NGHIÊN CỨU ĐA PHƯƠNG PHÁP CHO PHÁT HIỆN Ý ĐỊNH TRONG TƯƠNG TÁC NGƯỜI–MÁY

Nguyễn Thị Thúy A *, Huỳnh Lý Tân Khoa , Nguyễn Minh Ý , Tiếu Phùng Mai Sương

Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ - Tin học TP.HCM

antt@huflit.edu.vn, *22dh114590@st.huflit.edu.vn*, *22dh114418@st.huflit.edu.vn*, *suongtpm@huflit.edu.vn*

TÓM TẮT— Phát hiện ý định là một thành phần cốt lõi trong các hệ thống tương tác người–máy và người–robot, quyết định khả năng hiểu đúng yêu cầu và phản hồi phù hợp. Nghiên cứu này thực hiện so sánh có hệ thống bốn nhóm phương pháp cho bài toán phân loại ý định gồm học máy truyền thống, học sâu, Transformer và mô hình ngôn ngữ lớn dựa trên kĩ thuật lời nhắc. Thực nghiệm được tiến hành trên bộ dữ liệu HRI gồm 8.453 câu với 6 nhãn ý định, chia dữ liệu theo tỷ lệ 80% huấn luyện, 10% kiểm định và 10% kiểm tra, đồng thời giữ cấn bằng tỷ lệ lớp giữa các tập. Trên tập kiểm tra 10%, BERT-large đạt độ chính xác 0,9882 và F1 macro 0,9881; Llama 3.1 8B huấn luyện thích nghi nhẹ đạt độ chính xác 0,9882 và F1 macro 0,9877; RoBERTa-base đạt độ chính xác 0,9764 và F1 macro 0,9743. Trong khi đó, các mô hình nhẹ hơn có mức hiệu năng thấp hơn nhưng dễ triển khai, với Naive Bayes kết hợp TF–IDF đạt độ chính xác 0,8664, TextCNN đạt 0,8180 và BiLSTM đạt 0,6738. Đối với nhóm mô hình ngôn ngữ lớn sử dụng prompting, nghiên cứu đánh giá sáu biến thể nhắc lệnh trên một tập con cố định gồm 200 mẫu từ tập kiểm tra nhằm kiểm soát chi phí suy luận, qua đó làm rõ đánh đổi giữa hiệu suất, chi phí tính toán, độ phức tạp triển khai và khả năng tổng quát hóa. Kết quả cho thấy Transformer fine-tuning và huấn luyện thích nghi nhẹ cho hiệu năng cao nhất, trong khi prompting phù hợp khi cần triển khai nhanh hoặc khai thác khả năng zero-shot trong các kịch bản có ràng buộc về dữ liệu huấn luyện.

Từ khóa— Intent Classification, Human-Robot Interaction, Transformer, Large Language Models, Prompting

I. GIỚI THIỆU

A. BỐI CẢNH NGHIÊN CỨU

Trong các hệ thống tương tác người–máy (Human–Machine Interaction) và đặc biệt là tương tác người–robot (Human–Robot Interaction – HRI), việc hiểu đúng ý định của người dùng là điều kiện tiên quyết để hệ thống có thể đưa ra phản hồi phù hợp, kịp thời và tự nhiên. Khác với các bài toán xử lý ngôn ngữ truyền thống như phân loại chủ đề hay phân tích cảm xúc, phát hiện ý định thường đòi hỏi mô hình không chỉ nhận diện các từ khóa bề mặt mà còn phải suy luận được mục tiêu giao tiếp ẩn sau phát ngôn của người dùng [1]. Điều này trở nên đặc biệt quan trọng trong bối cảnh HRI, nơi các tương tác thường mang tính ngắn gọn, không đầy đủ ngữ cảnh và chịu ảnh hưởng mạnh từ ngôn ngữ đời thường [2].

Trong giai đoạn đầu, các phương pháp học máy truyền thống dựa trên đặc trưng thủ công và các giả định thống kê đơn giản đã được sử dụng rộng rãi nhờ tính dễ triển khai và chi phí thấp. Tuy nhiên, các phương pháp này thường gặp hạn chế trong việc nắm bắt ngữ cảnh dài hạn và các quan hệ ngữ nghĩa phức tạp. Sự phát triển của học sâu đã mở ra khả năng tự động học biểu diễn từ dữ liệu [3], trong đó các mô hình tuần tự như LSTM và các kiến trúc convolution cho văn bản đã cải thiện đáng kể khả năng trích xuất đặc trưng ngữ nghĩa.

Gần đây, sự ra đời của các mô hình dựa trên Transformer và đặc biệt là các mô hình ngôn ngữ lớn đã tạo ra một bước ngoặt quan trọng. Các mô hình này không chỉ đạt hiệu suất cao thông qua tinh chỉnh mà còn cho phép giải quyết bài toán phát hiện ý định trong chế độ zero-shot hoặc few-shot thông qua kĩ thuật lời nhắc (prompting)[1]. Tuy nhiên, việc lựa chọn phương pháp phù hợp cho một hệ thống HRI cụ thể vẫn là một câu hỏi mở, đòi hỏi phải xem xét đồng thời nhiều yếu tố ngoài độ chính xác, bao gồm chi phí tính toán, độ trễ, khả năng giải thích và tính bền vững khi triển khai lâu dài.

B. MỤC TIÊU NGHIÊN CỨU

Xuất phát từ các vấn đề nêu trên, nghiên cứu này đặt ra ba mục tiêu chính. Thứ nhất, tiến hành so sánh một cách toàn diện các nhóm phương pháp phát hiện ý định từ cơ bản đến hiện đại, nhằm đóng góp một cái nhìn tổng thể về khả năng của từng phương pháp, từ đó góp phần tham khảo và đánh giá hướng đi tiềm năng cho bài toán phát hiện ý định, vì vốn dĩ việc có tập dữ liệu thực tế hoàn toàn cân bằng là không khả thi nên chúng tôi đánh giá cao thang đo F1 ở các thực nghiệm trong nghiên cứu này. Thứ hai, đánh giá tiềm năng của các mô hình ngôn ngữ lớn trong bối cảnh phát hiện ý định, đặc biệt là khả năng prompting, việc này mang lại hiệu quả tiếp cận cho độc giả có thể suy xét hướng đi trong nghiên cứu hai hướng promting – fine-tuning. Cuối cùng, nghiên cứu phân tích các đánh đổi giữa hiệu suất, chi phí và độ phức tạp triển khai, từ đó cung cấp các gợi ý thực tiễn cho việc lựa chọn phương pháp trong các kịch bản HRI khác nhau.

Phần còn lại của bài báo được trình bày như sau: Mục II giới thiệu về các nghiên cứu liên quan; mục III trình bày tập dữ liệu và phương pháp chúng tôi đánh giá. Ở mục IV chúng tôi sẽ trình bày về các mô hình, phương pháp thực nghiệm trong nghiên cứu. Kết quả của quá trình thực nghiệm sẽ được trình bày, giải thích và thảo luận ở phần V và cuối cùng là phần VI kết luận.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong ngữ cảnh tương tác người–máy (Human-Machine Interaction - HMI) và tương tác người–robot (Human-Robot Interaction - HRI), việc xác định ý định của người dùng từ ngôn ngữ tự nhiên là yếu tố then chốt để đảm bảo hệ thống hoạt động hiệu quả và an toàn. Quá trình phát triển của lĩnh vực này đã trải qua sự chuyển dịch mạnh mẽ từ các mô hình thống kê truyền thống đến các kiến trúc học sâu (Deep Learning) và gần đây nhất là các mô hình ngôn ngữ lớn (Large Language Models - LLMs).

