RESEARCH ARTICLE

NGHIÊN CỨU HỆ ĐA TÁC NHÂN DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN CHO BÀI TOÁN KIỂM CHỨNG THÔNG TIN

Lê Huỳnh Nghiêm[1], Trần Xuân Hiệp[2], Tiếu Phùng Mai Sương[1*]

[1] Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ - Tin học TP.HCM
[2] Khoa Kỹ thuật Công nghệ , Trường Đại học Phú Yên
21dh114462@st.huflit.edu.vn, hieptranxuan@pyu.edu.vn, suongtpm@huflit.edu.vn,*

TÓM TẮT — Bài báo đề xuất một phương pháp kiểm chứng thông tin mới dựa trên hệ thống đa tác nhân sử dụng mô hình ngôn ngữ lớn (LLM), cụ thể là GPT, nhằm nâng cao tính chính xác và khả năng giải thích trong đánh giá tính xác thực của các tuyên bố. Phương pháp này khai thác khả năng nhập vai của LLM để tạo nên ba tác nhân phản biện với các lập trường và định hướng suy luận khác nhau. Qua đó, hệ thống cho phép mô phỏng quá trình tranh luận đa chiều, tạo nền tảng cho việc trích xuất đặc trưng sâu bằng mô hình học sâu. Kết quả thực nghiệm trên tập dữ liệu GossipCop cho thấy phương pháp của chúng tôi đạt hiệu năng cạnh tranh so với các mô hình tiên tiến hiện nay.

Từ khóa — Mô hình Ngôn ngữ lớn, Hệ đa tác nhân, Kiểm chứng thông tin.

I. GIỚI THIỆU

Sự phát triển không ngừng của các nền tảng truyền thông xã hội đã thúc đẩy việc lan truyền thông tin ở quy mô chưa từng có. Tuy nhiên, điều này cũng kéo theo sự gia tăng của các nội dung sai lệch và tin giả, đặt ra yêu cầu cấp thiết về các hệ thống có khả năng kiểm chứng thông tin một cách hiệu quả và minh bạch. Trong bối cảnh đó, việc khai thác sức mạnh của mô hình Ngôn ngữ lớn (LLM) để xây dựng các hệ thống kiểm chứng tự động không chỉ là một hướng đi đầy hứa hẹn mà còn phù hợp với xu hướng phát triển của trí tuệ nhân tạo hiện đại.

Khác với các mô hình truyền thống thường hoạt động như "hộp đen", khó diễn giải và thiếu tính minh bạch, các LLM hiện đại có khả năng lập luận và tạo văn bản có tính giải thích cao. Tuy nhiên, chính khả năng sáng tạo ngôn ngữ mạnh mẽ cũng khiến chúng dễ rơi vào hiện tượng “ảo giác” – tạo ra thông tin có vẻ hợp lý nhưng thực chất không chính xác.

Trong nghiên cứu này, chúng tôi phát triển một hệ thống đa tác nhân kết hợp LLM nhằm mô phỏng quá trình phản biện như một hội đồng kiểm chứng thông tin. Mỗi tác nhân được thiết kế với vai trò và định hướng cụ thể, cho phép mô hình phân tích thông tin từ nhiều góc độ. Kết hợp với các kỹ thuật học sâu, hệ thống có thể trích xuất các đặc trưng ngữ nghĩa và lập luận đa chiều, góp phần nâng cao độ chính xác và khả năng giải thích của quá trình kiểm chứng.

Phần còn lại của bài báo được tổ chức như sau: Mục 2 trình bày cơ sở lý thuyết và các công trình liên quan; Mục 3 giới thiệu phương pháp đề xuất; Mục 4 trình bày các kết quả thực nghiệm; và cuối cùng, Mục 5 là phần kết luận của bài báo.

II. CƠ SỞ LÝ THUYẾT VÀ CÁC CÔNG TRÌNH LIÊN QUAN

A. CƠ SỞ LÝ THUYẾT

1. TÁC VỤ KIỂM TRA THÔNG TIN VÀ PHÁT HIỆN TIN GIẢ

Tác vụ kiểm tra thông tin (fact-checking) đề cập đến quá trình đánh giá tính xác thực của một tuyên bố thông qua việc suy luận dựa trên các nguồn dữ liệu đóng vai trò làm bằng chứng. Tuy nhiên, việc truy xuất và xác minh các bằng chứng này trong môi trường thực tế thường gặp nhiều thách thức, đặc biệt khi yêu cầu tính sẵn sàng và phản hồi theo thời gian thực của các hệ thống tự động.

