. [Online]. Available: <https://github.com/KaiDMML/FakeNewsNet>.
- [14] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
- [15] Statista, *Prevalence of online misinformation worldwide*, 2022.
- [16] V. D. Vinh and P. Do, “Detecting vietnamese fake news,” *Can Tho University Journal of Science*, vol. 58, no. 1, pp. 55–64, 2022. [Online]. Available: <https://ctujs.ctu.edu.vn/index.php/ctujs/article/view/680/650>.

- [17] Y. W. Wang, ““liar, liar pants on fire”: A new benchmark dataset for fake news detection,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 422–426.
- [18] C. Wardle and H. Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe, 2017.
- [19] J. Welbl and S. Stenning, *Fine-tuned ‘small’ llms (still) significantly outperform zero-shot generative ai models in text classification*, arXiv preprint, arXiv:2402.01325, 2024.
- [20] M. Wooldridge, *An Introduction to MultiAgent Systems*. Wiley, 2009.
- [21] A. Yadav, A. Kumar, and V. Singh, “Open-source intelligence: A comprehensive review of the current state, applications and future perspectives in cyber security,” *Cybersecurity Journal*, 2023.
- [22] M. Zhang, D. Lee, and Y. Zhao, *Analysis of disinformation and fake news detection using fine-tuned large language model*, arXiv preprint, arXiv:2309.04704, 2023. [Online]. Available: <https://arxiv.org/html/2309.04704>.