A. CÁC PHƯƠNG PHÁP HỌC MÁY TRUYỀN THỐNG

Các nghiên cứu ban đầu trong HRI thường dựa trên học máy truyền thống, sử dụng các đặc trưng thủ công như TF-IDF kết hợp với các bộ phân loại như Naive Bayes hoặc Support Vector Machine (SVM). Ví dụ, Gervits et al. [4] đã xây dựng hệ thống đối thoại HRI sử dụng bộ phân loại thống kê với tập dữ liệu huấn luyện hạn chế cho mỗi ngữ cảnh. Mặc dù các phương pháp này có ưu điểm về tốc độ tính toán và yêu cầu tài nguyên thấp, chúng thường bộc lộ hạn chế khi xử lý dữ liệu đầu vào có tính đa dạng cao hoặc chứa nhiều nhiễu.

B. SỰ TRỖI DẬY CỦA HỌC SÂU VÀ ĐA PHƯƠNG THỨC

Sự ra đời của mạng nơ-ron hồi tiếp (RNN) như LSTM, GRU và mạng tích chập (CNN) đã cho phép hệ thống tự động học các đặc trưng ngữ cảnh và tuần tự.

Trong bối cảnh HRI, các kiến trúc lai giữa LSTM và Transformer cũng được ứng dụng để dự đoán ý định chuyển động. Mathew et al. [5] đã đạt độ chính xác trên 98% trong bài toán phân loại 16 nhãn bằng cách kết hợp thông tin ngôn ngữ với dữ liệu hình ảnh môi trường, đặc biệt hữu ích cho các robot hỗ trợ người khuyết tật.

C. TRANSFORMER VÀ MÔ HÌNH NGÔN NGỮ LỚN (LLMS)

Sự xuất hiện của kiến trúc Transformer và các mô hình tiền huấn luyện như BERT hay RoBERTa đã thiết lập những tiêu chuẩn mới trong việc hiểu ngữ cảnh ngôn ngữ phong phú. Hiện nay, xu hướng nghiên cứu chuyển dịch sang việc tinh chỉnh tham số hiệu quả (Parameter-Efficient Fine-Tuning - PEFT) cho các LLM [2].

Kỹ thuật Low-Rank Adaptation (LoRA) là một minh chứng điển hình, cho phép tối ưu hóa LLM với tài nguyên tính toán thấp nhưng vẫn đảm bảo hiệu suất. Phương pháp MIDLM của Yin et al. [6] đã ứng dụng LoRA để tinh chỉnh mô hình ngôn ngữ lớn hai chiều cho bài toán phát hiện đa ý định. Việc ứng dụng LLM tinh chỉnh bằng LoRA đặc biệt phù hợp với HRI nhờ khả năng cân bằng giữa sức mạnh tính toán và tính linh hoạt của robot.

D. PHƯƠNG PHÁP PROMPTING VÀ HỌC ÍT MẪU (FEW-SHOT LEARNING)

Song song với việc tinh chỉnh, các chiến lược Prompting đang trở thành một hướng đi tiềm năng. Thay vì huấn luyện lại mạng, người ta sử dụng mô tả ý định bằng ngôn ngữ tự nhiên để hướng dẫn mô hình.

Zero-shot Learning: Hong et al. [7] chứng minh rằng mô hình FLAN-T5 có khả năng phân loại các ý định chưa từng xuất hiện trong quá trình huấn luyện thông qua các mô tả chi tiết.

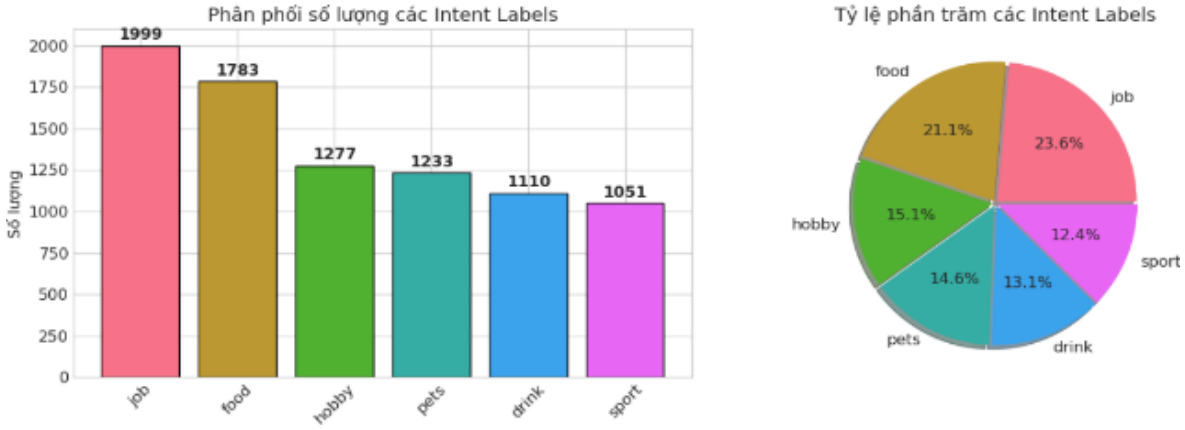
Few-shot và Chain-of-Thought: Parikh et al. [8] đã khảo sát các kỹ thuật trên LLM và chỉ ra rằng phương pháp T-Few trên các mô hình được huấn luyện theo dạng chỉ dẫn (Instruct-tuned) đạt hiệu quả tối ưu ngay cả khi chỉ có một mẫu huấn luyện duy nhất. Các chiến lược như chuỗi suy luận (Chain-of-Thought) cũng giúp cải thiện khả năng giải thích và suy luận cho robot trong các tác vụ phức tạp.

Quá trình tiến hóa của các phương pháp nhận dạng ý định trong HRI từ học máy cổ điển đến LLM đã mở ra khả năng tương tác tự nhiên và an toàn hơn. Trong khi các mô hình truyền thống vẫn giữ ưu thế về độ gọn nhẹ, các tiếp cận dựa trên LLM và Prompting mang lại khả năng thích nghi mạnh mẽ trong điều kiện dữ liệu hạn chế (Zero/Few-shot). Đối với các hệ thống HRI hiện đại, việc tích hợp khả năng hiểu ngôn ngữ sâu rộng của LLM với các kỹ thuật tối ưu hóa tham số là hướng đi then chốt để robots có thể hoạt động linh hoạt trong môi trường thực tế. Từ đó nghiên cứu của chúng tôi mang đến cái nhìn tổng quan hơn, giúp người đọc có thể nhìn rõ được mục đích và tính linh hoạt của từng phương pháp nhằm giúp lựa chọn hướng đi phù hợp với điểm mạnh nghiên cứu của mình.

III. TẬP DỮ LIỆU THỰC NGHIỆM VÀ PHƯƠNG PHÁP ĐÁNH GIÁ

A. TẬP DỮ LIỆU

Tập dữ liệu được sử dụng trong nghiên cứu này được xây dựng cho bài toán phát hiện ý định trong bối cảnh tương tác người–máy, với định hướng ban đầu là hỗ trợ các hệ thống tương tác xã hội, đặc biệt là robot dịch vụ và trợ lý ảo. Dữ liệu bao gồm tổng cộng 8.453 câu văn bản ngắn chia thành 6 lớp, mỗi câu tương ứng với một phát ngôn độc lập của người dùng và được gán một nhãn ý định duy nhất.



Hình 1. *Phân phối số lượng giữa các lớp trong tập dữ liệu*

Về mặt cấu trúc, tập dữ liệu được lưu trữ dưới định dạng CSV với hai trường chính: (i) câu đầu vào (Sentence) và (ii) nhãn ý định (Intent_label). Thiết kế tối giản này phản ánh một kịch bản thực tế phổ biến trong các hệ thống HRI, nơi mô hình cần đưa ra quyết định nhanh chóng dựa trên một phát ngôn đơn lẻ, không kèm theo thông tin hội thoại trước đó hay các nhãn slot chi tiết.

Tập dữ liệu bao gồm sáu lớp ý định chính, đại diện cho các chủ đề sinh hoạt thường ngày như thú cưng, thực phẩm, nghề nghiệp, sở thích, thể thao và đồ uống, tuy có sự mất cân bằng nhẹ ở lớp job và food cao hơn so với mặt bằng chung nhưng chúng tôi vẫn quyết định không xử lí mạnh tay mà chỉ đặt mức cân bằng trọng số (class weight) trên 2 lớp này nhẹ ở mức 0.7 để vẫn thử thách mức độ hiểu của mô hình.