Trong khi đó, phát hiện tin giả (fake news detection) được xem là một nhánh chuyên biệt của kiểm tra thông tin, chủ yếu tập trung vào việc nhận diện các thông tin sai lệch có chủ đích, thường được tạo ra nhằm thao túng nhận thức hoặc gây ảnh hưởng tiêu cực đến cộng đồng. Không giống như kiểm tra thông tin truyền thống vốn dựa vào suy luận logic từ bằng chứng, các phương pháp phát hiện tin giả thường khai thác các đặc trưng ngôn ngữ học hoặc hành vi lan truyền của văn bản để đưa ra phán đoán. Do không yêu cầu truy xuất bằng chứng trực tiếp, độ chính xác của các mô hình này có thể bị hạn chế. Tuy vậy, nhờ khả năng triển khai nhanh và phù hợp với các tình huống thời gian thực, tác vụ phát hiện tin giả vẫn thu hút sự quan tâm đáng kể trong thực tiễn ứng dụng. Đặc biệt, trong bối cảnh thông tin lan truyền nhanh chóng trên mạng xã hội, việc thu thập và xác thực đầy đủ bằng chứng cho từng tuyên bố là một thách thức lớn, khiến các giải pháp tự động hóa trở nên cần thiết.

2. MÔ HÌNH NGÔN NGỮ LỚN

Mô hình Ngôn ngữ lớn là các mô hình học sâu với số lượng tham số cực lớn, thường từ hàng trăm triệu đến hàng trăm tỷ, được huấn luyện trên các tập dữ liệu văn bản khổng lồ từ nhiều nguồn khác nhau như sách, báo, trang web, và mạng xã hội. Nhờ khả năng học biểu diễn ngữ nghĩa và cú pháp phức tạp trong ngôn ngữ tự nhiên, LLMs có thể sinh văn bản mạch lạc, giàu ngữ cảnh và phù hợp với mục đích sử dụng.

Các mô hình như GPT-3.5, GPT-4 (do OpenAI phát triển), LLaMA (Meta), Claude (Anthropic) hay Gemini (Google DeepMind) đã chứng minh hiệu quả vượt trội trong nhiều tác vụ xử lý ngôn ngữ tự nhiên (NLP), bao gồm dịch máy, trả lời câu hỏi, tóm tắt văn bản, phân tích cảm xúc, viết mã lập trình, và thậm chí là lập luận logic. Ngoài ra, thông qua kỹ thuật prompting, các mô hình này còn có thể thích ứng nhanh với nhiều tác vụ mới mà không cần huấn luyện lại, mở ra hướng tiếp cận linh hoạt và tiết kiệm chi phí cho các ứng dụng thực tiễn.

Sự phát triển của LLMs không chỉ đánh dấu bước tiến trong lĩnh vực NLP, mà còn thúc đẩy nhiều hướng nghiên cứu liên ngành như học tăng cường với phản hồi từ con người (Reinforcement Learning with Human Feedback – RLHF), học tự giám sát (self-supervised learning), và trí tuệ nhân tạo đa phương thức, góp phần định hình thế hệ AI tiếp theo.

3. HỆ THỐNG ĐA TÁC NHÂN TÍCH HỢP MÔ HÌNH NGÔN NGỮ LỚN

Hệ thống đa tác nhân (*Multi-Agent System – MAS*) [1] là kiến trúc gồm nhiều tác nhân (*agents*) hoạt động độc lập, có khả năng tương tác, trao đổi thông tin và phối hợp nhằm giải quyết các vấn đề phức tạp. Mỗi tác nhân có thể được thiết kế với mục tiêu riêng biệt và chiến lược hành động riêng, từ đó tạo ra một môi trường hợp tác hoặc cạnh tranh linh hoạt.

Khi tích hợp với LLM, agents không chỉ thực hiện các hành động tự động hóa truyền thống, mà còn có thể tận dụng khả năng hiểu và sinh ngôn ngữ tự nhiên để mô phỏng lập luận chuyên sâu, phản biện có hệ thống, và đưa ra các quan điểm đa chiều trong quá trình ra quyết định. Điều này đặc biệt hữu ích trong các tác vụ yêu cầu đánh giá tính xác thực của thông tin, chẳng hạn như fact-checking, phát hiện tin giả, hoặc giải thích nội dung.

Sự phối hợp giữa nhiều LLM và agents còn cho phép triển khai các chiến lược tranh luận (*debate-style reasoning*), phản biện chéo (*cross-examination*), hoặc phân chia vai trò như người xác minh, người phản biện, và người trung lập, từ đó nâng cao tính khách quan và độ chính xác trong các hệ thống AI giải thích và ra quyết định.

