

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**BỘ NÔNG NGHIỆP VÀ MÔI  
TRƯỜNG**

**TRƯỜNG ĐẠI HỌC THỦY LỢI**



**CAO NGUYỄN VŨ**

**NGHIÊN CỨU VÀ XÂY DỰNG MÔ HÌNH CHATBOT  
TƯ VẤN NHA KHOA SỬ DỤNG MẠNG KẾT HỢP  
CHUYÊN GIA**

**ĐỒ ÁN TỐT NGHIỆP**

**HÀ NỘI, NĂM 2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO    BỘ NÔNG NGHIỆP VÀ MÔI  
TRƯỜNG**

**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**CAO NGUYỄN VŨ**

**NGHIÊN CỨU VÀ XÂY DỰNG MÔ HÌNH CHATBOT  
TƯ VẤN NHA KHOA SỬ DỤNG MẠNG KẾT HỢP  
CHUYÊN GIA**

Ngành: Trí tuệ nhân tạo và  
khoa học dữ liệu

Mã số: 3074

**NGƯỜI HƯỚNG DẪN    ThS. Trần Anh Đạt**

**HÀ NỘI, NĂM 2025**

## **LỜI CAM ĐOAN**

Tác giả xin cam đoan đây là Đồ án tốt nghiệp của bản thân tác giả. Các kết quả trong Đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào. Việc tham khảo các nguồn tài liệu (nếu có) đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

**Tác giả ĐATN**

**Cao Nguyên Vũ**

## LỜI CẢM ƠN

Đồ án tốt nghiệp là một cột mốc quan trọng trong quá trình học tập tại Đại học Thủy Lợi, nơi em có cơ hội thể hiện những kiến thức và kỹ năng mà em đã học được trong bốn năm qua. Trong quá trình thực hiện đồ án, em rất biết ơn sự hỗ trợ và đồng hành của nhiều người, đặc biệt là sự hướng dẫn từ các thầy cô giáo, sự ủng hộ từ gia đình, và sự chia sẻ kiến thức từ bạn bè. Những đóng góp và tận tâm của mọi người đã góp phần quan trọng vào thành công của đồ án này.

Trước hết, em xin gửi lời cảm ơn đến các thầy cô giáo khoa Công nghệ thông tin tại Trường Đại học Thủy Lợi, đã chia sẻ, truyền đạt những kiến thức quý báu trong suốt bốn năm học. Những kiến thức này đã đóng vai trò quan trọng, hỗ trợ em thành công trong việc thực hiện đồ án ngày hôm nay.

Đặc biệt, em cảm ơn thầy Trần Anh Đạt - người đã tận tình hướng dẫn em thực hiện đồ án về chủ đề “Nghiên cứu và xây dựng mô hình chatbot tư vấn nha khoa sử dụng mạng kết hợp chuyên gia”. Những lời khuyên, góp ý quý báu của thầy giúp em vượt qua nhiều khó khăn, hoàn thiện đồ án. Em xin cảm ơn thầy vì đã là người thầy tuyệt vời của em.

Em cũng xin cảm ơn gia đình, bạn bè và các bạn cùng lớp đã luôn ủng hộ, quan tâm và giúp đỡ em trong suốt quá trình học tập và làm đồ án. Nhờ có sự động viên và chia sẻ của mọi người, em đã vượt qua được những khó khăn và áp lực trong công việc.

Em xin thừa nhận rằng đồ án của mình vẫn còn khuyết điểm và hạn chế, chủ yếu do kiến thức hạn chế, thời gian có hạn và thiếu kinh nghiệm thực tế. Em mong nhận được sự thông cảm và góp ý từ các thầy cô giáo cũng như các bạn để có cơ hội cải thiện đồ án. Em xin chân thành cảm ơn tất cả những người đã hỗ trợ em suốt hành trình này.