Nguồn dữ liệu được tổng hợp chủ yếu từ Wikipedia và các nguồn công khai khác. Do đó, các câu trong tập dữ liệu có xu hướng mang tính mô tả, định nghĩa hoặc cung cấp thông tin tổng quát. Đặc điểm này vừa mang lại lợi thế về độ rõ ràng ngữ nghĩa, vừa tiềm ẩn nguy cơ thiên lệch về phong cách ngôn ngữ so với các hội thoại tự nhiên, một yếu tố cần được cân nhắc khi diễn giải kết quả thực nghiệm.

B. CÁC LỚP Ý ĐỊNH VÀ ĐẶC ĐIỂM NGỮ NGHĨA

Các lớp ý định trong tập dữ liệu được thiết kế sao cho phản ánh các miền nội dung quen thuộc trong tương tác người–robot. Tuy nhiên, ranh giới giữa các lớp không hoàn toàn tách biệt, do tồn tại mức độ chồng lấn ngữ nghĩa nhất định. Ví dụ, các phát ngôn liên quan đến đồ uống có thể đồng thời mang đặc trưng của thực phẩm, trong khi các phát ngôn về sở thích có thể giao thoa với thể thao hoặc nghề nghiệp.

Sự chồng lấn này tạo ra một bài toán phân loại mang tính thực tế hơn, buộc mô hình không chỉ dựa vào các từ khóa bề mặt mà còn phải học được biểu diễn ngữ nghĩa sâu hơn để phân biệt mục tiêu giao tiếp chính của người dùng. Do đó, tập dữ liệu đặc biệt phù hợp để đánh giá khả năng hiểu ngữ cảnh và suy luận ngữ nghĩa của các mô hình hiện đại, bao gồm các kiến trúc dựa trên Transformer và các mô hình ngôn ngữ lớn.

C. CHIA DỮ LIỆU DỮ LIỆU THỰC NGHIỆM

Dữ liệu được chia thành các tập huấn luyện (80%), kiểm định (10%) và kiểm tra (10%) theo chiến lược lấy mẫu phân tầng nhằm duy trì phân bố lớp nhất quán giữa các tập. Một tập kiểm tra cố định 10% mẫu được sử dụng xuyên suốt toàn bộ quá trình đánh giá cho phương pháp fine-tuning, đảm bảo rằng mọi phương pháp đều được so sánh trên cùng một tập dữ liệu đầu vào.

Ở phần thực nghiệm prompting chỉ được thực hiện trên 200 mẫu của 10% tập kiểm tra đã dùng cho phương pháp fine-tuning, điều này sẽ giúp cho việc đánh giá được sức mạnh của các kĩ thuật prompting nhưng cũng không quá nặng vấn đề về chi phí vận hành.

Cách chia này giúp giảm thiểu ảnh hưởng của yếu tố ngẫu nhiên trong quá trình huấn luyện và cho phép rút ra các kết luận so sánh có độ tin cậy cao hơn. Đồng thời, việc giữ nguyên giao thức thực nghiệm cho tất cả các mô hình cũng góp phần làm rõ vai trò của kiến trúc và phương pháp tiếp cận, thay vì các khác biệt do tiền xử lý hay chia dữ liệu.

D. HÀM Ý VÀ GIỚI HẠN CỦA TẬP DỮ LIỆU

Mặc dù tập dữ liệu cung cấp một nền tảng phù hợp để so sánh các phương pháp phát hiện ý định, vẫn tồn tại một số giới hạn cần được lưu ý. Thứ nhất, các câu văn bản chủ yếu là phát ngôn đơn lẻ, không phản ánh đầy đủ ngữ cảnh hội thoại nhiều lượt thường gặp trong HRI. Thứ hai, phong cách ngôn ngữ mang tính mô tả có thể khiến một số mô hình đạt hiệu suất cao hơn so với khi triển khai trong môi trường hội thoại tự nhiên.

Tuy nhiên, chính những đặc điểm này cũng khiến tập dữ liệu trở thành một bài toán chuẩn hóa hữu ích, cho phép đánh giá tương đối công bằng khả năng biểu diễn ngữ nghĩa và phân loại của các mô hình khác nhau trước khi mở rộng sang các kịch bản phức tạp hơn.

E. PHƯƠNG PHÁP ĐÁNH GIÁ

Trong bài toán phân loại ý định, bốn thang đo Accuracy, Precision, Recall và F1-score được dùng để phản ánh hiệu năng ở các góc nhìn khác nhau. Accuracy đo tỷ lệ dự đoán đúng trên toàn bộ mẫu, vì vậy trực quan nhưng có thể "đẹp giả" khi dữ liệu lệch lớp hoặc khi một số lớp dễ hơn các lớp còn lại. Để nhìn sâu hơn theo từng lớp, ta xét Precision và Recall theo cách "mỗi lớp so với phần còn lại" (one-vs-rest), với các khái niệm: TP (đúng dương) là số mẫu thuộc lớp đó và dự đoán đúng; FP (dương giả) là số mẫu không thuộc lớp đó nhưng bị dự đoán nhầm vào lớp đó; FN (âm giả) là số mẫu thuộc lớp đó nhưng bị dự đoán sang lớp khác. Khi đó, precision phản ánh "mỗi khi mô hình dự đoán vào một lớp thì đáng tin đến đâu" (ít mẫu âm đúng), còn recall phản ánh "mô hình có bỏ sót nhiều mẫu của lớp đó không" (ít FN). F1-score là trung bình điều hòa giữa precision và recall, giúp cân bằng hai xu hướng "dự đoán quá tay" (Precision thấp) và "bỏ sót" (Recall thấp). Trong nghiên cứu này, các chỉ số được báo cáo ở dạng macro bằng cách tính theo từng lớp rồi lấy trung bình, nhằm đảm bảo mỗi lớp được xem trọng như nhau.

1. ACCURACY

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. PRECISION

$$Precision = \frac{TP}{TP + FP}$$

3. RECALL

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. F1-SCORE

$$F1\_score = 2^{*}\frac{Precision * Recall}{Precision + Recall}$$

IV. CÁC PHƯƠNG PHÁP THỰC NGHIỆM

A. HỌC MÁY TRUYỀN THỐNG (TRADITIONAL MACHINE LEARNING)

1. MULTINOMIAL NAIVE BAYES KẾT HỢP TF-IDF

Phương pháp Multinomial Naive Bayes kết hợp với biểu diễn TF-IDF được sử dụng như một baseline tiêu chuẩn cho bài toán phân loại văn bản. Mô hình này dựa trên giả định độc lập điều kiện giữa các đặc trưng, cho phép ước lượng xác suất một cách hiệu quả ngay cả khi số lượng đặc trưng lớn. Mặc dù giả định này không hoàn toàn phù hợp với ngôn ngữ tự nhiên, phương pháp vẫn thường cho kết quả ổn định trong các tập dữ liệu văn bản ngắn, đồng thời có ưu điểm về tốc độ huấn luyện và khả năng diễn giải.

Siêu tham số trong thực nghiệm: *ngram_range (1,2), max_features 50k, min_df=2, alpha=0.5.*

B. HỌC SÂU (DEEP LEARNING)

1. BIDIRECTIONAL LSTM

Mô hình Bidirectional LSTM được áp dụng nhằm khai thác thông tin ngữ cảnh theo cả hai chiều của chuỗi văn bản. Bằng cách xử lý chuỗi từ trái sang phải và từ phải sang trái, mô hình có khả năng nắm bắt các phụ thuộc dài hạn mà các phương pháp dựa trên bag-of-words không thể biểu diễn. Tuy nhiên, hiệu quả của mô hình phụ thuộc đáng kể vào kích thước dữ liệu và việc điều chỉnh các siêu tham số.