B. CÁC CÔNG TRÌNH LIÊN QUAN

Trong những năm gần đây, nhiều hướng nghiên cứu đã được đề xuất nhằm cải thiện hiệu năng và tính minh bạch trong tác vụ kiểm chứng thông tin. Các công trình có thể được phân loại theo ba nhóm chính:

1. KIỂM TRA THÔNG TIN BẰNG HỌC MÁY TRUYỀN THỐNG VÀ HỌC SÂU

Các phương pháp kiểm tra thông tin ban đầu [2] chủ yếu dựa trên các mô hình học máy truyền thống sử dụng đặc trưng thủ công (hand-crafted features), bao gồm đặc trưng từ vựng, tần suất xuất hiện của cụm từ, cú pháp, và các tín hiệu ngôn ngữ đơn giản. Những đặc trưng này sau đó được đưa vào các bộ phân loại như SVM, Naïve Bayes hoặc Random Forest để dự đoán tính đúng sai của tuyên bố.

Sự phát triển của học sâu đã mở ra hướng tiếp cận mới khi các mô hình như Convolutional Neural Networks (CNN) [3] và Recurrent Neural Networks (RNN) [4] được áp dụng nhằm tự động trích xuất đặc trưng ngữ cảnh sâu hơn từ văn bản. Những mô hình này giúp cải thiện đáng kể độ chính xác so với các phương pháp truyền thống, đặc biệt trong việc nhận diện các mẫu ngôn ngữ phức tạp và ngữ nghĩa ngầm ẩn.

Một số nghiên cứu sau đó [5] đã đề xuất kết hợp biểu diễn ngữ cảnh (contextual embeddings) với thông tin về độ tin cậy của nguồn tin, tạo ra bước tiến quan trọng trong việc liên kết giữa tuyên bố và các nguồn dẫn xác thực. Tuy nhiên, các mô hình học sâu này vẫn tồn tại hạn chế đáng kể về tính giải thích được, khi quá trình ra quyết định thường diễn ra như một "black-box", khiến người dùng khó hiểu được lý do phía sau mỗi dự đoán.

2. ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN TRONG KIỂM CHỨNG

Gần đây, các mô hình ngôn ngữ lớn (LLM) như GPT [6] và PaLM [7] đã được nghiên cứu và ứng dụng trong bài toán kiểm chứng thông tin, nhờ vào khả năng hiểu ngữ nghĩa, suy luận và sinh lập luận bằng ngôn ngữ tự nhiên. Các công trình thực nghiệm cho thấy, khi được cung cấp ngữ cảnh đầy đủ hoặc tài liệu tham chiếu phù hợp, LLM có thể trả lời các truy vấn kiểm chứng với độ chính xác cao, thậm chí đưa ra lời giải thích một cách mạch lạc và thuyết phục.

Tuy nhiên, một trong những thách thức lớn khi áp dụng LLM trong thực tiễn là hiện tượng hallucination– tức mô hình có thể sinh ra thông tin sai lệch hoặc không có thật nhưng lại trình bày dưới dạng văn bản rất tự tin và hợp lý [8]. Hiện tượng này không chỉ ảnh hưởng đến độ tin cậy của mô hình, mà còn đặt ra rào cản lớn trong việc triển khai các hệ thống kiểm chứng thông tin tự động ở quy mô lớn. Ngoài ra, khả năng mô hình đưa ra những lập luận có vẻ hợp lý nhưng thực chất lại sai sự thật cho thấy tính giải thích được và kiểm chứng nội dung đầu ra vẫn là những vấn đề chưa được giải quyết triệt để.