**Em xin chân thành cảm ơn!**

*Sinh viên*

**Cao Nguyên Vũ**

## MỤC LỤC

DANH MỤC CÁC HÌNH ẢNH.....	V
DANH MỤC BẢNG BIỂU .....	VI
DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ .....	VII
MỞ ĐẦU .....	1
CHƯƠNG 1 : GIỚI THIỆU TỔNG QUAN BÀI TOÁN.....	3
1.1 Lý do chọn đề tài .....	3
1.1.1 Bối cảnh và nhu cầu phát triển chatbot tư vấn trong lĩnh vực nha khoa .....	3
1.1.2 Những hạn chế của các hệ thống hiện tại.....	4
1.1.3 Tính cấp thiết và ý nghĩa ứng dụng.....	5
1.2 Mục tiêu nghiên cứu .....	6
1.2.1 Mục tiêu tổng quát.....	6
1.2.2 Mục tiêu cụ thể.....	7
1.3 Đối tượng nghiên cứu và phạm vi nghiên cứu .....	8
1.3.1 Đối tượng nghiên cứu.....	8
1.3.2 Phạm vi nghiên cứu.....	9
1.4 Câu hỏi nghiên cứu và giả thuyết .....	10
1.4.1 Các vấn đề cần giải quyết.....	10
1.4.2 Giả thuyết nghiên cứu .....	11
1.5 Phương pháp và công cụ nghiên cứu.....	12
1.5.1 Phương pháp nghiên cứu.....	12
1.5.2 Công cụ nghiên cứu.....	13
1.6 Cấu trúc đề tài và dàn ý các chương.....	15
CHƯƠNG 2 : KIẾN THỨC NỀN TẢNG VÀ MÔ HÌNH DEEPSEEK.....	17
2.1 Tổng quan kiến thức nền tảng .....	17
2.1.1 Các khái niệm cơ bản về trí tuệ nhân tạo, Transformer và ứng dụng trong chatbot .....	17
2.1.2 Tổng quan về mạng kết hợp chuyên gia.....	20
2.2 Giới thiệu mô hình DeepSeek.....	21
2.2.1 Lịch sử phát triển và các phiên bản của DeepSeek .....	22
2.2.2 Kiến trúc và các thành phần chính .....	23

2.2.3 Các cải tiến về hiệu năng.....	25
2.3 Cơ chế hoạt động và ưu nhược điểm .....	26
2.4 Ứng dụng của DeepSeek trong các bài toán NLP .....	28
2.5 Phương pháp SFT và OPRO.....	29
CHƯƠNG 3 : THU THẬP, XỬ LÝ DỮ LIỆU NHA KHOA.....	31
3.1 Giới thiệu nguồn và đặc thù dữ liệu nha khoa .....	31
3.2 Phương pháp thu thập dữ liệu.....	32
3.3 Tiền xử lý dữ liệu .....	34
3.4 Prompt Engineering - giai đoạn tiền xử lý.....	39
3.5 Tăng cường dữ liệu.....	40
3.6 Phân tích dữ liệu và đánh giá chất lượng bộ dữ liệu nha khoa.....	41
3.6.1 Tổng quan về tập dữ liệu.....	41
3.6.2 Phân tích định lượng dữ liệu .....	42
3.6.3 Đánh giá chất lượng dữ liệu .....	44
CHƯƠNG 4 : TÍNH CHỈNH VÀ HUẤN LUYỆN .....	47
4.1 Tổng quan về quy trình huấn luyện và tính chỉnh .....	47
4.2 Phương pháp fine-tuning mô hình .....	48
4.3 Quá trình chuẩn bị dữ liệu huấn luyện.....	50
4.4 Prompt Engineering - giai đoạn huấn luyện .....	50
4.5 Thiết lập thí nghiệm fine-tuning với học có giám sát.....	51
4.6 Cải thiện khả năng phản hồi và suy luận với RLHF.....	55
4.7 Phân tích kết quả thí nghiệm .....	56
4.8 Đánh giá và thảo luận .....	58
CHƯƠNG 5 : TRIỂN KHAI VÀ ỨNG DỤNG THỰC TẾ .....	67
5.1 Kiến trúc hệ thống triển khai chatbot .....	67
5.2 Các tính năng .....	68
5.3 Triển khai thực nghiệm và tích hợp thực tế.....	75
5.4 Đánh giá chatbot.....	77
5.5 Hướng phát triển và mở rộng.....	78
KẾT LUẬN .....	80
TÀI LIỆU THAM KHẢO .....	82