Siêu tham số trong thực nghiệm: *embedding_dim=128; hidden_dim=128; num_layers=1; bidirectional; dropout=0.3; max_len=64; vocab_size≤50,000; min_word_freq=2; batch_size=32; lr=1e-3; optimizer=Adam; loss=CrossEntropyLoss; epochs=10; early stopping patience=3.*

2. TEXTCNN

TextCNN sử dụng các bộ lọc convolution với nhiều kích thước khác nhau để trích xuất các mẫu n-gram cục bộ trong văn bản. Cách tiếp cận này đặc biệt phù hợp với các câu ngắn, nơi các cụm từ ngắn mang nhiều thông tin ngữ nghĩa quan trọng. So với các mô hình tuần tự, TextCNN thường có cấu trúc đơn giản hơn và thời gian huấn luyện ngắn hơn, trong khi vẫn duy trì khả năng trích xuất đặc trưng hiệu quả.

Siêu tham số trong thực nghiệm: *embedding_dim=128; kernel_sizes=(3,4,5); num_filters=100 mỗi kernel; dropout=0.5; max_len=64; vocab_size≤50,000; min_word_freq=2; batch_size=32; lr=1e-3; optimizer=Adam; loss=CrossEntropyLoss; epochs=10; early stopping patience=3.*

C. TINH CHỈNH MÔ HÌNH TRANSFORMER (FINE-TUNING TRANSFORMERS)

1. BERT-LARGE

BERT-large được fine-tune cho nhiệm vụ phân loại chuỗi nhằm tận dụng các biểu diễn ngữ cảnh hai chiều đã được học trong giai đoạn tiền huấn luyện. Khả năng mô hình hóa mối quan hệ phức tạp giữa các từ giúp mô hình này đặc biệt hiệu quả trong việc phân biệt các ý định có nội dung ngữ nghĩa gần nhau.

Siêu tham số trong thực nghiệm: *pretrained=bert-large-uncased; max_len=128; batch_size=8; lr=2e-5; optimizer=AdamW; scheduler=linear warmup (warmup_steps=0) → linear decay; loss=CrossEntropyLoss (BertForSequenceClassification); epochs=10; early stopping patience=3*.

2. ROBERTA

RoBERT là một biến thể cải tiến của BERT với chiến lược huấn luyện được tối ưu hóa, bao gồm việc loại bỏ một số ràng buộc trong quá trình tiền huấn luyện và sử dụng "dynamic masking". Những cải tiến này giúp mô hình học được các biểu diễn ngữ nghĩa mạnh mẽ và ổn định hơn trong quá trình fine-tuning.

Siêu tham số trong thực nghiệm: *pretrained=roberta-base; max_len=128; batch_size=16; lr=2e-5; optimizer=AdamW; scheduler=linear warmup (warmup_steps=0) → linear decay; loss=CrossEntropyLoss; epochs=10; early stopping patience=3*.

D. TINH CHỈNH MÔ HÌNH NGÔN NGỮ LỚN (INSTRUCTION TUNING LLMS)

Ngoài các mô hình Encoder truyền thống, chúng tôi thực nghiệm tinh chỉnh mô hình sinh văn bản Meta-Llama-3.1-8B-Instruct.

Kỹ thuật: Sử dụng Low-Rank Adaptation (LoRA) - một kỹ thuật tinh chỉnh hiệu quả tham số (PEFT). Thay vì cập nhật toàn bộ 8 tỷ tham số, LoRA đóng băng trọng số mô hình gốc và chỉ huấn luyện các ma trận thích ứng hạng thấp (low-rank matrices) được chèn vào các lớp attention (Query, Value).

Mục tiêu: Mô hình được huấn luyện để sinh ra nhãn ý định trực tiếp dựa trên câu lệnh đầu vào, tận dụng khả năng hiểu ngôn ngữ và tuân theo chỉ dẫn (instruction following) vượt trội của dòng Llama-3.1.

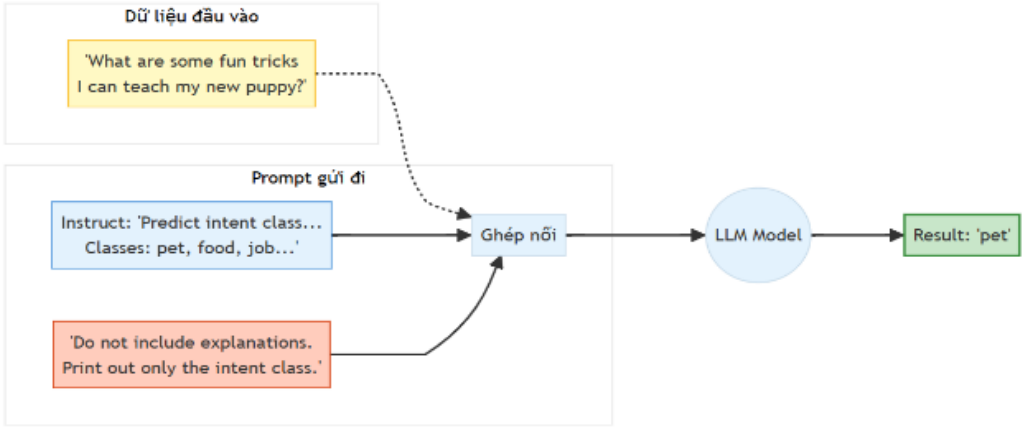
Siêu tham số trong thực nghiệm: *pretrained=meta-llama/Llama-3.1-8B-Instruct; max_len=256; batch_size=4; lr=2e-4; optimizer=AdamW; weight_decay=0.01; warmup_ratio=0.1; max_grad_norm=1.0; precision=bf16; loss=CrossEntropyLoss; epochs=3; LoRA r=16, alpha=32, dropout=0.1*.

E. LARGE LANGUAGE MODELS VỚI PROMPTING (IN-CONTEXT LEARNING)

1. ZERO-SHOT STANDARD

Đây là phương pháp prompting cơ bản nhất trong thiết lập zero-shot, trong đó mô hình được cung cấp mô tả tác vụ phân loại mà không kèm theo bất kỳ ví dụ minh họa nào.

Đặc điểm kỹ thuật: Xác định rõ ràng tác vụ và các lớp ý định, yêu cầu đầu ra ngắn gọn, không kèm giải thích, đánh giá khả năng khái quát hóa của mô hình dựa trên kiến thức tiền huấn luyện.

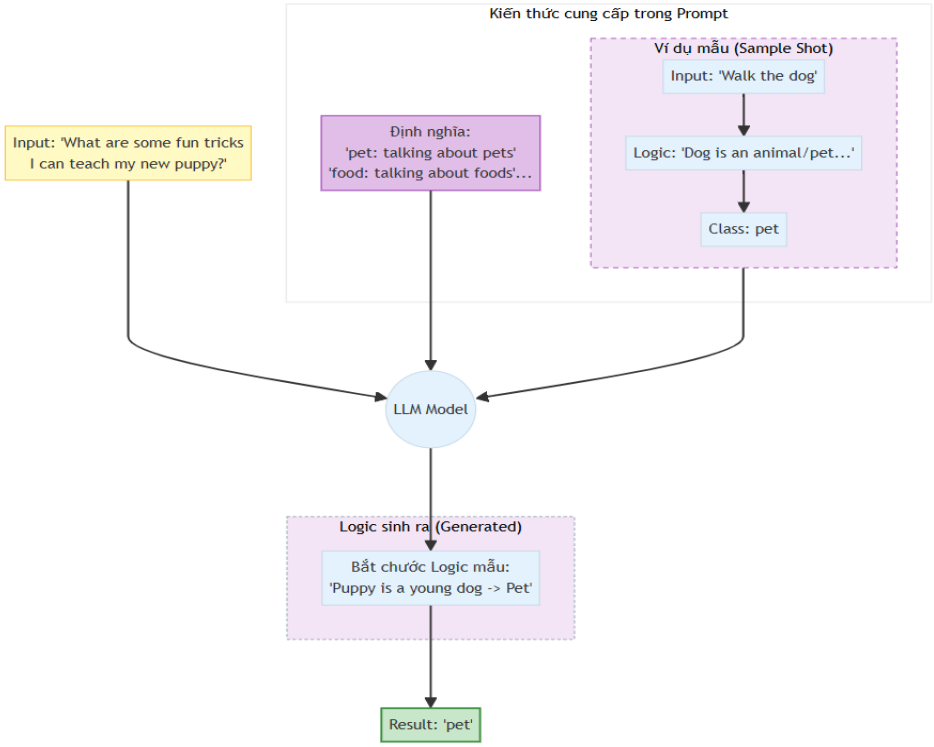


Hình 2. *Chuỗi hoạt động của zero-shot*

2. FEW-SHOT CHAIN-OF-THOUGHT

Kỹ thuật này được đề xuất trong nghiên cứu "*Chain-of-thought prompting elicits reasoning in large language models*" của Wei và cộng sự [9]. Mô hình được cung cấp mô tả chi tiết về ngữ nghĩa của từng lớp ý định, giúp định hướng quá trình suy luận theo chuỗi (chain-of-thought).