3. MÔ HÌNH ĐA TÁC NHÂN VÀ LẬP LUẬN PHÂN TÁN

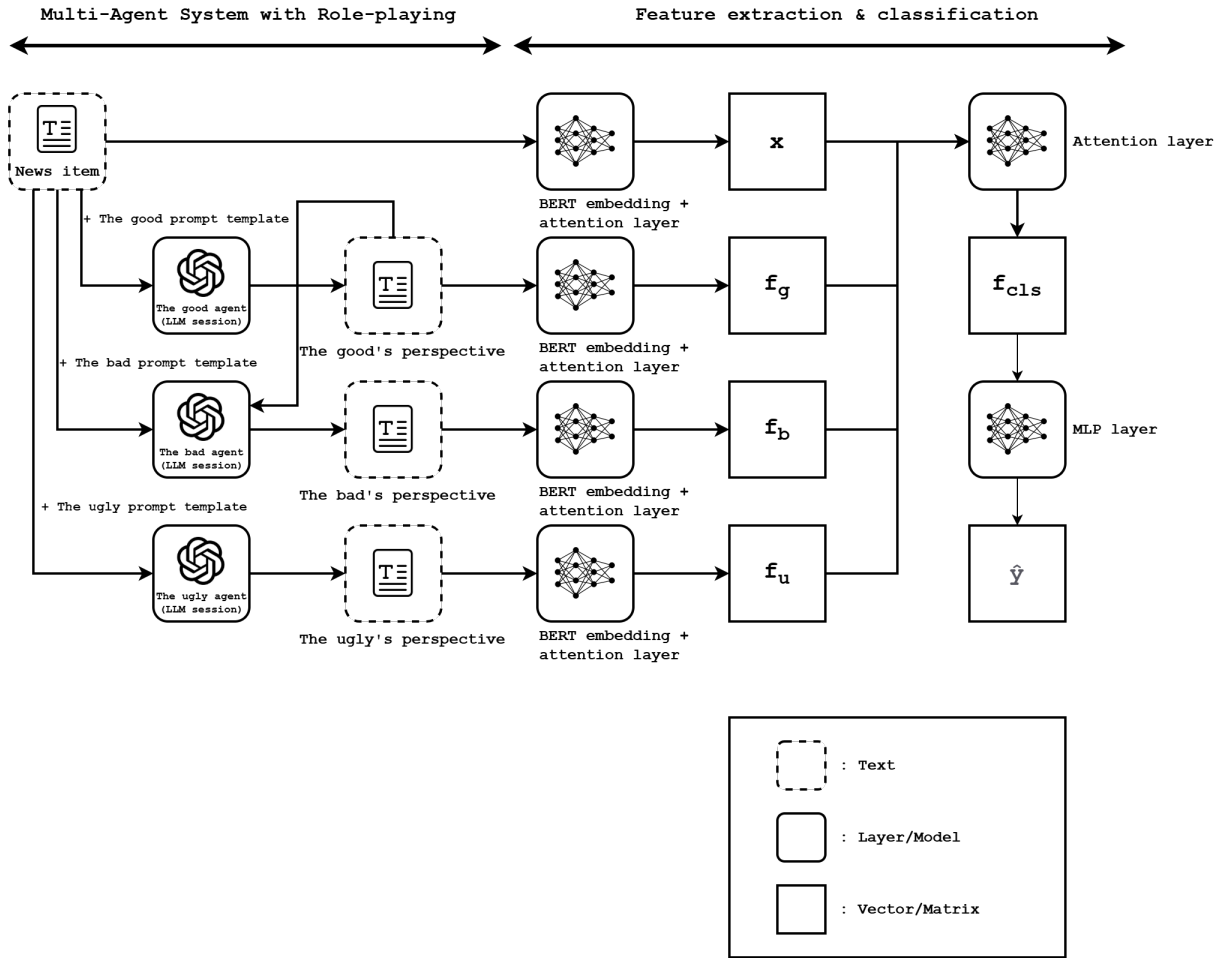
Hệ thống Multi - Agent đã được ứng dụng rộng rãi trong nhiều lĩnh vực như lập kế hoạch tự động, ra quyết định phân tán và gần đây là NLP. Trong bối cảnh này, mỗi tác nhân có thể đảm nhiệm một vai trò cụ thể, tương tác với các tác nhân khác thông qua ngôn ngữ tự nhiên để giải quyết các bài toán phức tạp một cách hiệu quả hơn so với hệ thống đơn tác nhân.

Gần đây, nhiều nghiên cứu đã khai thác tiềm năng của các mô hình ngôn ngữ lớn đa tác nhân bằng cách triển khai nhiều tác nhân GPT [9] đóng vai khác nhau để cùng thảo luận, tranh luận và ra quyết định trong các tình huống xã hội mô phỏng. Phương pháp này không chỉ cho phép mô hình hóa các quan điểm đa chiều mà còn thúc đẩy quá trình lập luận theo hướng phân tán.

Một ví dụ tiêu biểu là nghiên cứu của Kim, Kyungha et al. [10], trong đó đề xuất hệ thống "debate agents" là các tác nhân phản biện lẫn nhau nhằm xác định lập luận hợp lý và đáng tin cậy nhất. Kết quả cho thấy cách tiếp cận này giúp giảm đáng kể hiện tượng "ảo giác" trong phản hồi của LLM và nâng cao chất lượng suy luận. Tuy nhiên, hầu hết các công trình hiện tại mới chỉ tập trung vào các nhiệm vụ tổng quát như giải bài toán đạo đức, lên kế hoạch xã hội, hoặc chọn phương án hợp lý nhất trong tình huống giả lập. Việc áp dụng mô hình đa tác nhân vào bài toán kiểm chứng thông tin vẫn còn là một hướng nghiên cứu tiềm năng chưa được khai thác đầy đủ.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Chúng tôi đề xuất một mô hình hỗn hợp gồm 2 giai đoạn: hệ thống đa tác nhân nhập vai và sử dụng mô hình học sâu để trích xuất các đặc trưng. Mô hình khai thác được tính ưu việt của các hệ đa tác nhân với trung tâm là các mô hình ngôn ngữ hiện đại. Đồng thời, mô hình này cũng tận dụng khả năng suy luận từ khả năng học hỏi ngữ cảnh và đặc điểm ngôn ngữ mạnh mẽ của các mô hình deep learning hiện đại. Hình 1 biểu diễn kiến trúc tổng quan của mô hình.