## DANH MỤC CÁC HÌNH ẢNH

Hình 2.1 Mô tả quá trình hoạt động của Transformer.....	18
Hình 2.2 Mô tả cơ chế Attention .....	20
Hình 2.3 Mô tả mạng kết hợp chuyên gia .....	21
Hình 2.4 Kiến trúc DeepSeeK.....	23
Hình 3.1 So sánh IO prompt với CoT prompt.....	35
Hình 3.2 Các câu hỏi thường được dùng trong 500.000 dữ liệu đầu tiên .....	44
Hình 4.1 Mô tả quá trình huấn luyện.....	48
Hình 4.2 Mô tả cách hoạt động kỹ thuật Qlora .....	49
Hình 4.3 Diễn biến các chỉ số ngữ nghĩa trong 5000 bước huấn luyện .....	57
Hình 4.4 Các độ đo ROUGE theo 5000 bước huấn luyện .....	58
Hình 5.1 Tổng quan kiến trúc hoạt động hệ thống chatbot DentalGPT .....	67
Hình 5.2 Mô tả cách chức năng suy luận, suy luận sâu hoạt động.....	72
Hình 5.3 Mô tả cách chức năng RAG Search hoạt động.....	73
Hình 5.4 Mô tả cách cách chức năng RAG Search hoạt động .....	74

## DANH MỤC BẢNG BIỂU

Bảng 1.1 Mô hình sử dụng và tối ưu cho nghiên cứu .....	14
Bảng 1.2 Dịch vụ sử dụng trong nghiên cứu.....	15
Bảng 1.3 Thiết bị dùng để nghiên cứu .....	15
Bảng 3.1 Mô tả cách hoạt động chuỗi suy luận.....	37
Bảng 3.2 Mô tả vai trò trong cuộc hội thoại.....	40
Bảng 3.3 Đánh giá định lượng dữ liệu huấn luyện.....	43
Bảng 3.4 Đánh giá chất lượng dữ liệu huấn luyện .....	45
Bảng 4.1 Liệt kê một số tham số cấu hình cho Mô hình DeepSeek.....	54
Bảng 4.2 Liệt kê một số tham số cấu hình cho huấn luyện có giám sát.....	55
Bảng 4.3 Liệt kê một số tham số cấu hình cho huấn luyện học tăng cường .....	56
Bảng 4.4 Kết quả đánh giá mô hình sau khi huấn luyện .....	61
Bảng 4.5 Kết quả chấm điểm trung bình của năm nhóm đối tượng.....	61
Bảng 4.6 Kết quả chấm điểm sau khi làm bài kiểm tra .....	62
Bảng 4.7 So sánh giữa DentalGPT và các mô hình trong bài báo gốc [5].....	64
Bảng 4.8 So sánh giữa DentalGPT và các mô hình tương tự và suy luận khác [5].	65
Bảng 4.9 So sánh giữa DentalGPT và các mô hình tăng cường cho suy luận [5]....	66
Bảng 5.1 Trả lời thường khi đã có trước chuỗi suy luận cố định. ....	70
Bảng 5.2 Trả lời với chuỗi suy luận. ....	71



## DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ

Thuật ngữ	Nghĩa tiếng anh	Nghĩa tiếng việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
DL	Deep Learning	Học sâu
KV	Key-Value	bộ nhớ Khóa - Giá trị trong kiến trúc attention
LLM	Large Language Model	Mô hình ngôn ngữ lớn
LoRA	Low-Rank Adaptation	Thích ứng cấp thấp
MLA	Multi-head Latent Attention	Cơ chế chú ý tiềm ẩn đa đầu
ML	Machine Learning	Học máy
MoE	Mixture of Experts	Mạng mô hình kết hợp các chuyên gia
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
ORPO	Odds Ratio Preference Optimization	Tối ưu hóa ưu tiên Tỷ lệ Cược
QLoRA	Quantized Low-Rank Adaptation	Thích ứng cấp thấp lượng tử hóa
RHLF	Reinforcement Learning with Human Feedback	Học tăng cường từ phản hồi con người
RoPE	Rotary Position Embedding	Nhúng vị trí xoay
SFT	Supervised Fine-Tuning	Tinh chỉnh có giám sát
TRL	Transformers Reinforcement Learning library	Thư viện Học tăng cường cho Transformers