Đặc điểm kỹ thuật: Bổ sung định nghĩa ngữ nghĩa cho mỗi lớp (pet: liên quan đến vật nuôi, food: liên quan đến thực phẩm, v.v.), cung cấp quy tắc xử lý trường hợp biên (câu không thuộc lớp nào được gán nhãn other), tăng cường khả năng phân biệt giữa các lớp ý định tương tự.

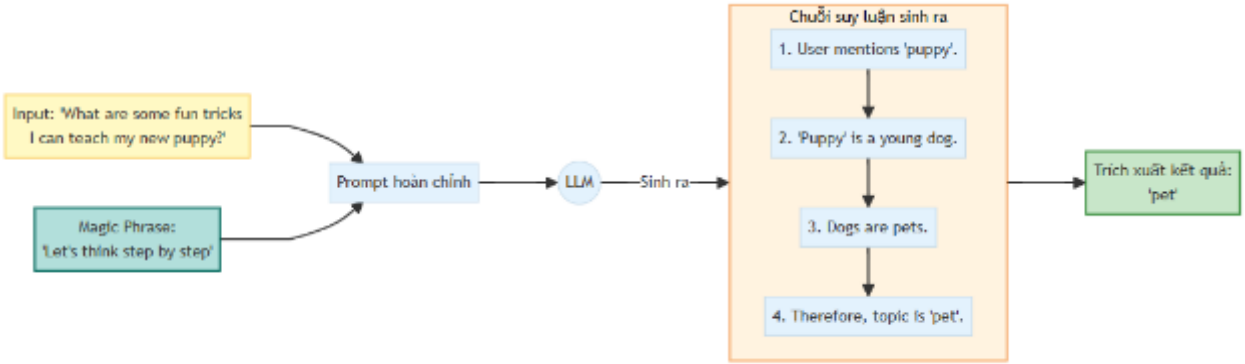


Hình 3. *Chuỗi hoạt động của few-shot chain-of-thought*

3. ZERO-SHOT CHAIN-OF-THOUGHT

Phương pháp này được giới thiệu trong nghiên cứu "Large language models are zero-shot reasoners" của Kojima và cộng sự [10]. Kĩ thuật này giúp kích hoạt khả năng suy luận của mô hình thông qua câu khởi động đặc trưng: "Let's think step by step".

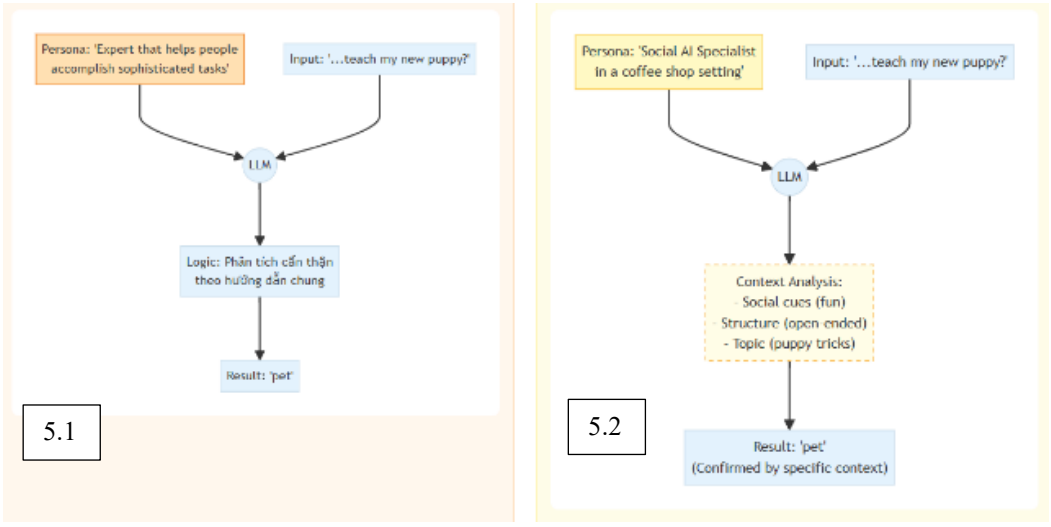
Đặc điểm kỹ thuật: Không yêu cầu ví dụ minh họa (zero-shot), khuyến khích mô hình tiến hành phân tích từng bước trước khi đưa ra quyết định, cải thiện độ chính xác trong các tác vụ yêu cầu suy luận logic.



Hình 4. *Chuỗi hoạt động của zero-shot chain-of-thought*

4. EXPERTPROMPTING

Kĩ thuật ExpertPrompting đã được đề xuất trong nghiên cứu "ExpertPrompting: Instructing large language models to be distinguished experts" [11]. Dự án triển khai hai biến thể của ExpertPrompting:



Hình 5. *Chuỗi hoạt động của expertprompting*

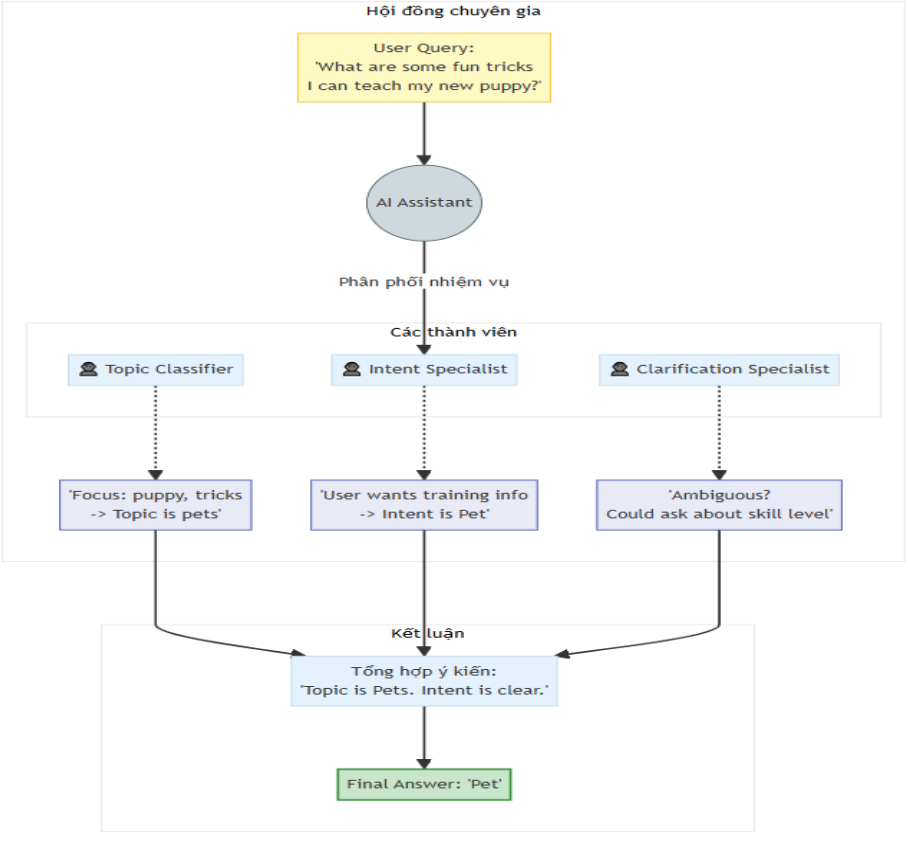
a) Expert-General (Hình 5.1)