Hình 1. *Sơ đồ kiến trúc tổng quát của mô hình đề xuất*

A. GIAI ĐOẠN 1: KIẾN TRÚC HỆ THỐNG ĐA TÁC NHÂN NHẬP VAI

Hệ thống đề xuất được tổ chức dựa trên kiến trúc ba tác nhân nhập vai (*role-playing agents*), hoạt động trên nền tảng LLM. Mỗi tác nhân đại diện cho một hướng suy luận riêng biệt, nhằm tái hiện quá trình tranh luận đa chiều tương tự như trong thực tế. Cơ chế nhập vai cho phép các tác nhân hình thành phong cách diễn đạt và cấu trúc lý luận phù hợp với mục tiêu giao tiếp cụ thể. Cụ thể, tác nhân *The Good* đóng vai người phát ngôn của tuyên bố, bảo vệ lập luận từ góc nhìn chủ quan, thậm chí bao gồm cả thiên kiến tiềm ẩn; tác nhân *The Bad* đảm nhận vai trò phản biện, đưa ra các lập luận phản đối một cách khách quan và trung lập, tạo ra đối trọng hợp lý với quan điểm ban đầu; trong khi đó, tác nhân *The Ugly* tạo ra các tình huống giả định mang tính sai lệch hoặc độc hại trong cùng ngữ cảnh, nhằm kiểm thử mức độ phản ứng và độ vững chắc trong lập luận của hệ thống. Các tác nhân được triển khai theo kiến trúc phân tán, nghĩa là mỗi tác nhân hoạt động độc lập với prompt riêng và tự đưa ra phản hồi. Sau đó,

các kết quả được tổng hợp lại để tiến hành phân tích tập trung. Thiết kế này cho phép hệ thống mô phỏng một cuộc tranh biện có tổ chức, tái hiện cách con người tiếp cận thông tin từ nhiều chiều cạnh khác nhau trước khi đưa ra đánh giá cuối cùng về tính xác thực của một tuyên bố.

B. GIAI ĐOẠN 2: SỬ DỤNG MÔ HÌNH HỌC SÂU ĐỂ TRÍCH XUẤT CÁC ĐẶC TRƯNG

Mẫu tin tức và các luận điểm của các tác nhân được trích xuất đặc trưng bằng cách sử dụng mô hình mã hóa BERT [11] và các lớp attention tiếp nối nhằm khai thác các đặc điểm về ngữ cảnh và ngữ nghĩa của các đoạn văn bản. Nghiên cứu triển khai học chuyển giao trên 4 mô hình BERT. Các BERT encoder được đóng băng, ngoại trừ tầng cuối. Tầng này vẫn được tham gia vào quá trình huấn luyện của giai đoạn 2. Điều này giúp mô hình thích ứng với ngữ cảnh lập luận trong khi vẫn giữ được khả năng biểu diễn ngôn ngữ chung của BERT.

Đầu ra của các bộ encoder này là các ma trận ngữ nghĩa. Nghiên cứu sử dụng thêm 1 layer mask attention đóng vai trò như một lớp Pooling Attention có thể huấn luyện được nhằm giảm số chiều của toàn bộ ma trận ngữ nghĩa. Vector này đại diện cho "tầm quan trọng tổng thể" của chuỗi đó sau khi đã tập trung vào các phần có ý nghĩa.

Vector đặc trưng được tổng hợp theo biểu thức (1) và phân loại thành dự đoán, sử dụng một mạng MLP. Với 1 item tin tức x có nhãn $y \in \{0,1\}$, mô hình tổng hợp các vector x, f_g, f_b, f_u

$$f_{cls} = \omega_x^{cls} \cdot x + \omega_g^{cls} \cdot f_g + \omega_b^{cls} \cdot f_b + \omega_u^{cls} \cdot f_u \ (1)$$

Với x, f_g, f_b, f_u lần lượt đại diện cho vector đặc trưng của mẫu tin tức và phản hồi của tác nhân *the good, the bad* và *the ugly.* Trong đó, các hệ số trọng số ω_*^{cls}được học trong quá trình huấn luyện, nhằm điều chỉnh mức độ đóng góp của từng nguồn thông tin vào vector phân loại cuối cùng. Hàm kích hoạt sigmoid được sử dụng ở lớp cuối cùng để biến đổi đầu ra thành một xác suất trong khoảng [0,1], đại diện cho độ tin cậy của mẫu tin tức thuộc lớp Real/Fake. Cụ thể, giá trị gần 1 biểu thị mức độ tin cậy cao rằng thông tin là thật, trong khi giá trị gần 0 biểu thị khả năng cao là thông tin giả. Quá trình huấn luyện được tối ưu thông qua hàm mất mát cross-entropy nhị phân:

$$L_{CE} = CE(MLP(f_{cls}, y)) \ (2)$$

Trong đó, y ∈ {0,1} là nhãn thực tế tương ứng với thông tin thật hoặc giả. Hàm *cross-entropy* đo lường sự sai khác giữa phân phối xác suất dự đoán và nhãn thực tế, giúp mô hình học cách điều chỉnh các trọng số để giảm lỗi phân loại trong quá trình huấn luyện.