## MỞ ĐẦU

Trong những năm gần đây, sự phát triển nhanh chóng của các mô hình ngôn ngữ lớn đã mở ra nhiều hướng đi mới trong việc xây dựng các hệ thống hội thoại tự động, đặc biệt là trong lĩnh vực y tế. Tuy nhiên, các ứng dụng hiện có vẫn còn nhiều hạn chế khi triển khai cho những chuyên ngành đặc thù như nha khoa, nơi đòi hỏi mức độ chính xác cao, khả năng hiểu ngữ cảnh chuyên sâu và tương tác bằng ngôn ngữ tiếng Việt. Trên cơ sở đó, đề tài được thực hiện nhằm đề xuất một giải pháp chatbot tư vấn nha khoa có khả năng phản hồi chính xác, phù hợp với ngữ cảnh và đảm bảo tính chuyên môn trong nội dung giao tiếp. Đề tài được đặt trong **bối cảnh** nhu cầu tiếp cận thông tin y tế ngày càng tăng, trong khi hệ thống tư vấn truyền thống chưa đáp ứng hiệu quả ở cả khía cạnh quy mô và tính cá nhân hóa. Việc xây dựng một hệ thống hội thoại ứng dụng trí tuệ nhân tạo có khả năng hoạt động như một trợ lý ảo chuyên biệt trong lĩnh vực nha khoa sẽ góp phần nâng cao chất lượng dịch vụ y tế, đặc biệt trong xu thế chuyển đổi số hiện nay.

**Mục tiêu** xuyên suốt của nghiên cứu là thiết kế và triển khai một mô hình chatbot có thể vận hành hiệu quả trong môi trường hội thoại tự nhiên, hiểu và xử lý được ngôn ngữ chuyên ngành nha khoa, từ đó đưa ra phản hồi phù hợp với từng trường hợp cụ thể. Để hiện thực hóa mục tiêu này, đề tài tập trung vào các nhiệm vụ chính như xây dựng dữ liệu huấn luyện phù hợp, áp dụng các kỹ thuật tinh chỉnh hiện đại cho mô hình ngôn ngữ lớn, và triển khai mô hình trong môi trường thử nghiệm nhằm đánh giá khả năng phản hồi trong thực tế. **Phạm vi nghiên cứu** của đề tài giới hạn trong các hội thoại liên quan đến tư vấn sức khỏe răng miệng bằng tiếng Việt dưới dạng văn bản. Các chức năng tương tác đa phương thức hoặc triển khai thực tế tại cơ sở y tế sẽ được xem xét trong các hướng phát triển tiếp theo. Về mặt **phương pháp**, nghiên cứu lựa chọn mô hình DeepSeek-R1 làm nền tảng, kết hợp với kiến trúc mạng chuyên gia nhằm tối ưu hóa khả năng suy luận và cá nhân hóa phản hồi. Quá trình huấn luyện mô hình được tiến hành theo hai giai đoạn chính: huấn luyện có giám sát từ dữ liệu gán nhãn và điều chỉnh phản hồi thông qua học tăng cường từ tương tác. Đồ án **được tổ chức thành năm chương**, trong đó trình bày lần lượt tổng quan bài toán, các khái niệm nền tảng, phương pháp triển khai, kết quả thực nghiệm và các định hướng phát triển tiếp theo. Cấu trúc này nhằm đảm bảo tính logic, thống nhất và hỗ trợ việc theo dõi nội dung một cách rõ ràng, mạch lạc.

## CHƯƠNG 1 : GIỚI THIỆU TỔNG QUAN BÀI TOÁN

Chương này đặt nền tảng cho toàn bộ đề tài bằng cách nêu bật bối cảnh ứng dụng AI trong lĩnh vực nha khoa, chỉ ra các hạn chế của những hệ thống hiện tại, và từ đó hình thành nhu cầu xây dựng một hệ thống chatbot chuyên ngành nha khoa. Nội dung bao gồm lý do chọn đề tài, mục tiêu nghiên cứu, phạm vi, phương pháp và cấu trúc đồ án.