Mô hình được gán vai trò như một chuyên gia tổng quát có khả năng hỗ trợ người dùng hoàn thành các tác vụ phức tạp. Đặc điểm: Kích hoạt kiến thức chuyên môn tổng quát của mô hình, yêu cầu phân tích cẩn thận để đưa ra phản hồi chính xác nhất.

b) Expert-Specific (Hình 5.2)

Mô hình được gán vai trò như một chuyên gia AI xã hội (Social AI Specialist) đang thiết kế AI hội thoại cho robot xã hội trong môi trường quán cà phê. Định nghĩa ngữ cảnh ứng dụng cụ thể (social robot, coffee shop setting), yêu cầu xem xét các yếu tố ngữ dụng học: cấu trúc câu, tín hiệu xã hội, mức độ trang trọng, tích hợp khả năng sử dụng tín hiệu phi ngôn ngữ và lịch sử tương tác (nếu có).

5. MULTI-PERSONA PROMPTING



Hình 6. Chuỗi hoạt động của multi-persona prompting

Kĩ thuật Multi-Persona Prompting được giới thiệu trong nghiên cứu "Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration" [12]. Phương pháp này thiết lập nhiều persona chuyên biệt cộng tác để phân tích và phân loại ý định người dùng.

V. KẾT QUẢ THỰC NGHIỆM

A. KẾT QUẢ CỦA CÁC KĨ THUẬT FINE-TUNING

Bảng 1. *Hiệu suất các mô hình qua phương pháp fine-tuning*

Tên Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted Precision	Weighted Recall	Weighted F1
NB + TF-IDF	0.8664	0.8959	0.8486	0.8656	0.8801	0.8664	0.8665
LSTM	0.6738	0.6533	0.6469	0.6462	0.6851	0.6738	0.6760
CNN Text	0.8180	0.8135	0.8074	0.8094	0.8175	0.8180	0.8167
BERT Large	0.9882	0.9880	0.9883	0.9881	0.9883	0.9882	0.9882
RoBERTa	0.9764	0.9747	0.9739	0.9743	0.9764	0.9764	0.9763
LlaMA 3.1	0.9882	0.9872	0.9883	0.9877	0.9884	0.9882	0.9882

Bảng 1 trình bày hiệu suất của các mô hình phát hiện ý định thuộc nhiều nhóm phương pháp khác nhau, từ học máy truyền thống, học sâu đến các mô hình Transformer và mô hình ngôn ngữ lớn được tinh chỉnh, được đánh giá trên cùng thiết lập dữ liệu. Nhìn chung, kết quả cho thấy sự phân hóa rõ rệt giữa nhóm mô hình dựa trên đặc trưng bề mặt và nhóm mô hình học biểu diễn ngữ nghĩa, đồng thời phản ánh tác động của kiến trúc mô hình đối với dữ liệu câu ngắn và ranh giới lớp tương đối rõ.

Trước hết, mô hình Multinomial Naive Bayes kết hợp TF–IDF đạt Accuracy 0.8664 và Macro F1 0.8656, cho thấy hiệu suất khá tốt dù dựa trên giả định đơn giản và không mô hình hóa ngữ cảnh sâu. Kết quả này phù hợp với đặc thù tập dữ liệu gồm các câu ngắn, nơi các từ khóa và cụm từ mang tính chỉ báo mạnh cho ý định. Do đó, NB + TF–IDF vẫn là một baseline đáng tin cậy trong các kịch bản cần triển khai gọn nhẹ, chi phí tính toán thấp và độ trễ nhỏ.

Đối với nhóm học sâu, LSTM cho kết quả thấp nhất (Accuracy 0.6738, Macro F1 0.6462), cho thấy mô hình tuần tự không tận dụng tốt lợi thế trong bối cảnh câu ngắn và tín hiệu phân biệt chủ yếu nằm ở các cụm từ ngắn. Ngoài ra, khi quy mô dữ liệu và độ đa dạng ngữ nghĩa chưa đủ lớn, LSTM dễ gặp khó khăn trong việc học biểu diễn ổn định so với các kiến trúc tiền huấn luyện. TextCNN cải thiện đáng kể so với LSTM (Accuracy 0.8180, Macro F1 0.8094), cho thấy các bộ lọc tích chập với nhiều kích thước kernel có khả năng trích xuất hiệu quả các mẫu n-gram và cụm từ đặc trưng cho ý định, phù hợp với bản chất của bài toán phân loại ý định.

Bước nhảy vọt về hiệu suất thể hiện rõ khi chuyển sang các mô hình tiền huấn luyện dựa trên Transformer. BERT-large đạt Accuracy 0.9882 và Macro F1 0.9881, thể hiện khả năng nắm bắt ngữ cảnh hai chiều và quan hệ ngữ nghĩa tốt, giúp phân biệt ổn định giữa các lớp. RoBERTa-base cũng đạt kết quả rất cao (Accuracy 0.9764, Macro F1 0.9743) nhưng thấp hơn BERT-large trong thiết lập này. Sự khác biệt này có thể đến từ chênh lệch về quy mô mô hình và cấu hình huấn luyện, khi BERT-large có số tham số lớn hơn và có thể phù hợp hơn với nhiệm vụ phân loại có ranh giới lớp tương đối rõ.

Đáng chú ý, mô hình meta-llama/Llama-3.1-8B-Instruct được tinh chỉnh bằng LoRA đạt Accuracy 0.9882 và Macro F1 0.9877, tiệm cận BERT-large và đứng trong nhóm cao nhất. Điều này cho thấy các kỹ thuật tinh chỉnh hiệu quả tham số như LoRA có thể khai thác tốt tri thức tiền huấn luyện của LLM để giải quyết bài toán phân loại, đồng thời giảm đáng kể chi phí huấn luyện so với fine-tuning toàn bộ mô hình. Kết quả cũng gợi ý rằng, trong bối cảnh có hạ tầng phù hợp, LLM tinh chỉnh nhẹ có thể là một lựa chọn cạnh tranh về hiệu năng.

Tổng thể, các kết quả cho thấy xu hướng chung là hiệu suất tăng khi chuyển từ các phương pháp dựa trên đặc trưng bề mặt sang các mô hình học biểu diễn ngữ nghĩa sâu. Tuy nhiên, lựa chọn mô hình cần cân nhắc đồng thời bối cảnh triển khai: NB + TF–IDF phù hợp cho hệ thống gọn nhẹ và chi phí thấp; TextCNN là phương án trung gian khi cần cải thiện hiệu năng nhưng vẫn giữ cấu trúc tương đối đơn giản; trong khi Transformer fine-tuning và LLM tinh chỉnh nhẹ cho hiệu suất cao nhất, phù hợp với các ứng dụng yêu cầu độ chính xác cao và có khả năng sử dụng tài nguyên tính toán mạnh.

B. KẾT QUẢ CỦA CÁC KĨ THUẬT PROMPTING (ACCURANCY)

Bảng 2. *Hiệu suất của các kĩ thuật prompting*

Các kĩ thuật prompting	Gemma	Claude 3 Opus	GPT-4 Turbo
Standard	39%	63%	96%
Few-shot	49%	99%	98%
Zero-shot	65%	100%	100%
Expert-general	77%	99%	100%
Expert-specific	68%	99%	98%
Multi-persona	76%	100%	99%

Bảng 2 trình bày hiệu suất của các kỹ thuật prompting khác nhau khi áp dụng cho ba Mô hình Ngôn ngữ Lớn đại diện là Gemma, Claude 3 Opus và GPT-4 Turbo. Không giống các mô hình fine-tuning được huấn luyện trực tiếp trên tập dữ liệu mục tiêu, các phương pháp prompting đánh giá khả năng của LLM, tức khả năng suy luận và phân loại ý định chỉ dựa trên tri thức đã được tiền huấn luyện và thông tin cung cấp trong prompt.