C. HẠN CHẾ CỦA MÔ HÌNH DEEP LEARNING TRUYỀN THỐNG VÀ MÔ HÌNH NGÔN NGỮ LỚN HIỆN ĐẠI

Một điểm nổi bật trong thiết kế của hệ thống đề xuất là khả năng trung hòa hai hạn chế vốn tồn tại song song trong các mô hình kiểm chứng thông tin hiện nay. Thứ nhất, các mô hình truyền thống, đặc biệt là những kiến trúc học sâu như CNN, RNN hay các biến thể của BERT, thường hoạt động như một "hộp đen", chúng đưa ra kết quả phân loại nhưng không kèm theo cơ chế giải thích rõ ràng về lý do tại sao một tuyên bố được xác định là đúng hay sai. Sự thiếu minh bạch này gây khó khăn cho việc đánh giá độ tin cậy của mô hình, đồng thời làm hạn chế khả năng ứng dụng trong các lĩnh vực nhạy cảm như truyền thông, pháp lý hoặc y tế. Thứ hai, mặc dù các LLM có khả năng tạo lập văn bản diễn giải thuyết phục, chúng lại dễ mắc phải hiện tượng "ảo giác" – tức là tạo ra lập luận hoặc bằng chứng nghe có vẻ hợp lý nhưng thực chất lại không đúng với sự thật. Khi hệ thống chỉ sử dụng một tác nhân đơn lẻ, nguy cơ chấp nhận và củng cố những thông tin sai lệch do chính mô hình sinh ra là rất cao, từ đó làm suy giảm độ chính xác và độ tin cậy của quá trình kiểm chứng.

Để giải quyết hai vấn đề này, hệ thống đề xuất triển khai ba tác nhân độc lập, mỗi tác nhân mang một vai trò và thiên hướng suy luận khác nhau, tạo nên một cơ chế phản biện nội bộ mang tính đa chiều. Cụ thể, *The Good* đóng vai người đưa ra tuyên bố, phản ánh các thiên kiến và động cơ chủ quan đi kèm theo phát ngôn. *The Bad* đại diện cho quan điểm khách quan, có nhiệm vụ phản biện lập luận từ *The Good*, buộc hệ thống phải tiếp cận và xử lý các góc nhìn đối lập. Cuối cùng, *The Ugly* đưa ra các ví dụ giả định sai lệch có chủ đích, đóng vai trò như một cơ chế "kiểm thử độ nhạy" của hệ thống đối với các thông tin độc hại. Thiết kế này không những giúp mô hình giảm thiểu hiện tượng ảo giác mà còn tăng cường khả năng giải thích, từ đó nâng cao độ tin cậy và ứng dụng thực tiễn của hệ thống trong các tình huống phức tạp.

IV. THỰC NGHIỆM

A. TẬP DỮ LIỆU VÀ THIẾT LẬP MÔ HÌNH

Bảng 1.*Thông tin thống kế của tập dữ liệu GossipCop*

	Train	**Val**	**Test**
Real	**2,878**	**1,030**	**1,024**
Fake	**1,006**	**244**	**234**
Tổng cộng	**3,884**	**1,274**	**1,258**

Nghiên cứu sử dụng tập dữ liệu GossipCop [12], bao gồm các tin tức có nhãn Real/Fake và mô tả chi tiết các tuyên bố. Bảng 1 là phân bố của tập dữ liệu. Mô hình ngôn ngữ chính sử dụng là GPT – 4.1 nano [13] với khả năng tạo lập văn bản nhanh, có kiểm soát và tối ưu chi phí.

B. MẪU PROMPT ĐỂ TẠO TÁC NHÂN THÔNG MINH

Gọi một item tin tức là x, các mẫu prompt ở từng tác nhân thông minh bao gồm:

- The Good:

> "Role-play as the subject (the entity being referred to) of the statements: "x". Explain step by step whether each statement is true or false and why. No need to explain the role-playing process. Limit the response to 300 words."

- The Bad:

> There is a statement: "x". An explanation from the subject making the statement says: *the-good-perspective*. Assume the role of a neutral expert, think step by step and provide a counterargument to this statement. Limit your response to 300 words.