Trước hết, với kỹ thuật standard zero-shot prompting, sự khác biệt rõ rệt giữa các mô hình được quan sát. Gemma chỉ đạt 39% accuracy, cho thấy việc mô tả tác vụ đơn thuần là chưa đủ để mô hình nắm bắt chính xác ranh giới ngữ nghĩa giữa các lớp ý định. Claude 3 Opus cải thiện đáng kể với 63%, trong khi GPT-4 Turbo đạt tới 96%, phản ánh năng lực suy luận và tuân thủ chỉ dẫn vượt trội của các LLM ở quy mô lớn. Kết quả này nhất quán với các phân tích trước đó về fine-tuning, trong đó chất lượng biểu diễn ngữ nghĩa đóng vai trò quyết định đối với bài toán intent classification.

Khi bổ sung few-shot prompting, hiệu suất của Gemma tăng nhẹ lên 49%, cho thấy các ví dụ minh họa giúp mô hình định hình tốt hơn không gian quyết định, nhưng vẫn chưa đủ để đạt độ chính xác cao. Ngược lại, Claude 3 Opus và GPT-4 Turbo gần như đạt mức bão hòa (99% và 98%), cho thấy với các mô hình mạnh, chỉ cần một số ít ví dụ đã đủ để kích hoạt khả năng suy luận phù hợp cho tác vụ phân loại ý định. Điều này củng cố giả thuyết rằng năng lực few-shot learning của LLM phụ thuộc rất lớn vào quy mô và chất lượng tiền huấn luyện.

Đáng chú ý, kỹ thuật zero-shot prompting kết hợp Chain-of-Thought mang lại sự cải thiện đáng kể cho Gemma, với accuracy tăng lên 65%. Việc yêu cầu mô hình "suy nghĩ từng bước" giúp giảm lỗi do suy luận vội vàng và cải thiện khả năng phân biệt các ý định có nội dung ngữ nghĩa gần nhau. Với Claude 3 Opus và GPT-4 Turbo, phương pháp này đạt mức 100%, cho thấy trong các mô hình đủ mạnh, việc kích hoạt chuỗi suy luận có thể thay thế cho việc cung cấp ví dụ huấn luyện.

Các kỹ thuật Expert Prompting tiếp tục nâng cao hiệu suất, đặc biệt đối với Gemma. Phiên bản expert-general đạt 77%, cao nhất trong các cấu hình prompting của Gemma, cho thấy việc gán vai trò chuyên gia giúp mô hình huy động kiến thức nền một cách có cấu trúc hơn. Tuy nhiên, expert-specific không mang lại cải thiện tương ứng và thậm chí giảm nhẹ hiệu suất (68%). Hiện tượng này có thể được lý giải bởi việc ràng buộc mô hình vào một ngữ cảnh ứng dụng quá cụ thể (robot xã hội trong quán cà phê) có thể gây nhiễu khi tập dữ liệu đánh giá mang tính khái quát hơn.

Với multi-persona prompting, Gemma đạt 76%, cho thấy sự cộng tác giữa nhiều persona giúp mô hình tiếp cận bài toán từ nhiều góc nhìn, cải thiện khả năng phân tích so với zero-shot hoặc few-shot đơn lẻ. Đối với Claude 3 Opus và GPT-4 Turbo, phương pháp này tiếp tục duy trì hiệu suất rất cao (100% và 99%), nhưng mức cải thiện so với các kỹ thuật prompting khác là không đáng kể, cho thấy lợi ích cận biên của các chiến lược prompting phức tạp giảm dần khi mô hình đã đủ mạnh.

So sánh tổng thể với các kết quả fine-tuning, có thể nhận thấy rằng LLM prompting có tiềm năng đạt hiệu suất rất cao mà không cần huấn luyện lại mô hình, đặc biệt với các LLM hàng đầu như Claude 3 Opus và GPT-4 Turbo. Tuy nhiên, hiệu suất này phụ thuộc mạnh vào chất lượng mô hình nền và thiết kế prompt, đồng thời đi kèm với chi phí suy luận cao và độ trễ lớn hơn, những yếu tố quan trọng trong bối cảnh HRI thời gian thực. Ngược lại, các mô hình fine-tuning, đặc biệt là Transformer encoder hoặc LLM được tinh chỉnh bằng LoRA, mang lại sự cân bằng tốt hơn giữa hiệu suất, chi phí và khả năng triển khai lâu dài.

VI. KẾT LUẬN

Nghiên cứu này đã trình bày một phân tích so sánh toàn diện các nhóm phương pháp phổ biến cho bài toán phát hiện ý định trong tương tác người–máy, bao gồm các phương pháp học máy truyền thống, học sâu, mô hình Transformer được fine-tune, mô hình ngôn ngữ lớn được tinh chỉnh hiệu quả tham số, và các kỹ thuật prompting

dựa trên in-context learning. Thông qua việc đánh giá trên cùng một tập dữ liệu và thiết lập thực nghiệm thống nhất, nghiên cứu đã làm rõ những ưu điểm, hạn chế và sự đánh đổi giữa các cách tiếp cận khác nhau.

Kết quả thực nghiệm cho thấy không tồn tại một phương pháp duy nhất vượt trội trong mọi kịch bản triển khai. Các phương pháp học máy truyền thống như Multinomial Naive Bayes kết hợp TF-IDF, mặc dù đơn giản, vẫn đạt hiệu suất ổn định và có chi phí tính toán thấp, phù hợp với các hệ thống HRI nhúng hoặc các ứng dụng yêu cầu độ trễ nhỏ. Nhóm học sâu, với các mô hình như Bidirectional LSTM và TextCNN, cho thấy khả năng học biểu diễn tốt hơn các phương pháp truyền thống, nhưng hiệu quả phụ thuộc đáng kể vào đặc điểm dữ liệu và quy mô tập huấn luyện.

Các mô hình Transformer được fine-tune, đặc biệt là BERT-large và RoBERTa, đạt hiệu suất cao nhất trong nhóm mô hình phân loại truyền thống, khẳng định vai trò quan trọng của biểu diễn ngữ cảnh hai chiều trong việc phân biệt các ý định có nội dung ngữ nghĩa gần nhau. Tuy nhiên, việc fine-tune các mô hình này đòi hỏi tài nguyên tính toán đáng kể, đặc biệt là GPU, cùng với kiến thức chuyên môn về lập trình, xử lý dữ liệu và điều chỉnh siêu tham số. Điều này có thể trở thành rào cản đối với các nhóm triển khai nhỏ hoặc các hệ thống cần mở rộng nhanh.

Trong khi đó, mô hình ngôn ngữ lớn Meta-Llama-3.1-8B-Instruct được tinh chỉnh bằng LoRA cho thấy khả năng đạt hiệu suất tiệm cận các mô hình Transformer tốt nhất, đồng thời giảm đáng kể số lượng tham số cần huấn luyện. Cách tiếp cận này đại diện cho một hướng trung gian hiệu quả, kết hợp được ưu điểm của fine-tuning truyền thống và sức mạnh biểu diễn ngôn ngữ của LLM, nhưng vẫn yêu cầu hạ tầng GPU và quy trình huấn luyện tương đối phức tạp.

Đối với các kỹ thuật prompting, kết quả cho thấy các LLM mạnh như Claude 3 Opus và GPT-4 Turbo có thể đạt hiệu suất rất cao, thậm chí tiệm cận hoặc vượt các mô hình fine-tune, mà không cần huấn luyện lại. Ưu điểm lớn nhất của cách tiếp cận này là tính linh hoạt và tốc độ triển khai nhanh, cho phép xây dựng hệ thống phát hiện ý định chỉ thông qua thiết kế prompt. Tuy nhiên, hiệu quả này đi kèm với chi phí sử dụng dịch vụ cao, phụ thuộc vào mô hình độc quyền, và độ trễ suy luận lớn hơn, những yếu tố cần được cân nhắc kỹ trong các hệ thống HRI thời gian thực hoặc triển khai lâu dài.