Với *the-good-perspective* là phản hồi từ tác nhân thông minh *The Good.*

- The Ugly:

> This is a news snippet: "x". Generate 3 pairs of statements and evidence, where each statement is unreliable, toxic, and harmful, and each pair is related to the snippet above. Only provide the sets of statements and evidence, no explanation or introduction needed. Limit to 300 words.

Nghiên cứu áp dụng kỹ thuật nhập vai (role-playing) [14] nhằm xây dựng các tác nhân với vai trò xác định tương ứng. Bên cạnh đó, kỹ thuật nhắc nhở dạng Zero-shot Chain-of-Thought [15] được triển khai để khởi tạo quá trình suy luận cho từng tác nhân. Ngoài ra, số lượng đầu ra được kiểm soát nhằm đảm bảo tính đồng nhất giữa các mẫu tin cũng như giữa các lần thực nghiệm.

C. ĐỐI CHỨNG

Trong quá trình đánh giá hiệu quả của hệ thống, nghiên cứu sử dụng ba mô hình đối chứng nhằm so sánh toàn diện giữa các hướng tiếp cận khác nhau:

- GPT-3.5-turbo với few-shot prompting [16]: đại diện cho hướng tiếp cận dựa trên LLM không cần huấn luyện thêm, mà chỉ cần cung cấp ví dụ hướng dẫn trong prompt để thực hiện kiểm chứng, mô hình này được chọn vì phản ánh xu hướng ứng dụng LLM phổ biến hiện nay nhờ khả năng xử lý linh hoạt và không cần fine-tuning.
- BERT (Bidirectional Encoder Representations from Transformers) với 1 layer MLP để thực hiện tác vụ phân loại. Việc sử dụng BERT làm đối chứng giúp đánh giá hiệu quả của mô hình đề xuất trong tương quan với một kiến trúc phổ biến đã được kiểm chứng trong nhiều bài toán NLP.
- Mô hình đề xuất (ours) kết hợp kiến trúc đa tác nhân nhập vai cùng khả năng trích xuất đặc trưng ngữ nghĩa từ BERT, lớp attention có huấn luyện, và mạng phân loại MLP.

Việc lựa chọn ba mô hình trên nhằm đảm bảo sự cân đối giữa ba hướng: (1) mô hình LLM không huấn luyện (GPT-3.5), (2) mô hình học sâu truyền thống (BERT+MLP), và (3) mô hình đề xuất tích hợp đa tác nhân và ngữ nghĩa. Điều này giúp đánh giá rõ ràng mức độ cải tiến của hệ thống trong cả khía cạnh độ chính xác và khả năng suy luận giải thích.

D. KẾT QUẢ THỰC NGHIỆM

Bảng 2. *Kết quả thực nghiệm*

Mô hình	macF1	Accuracy	$F1_{Real}$	$F1_{Fake}$
GPT-3.5-turbo [16]	**70.2%**	**81.3%**	**88.4%**	**51.9%**
BERT [16]	**76.5%**	**86.2%**	**91.6%**	**61.5%**
Ours	**76.49%**	**87.12%**	**92.30%**	**60.68%**

Kết quả thực nghiệm cho thấy mô hình đề xuất đạt hiệu suất tốt nhất trên hầu hết các chỉ số đánh giá so với hai mô hình đối chứng là GPT-3.5-turbo và BERT. Về accuracy, mô hình đề xuất đạt 87.12%, cao nhất trong ba mô hình, vượt nhẹ BERT (86.2%) và đáng kể so với GPT-3.5-turbo (81.3%). Điều này cho thấy kiến trúc đa tác nhân kết hợp với cơ chế tổng hợp đặc trưng góp phần nâng cao khả năng phân loại chính xác các mẫu tin tức. Xét theo chỉ số F1 của lớp "real", Ours tiếp tục dẫn đầu với 92.30%, cao hơn BERT (91.6%) và GPT-3.5 (88.4%), cho thấy

hiệu quả vượt trội trong việc nhận diện các thông tin thật. Ngược lại, ở lớp "fake", BERT đạt F1 cao nhất (61.5%), trong khi mô hình đề xuất có kết quả gần tương đương (60.68%) và GPT-3.5-turbo thể hiện rõ điểm yếu với F1 fake chỉ đạt 51.9%. Kết quả này phản ánh hiện tượng ảo giác phổ biến của LLMs và xu hướng thiên vị về phía thông tin thật khi không có kiểm soát chặt chẽ. Cuối cùng, xét về macro-F1 thì cả BERT (76.5%) và mô hình đề xuất (76.49%) đều vượt trội hơn GPT-3.5 (70.2%). Dù macro-F1 của mô hình đề xuất thấp hơn BERT một chút, sự cân bằng này được bù đắp bằng độ chính xác và F1 real cao hơn, khẳng định lợi thế toàn diện của kiến trúc đề xuất trong bài toán kiểm chứng thông tin.