Từ các kết quả trên, có thể thấy rằng việc lựa chọn phương pháp phát hiện ý định phù hợp phụ thuộc mạnh vào bối cảnh ứng dụng, tài nguyên tính toán, ngân sách và năng lực kỹ thuật của đội ngũ phát triển. Fine-tuning các mô hình học sâu và Transformer phù hợp với các hệ thống cần kiểm soát dữ liệu và chi phí vận hành lâu dài, trong khi prompting với LLM phù hợp với các kịch bản yêu cầu triển khai nhanh và độ linh hoạt cao. Do đó, thay vì tìm kiếm một giải pháp tối ưu duy nhất, nghiên cứu này nhấn mạnh tầm quan trọng của việc cân nhắc toàn diện các yếu tố kỹ thuật và thực tiễn khi thiết kế hệ thống phát hiện ý định trong HRI.

Trong tương lai, các hướng nghiên cứu tiếp theo có thể tập trung vào việc kết hợp linh hoạt giữa fine-tuning và prompting, khai thác ưu điểm của cả hai cách tiếp cận, cũng như mở rộng đánh giá trên các tập dữ liệu hội thoại nhiều lượt và các kịch bản HRI thực tế hơn nhằm nâng cao khả năng tổng quát hóa và tính ứng dụng của các mô hình.

VII. LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Trường Đại học Ngoại ngữ – Tin học Thành phố Hồ Chí Minh trong khuôn khổ Đề tài mã số H2024-04.

VIII. TÀI LIỆU THAM KHẢO

[1] G. Arora, S. Jain, and S. Merugu (2024). Intent Detection in the Age of LLMs, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, ACL, pp. 1559–1570.

[2] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao (2023). Large language models for human–robot interaction: A review, Biomimetic Intelligence and Robotics, Vol. 3, No. 4, p. 100131.

[3] X. Zhang and H. Wang (2016). A Joint Model of Intent Determination and Slot Filling, International Joint Conference on Artificial Intelligence, AAAI Press, pp. 2993–2999.

[4] F. Gervits, A. Leuski, C. Bonial, C. Gordon, and D. Traum (2019). A Classification-Based Approach to Automating Human-Robot Dialogue, International Workshop on Spoken Dialogue Systems Technology, Vol. 714, pp. 115–127. DOI: 10.1007/978-981-15-9323-9_10.

[5] K. V. Mathew, V. S. A. Tarigoppula, and L. Frermann (2021). Multi-modal Intent Classification for Assistive Robots with Large-scale Naturalistic Datasets, pp. 47–57. Accessed date: Jan. 10, 2026. Available: https://aclanthology.org/2021.alta-1.5/

[6] S. Y. (尹商鉴), P. H. (黄沛杰), and Y. X. (徐禹洪) (2025). MIDLM: Multi-Intent Detection with Bidirectional Large Language Models, pp. 2616–2625. Accessed date: Jan. 10, 2026. Available: https://aclanthology.org/2025.coling-main.179/

[7] T. Hong et al. (2024). Exploring the Use of Natural Language Descriptions of Intents for Large Language Models in Zero-shot Intent Classification, SIGDIAL 2024 - 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference, pp. 458–465. DOI: 10.18653/V1/2024.SIGDIAL-1.39.

[8] S. Parikh, P. Tumbade, Q. Vohra, and M. Tiwari (2023). Exploring Zero and Few-shot Techniques for Intent Classification, Proceedings of the Annual Meeting of the Association for Computational Linguistics, Vol. 5, pp. 744–751. DOI: 10.18653/v1/2023.acl-industry.71.

[9] J. Wei et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Advances in Neural Information Processing Systems, Vol. 35. Accessed date: Jan. 10, 2026. Available: https://arxiv.org/pdf/2201.11903

[10] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa (2022). Large Language Models are Zero-Shot Reasoners, Advances in Neural Information Processing Systems, Vol. 35. Accessed date: Jan. 10, 2026. Available: https://arxiv.org/pdf/2205.11916

[11] B. Xu et al. (2023). ExpertPrompting: Instructing Large Language Models to be Distinguished Experts. Accessed date: Jan. 10, 2026. Available: https://arxiv.org/pdf/2305.14688

[12] Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, and H. Ji (2024). Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024, Vol. 1, pp. 257–279. DOI: 10.18653/V1/2024.NAACL-LONG.15.

A MULTI-METHOD STUDY FOR INTENT DETECTION IN HUMAN–MACHINE INTERACTION

Thuy-A Nguyen, Huynh Ly Tan Khoa, Nguyen Minh Y, Tieu Phung Mai Suong

ABSTRACT— Intent detection is a core component of human–machine and human–robot interaction systems, determining the system's ability to correctly understand user requests and produce appropriate responses. This study conducts a systematic comparison of four groups of methods for intent classification: traditional machine learning, deep learning, Transformers, and large language models using prompting techniques. Experiments are carried out on an HRI dataset of 8,453 utterances with six intent labels. The data are split into 80% training, 10% validation, and 10% testing, while maintaining stratified class proportions across the splits. On the 10% test set, BERT-large achieves an accuracy of 0.9882 and a macro-F1 of 0.9881; Llama 3.1 8B with parameter-efficient fine-tuning achieves an accuracy of 0.9882 and a macro-F1 of 0.9877; RoBERTa-base achieves an accuracy of 0.9764 and a macro-F1 of 0.9743. Meanwhile, lighter models deliver lower performance but are easier to deploy, with Naive Bayes combined with TF–IDF reaching an accuracy of 0.8664, TextCNN achieving 0.8180, and BiLSTM achieving 0.6738. For the prompting-based large language model group, the study evaluates six prompting variants on a fixed subset of 200 samples from the test set to control inference costs, thereby clarifying the trade-offs among performance, computational cost, deployment complexity, and generalization. The results indicate that Transformer fine-tuning and lightweight adaptation training yield the best performance, whereas prompting is suitable when rapid deployment is needed or when leveraging zero-shot capabilities in scenarios with limited training data.

Keywords — Intent Classification, Human-Robot Interaction, Transformer, Large Language Models, Prompting

ThS. Nguyễn Thị Thúy A nhận bằng thạc sĩ ngành Khoa học máy tính tại Trường Đại học Khoa học tự nhiên, ĐHQG TP.HCM vào năm 2018. Hiện tại Thạc sĩ Thúy A đang là giảng viên tại khoa Công nghệ thông tin, trường Đại học Ngoại ngữ-Tin học TP. Hồ Chí Minh (HUFLIT).

Hướng nghiên cứu chính: Xử lý ngôn ngữ tự nhiên, Trí tuệ nhân tạo.

Huỳnh Lý Tân Khoa Hiện đang là sinh viên đại học năm 4 chuyên ngành Khoa học dữ liệu thuộc ngành Công nghệ thông tin tại Trường Đại học Ngoại ngữ - Tin học TP. Hồ Chí Minh (HUFLIT).

Hướng nghiên cứu chính: Học sâu, Khoa học dữ liệu, thị giác máy tính và Xử lý ngôn ngữ tự nhiên.

Tiếu Phùng Mai Sương là Thạc sĩ Khoa học máy tính tại Trường Đại học Khoa học tự nhiên, ĐHQG TP.HCM 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.

Nguyễn Minh Ý sinh viên năm 4 chuyên ngành Khoa học dữ liệu, ngành Công nghệ thông tin tại Trường Đại học Ngoại ngữ - Tin học TP Hồ Chí Minh (HUFLIT).
Hướng nghiên cứu chính: Trí tuệ Nhân tạo, Khai phá dữ liệu lớn, eLearning, Xử lý Ngôn ngữ tự nhiên và Hệ thống thông minh.