V. KẾT LUẬN

Kết luận cho thấy rằng hệ thống đa tác nhân tích hợp LLM có thể tạo ra một mô hình kiểm chứng thông tin không chỉ đạt hiệu quả cao mà còn có khả năng giải thích tốt. Sự phối hợp giữa các tác nhân nhập vai và kỹ thuật học sâu giúp mô hình nắm bắt được các đặc điểm ngữ nghĩa tinh vi và quan hệ lập luận đa chiều, vượt qua các giới hạn của mô hình đơn tác nhân truyền thống. Hướng phát triển tiếp theo sẽ tập trung vào việc tối ưu hóa cấu trúc và vai trò của các tác nhân, cũng như mở rộng phạm vi áp dụng sang các ngôn ngữ khác và các lĩnh vực chuyên ngành như y tế, pháp lý và giáo dục, nơi yêu cầu về tính chính xác và khả năng giải thích là đặc biệt quan trọng.

VI. LỜI CẢM ƠN

Nghiên cứu được tài trở bởi Trường Đại học Ngoại ngữ -Tin học Thành phố Hồ Chí Minh trong khuôn khổ Đề tài mã số HSV2024-07.

VII. TÀI LIỆU THAM KHẢO

[1] Wooldridge, M. (2009). *An introduction to multiagent systems.* John wiley & sons, 365p.

[2] Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science* (pp. 18-22).

[3] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems, 28*.

[4] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter, 19*, 22-36.

[5] Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

[6] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Preprint.

[7] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., & Mishra, G. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research, 24*(240), 1-113.

[8] Ji, Z., Liu, Z., Lee, N., Yu, T., Wilie, B., Zeng, M., & Fung, P. (2022). RHO: Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding. *arXiv preprint arXiv:2212.01588*.

[9] Park, J., O'Brien, J., Cai, C., Morris, M., Liang, P., & Bernstein, M. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp1-22.

[10] Kim, K., Lee, S., Huang, K.-H., Chan, H., Li, M., & Ji, H. (2024). Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. arXiv preprint arXiv:2402.07401.

[11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171-4186.

[12] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data, 8*(3), 171-188.

[13] Anonymous (2025). *GPT-4.1 nano*. (OpenAI) Retrieved 6 10, 2025, from https://platform.openai.com/docs/models/gpt-4.1-nano.

[14] Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., . . . Liu, Y. (2023). Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

[15] Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems, 35*, 22199-22213.

[16] Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., & Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(20), 22105-22113.

INVESTIGATING MULTI-AGENT SYSTEMS POWERED BY LARGE LANGUAGE MODELS FOR FACT-CHECKING

Le Huynh Nghiem, Tran Xuan Hiep, Tieu Phung Mai Suong

ABSTRACT—The paper proposes a novel fact-checking method based on a multi-agent system utilizing large language models (LLMs), specifically GPT, to enhance the accuracy and explainability in evaluating the authenticity of claims. This method leverages the role-playing capabilities of LLMs to create three adversarial agents with distinct stances and reasoning orientations. Through this, the system simulates a multi-faceted debate process, providing a foundation for extracting deep features using a deep learning model. Experimental results on the GossipCop dataset demonstrate that our method achieves competitive performance compared to current state-of-the-art models.

Keywords—Large Language Models, Multi-Agent Systems, Fact-Checking

Lê Huỳnh Nghiêm hiện là sinh viên đại học, chuyên ngành Kỹ thuật phần mềm thuộc chương trình Công nghệ thông tin tại Trường Đại học Ngoại ngữ – Tin học TP. Hồ Chí Minh (HUFLIT).

Trần Xuân Hiệp nhận bằng Thạc sĩ Truyền dữ liệu và Mạng máy tính tại Học viện Bưu chính viễn thông TP. Hồ Chí Minh vào năm 2011. Hiện nay, ông là giảng viên tại khoa Khoa học tự nhiên và Công nghệ thông tin, Trường Đại học Phú Yên. Các lĩnh vực nghiên cứu chính bao gồm: Mạng máy tính và Trí tuệ nhân tạo.

Tiếu Phùng Mai Sương là Thạc sĩ Khoa học máy tính tại trường Đại học Khoa học tự nhiên TP. Hồ Chí Minh vào năm 2017. Hiện nay, cô là giảng viên tại Trường Đại học Ngoại ngữ - Tin học TP. Hồ Chí Minh (HUFLIT). Các lĩnh vực nghiên cứu gồm: Xử lý ngôn ngữ tự nhiên và Trí tuệ nhân tạo.