### 1.1 Lý do chọn đề tài

#### 1.1.1 Bối cảnh và nhu cầu phát triển chatbot tư vấn trong lĩnh vực nha khoa

Trong những năm gần đây, sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), các mô hình xử lý ngôn ngữ tự nhiên (NLP), hay mô hình ngôn ngữ lớn (LLM) đã mở ra nhiều cơ hội mới trong lĩnh vực chăm sóc sức khỏe hay nha khoa. Đặc biệt, chatbot AI ngày càng được ứng dụng và sử dụng rộng rãi không chỉ trong các hệ thống hỗ trợ y tế mà còn giúp cung cấp thông tin nhanh chóng, tiết kiệm thời gian và nâng cao trải nghiệm của bệnh nhân. Hơn thế nữa là các AI Agent đã có thể lập kế hoạch, tự động đưa ra các phương án điều trị cho bệnh nhân hay hỗ trợ các chuyên gia trong y tế nghiên cứu về học thuật. Cho tới hiện nay đã có các công ty lớn như FPT AI đã ứng dụng chatbot vào lĩnh vực y tế có khả năng phân tích triệu chứng do người dùng cung cấp và đưa ra lời khuyên phù hợp, từ hướng dẫn chăm sóc tại nhà đến khuyến nghị khám bệnh chuyên sâu. Theo dự đoán của Juniper Research, tỷ lệ thành công của chatbot trong lĩnh vực y tế sẽ tăng từ 12%, lên hơn 75% ,vào năm 2023 [1]. Hay trong cuộc chiến chống dịch Covid-19, FPT.AI đã phối hợp cùng Sở Y tế Hồ Chí Minh xây dựng chatbot cung cấp thông tin chính thống từ Bộ Y tế, hướng dẫn chăm sóc sức khỏe F0 cách ly tại nhà, giúp giảm tải cho hệ thống y tế và nâng cao hiệu quả điều trị [1].Ngoài ra, còn có Trợ lý y tế ảo DrAid™ của VinBrain giúp tư vấn 24/7, tự động tạo báo cáo y tế, tóm tắt và trích xuất thông tin bệnh án thức thì [2]. Nhìn chung, những ứng dụng trên của FPT và VinBrain cho thấy chatbot đang đóng vai trò quan trọng trong việc nâng cao chất lượng chăm sóc sức khỏe, tối ưu hóa quy trình làm việc và cải thiện trải nghiệm của bệnh nhân.

Trong lĩnh vực nha khoa, nhu cầu tư vấn và hỗ trợ bệnh nhân từ xa ngày càng trở nên cấp thiết trong đời sống sinh hoạt và công việc. Các bệnh lý răng miệng phổ biến như sâu răng, viêm lợi, viêm nha chu hay các vấn đề về niềng răng đòi hỏi bệnh nhân phải có thông tin chính xác và rõ ràng với tình trạng để tự theo dõi tình trạng sức khỏe của mình.

Ngoài ra, việc tìm kiếm các tài liệu chất lượng cao, uy tín và các thông tin học thuật để hỗ trợ các nhà nghiên cứu, chuyên gia nha khoa đang là vấn đề gây khó khăn. Tuy nhiên, do hạn chế về nguồn nhân lực và cơ sở hạ tầng y tế nên việc tiếp cận tư vấn nha khoa chuyên sâu tại nhiều nơi còn gặp nhiều khó khăn, nhất là với người Việt. Trong bối cảnh đó, việc phát triển một hệ thống chatbot AI có khả năng tư vấn nha khoa sẽ giúp cải thiện chất lượng chăm sóc sức khỏe răng miệng, hỗ trợ người dùng, bệnh nhân trong việc nhận diện triệu chứng và đề xuất phương án điều trị tại nhà hoặc xử lý sơ bộ trước khi gặp bác sĩ. Đồng thời, có thể giải đáp các thắc mắc về học thuật cho các nhà nghiên cứu, chuyên gia về lĩnh vực nha khoa.

### **1.1.2 Những hạn chế của các hệ thống hiện tại**

Mặc dù đã có nhiều nghiên cứu, phát triển và ứng dụng chatbot trong lĩnh vực y tế hay các hệ thống y tế chuyên khoa, nhưng các hệ thống hiện tại vẫn tồn tại nhiều hạn chế khi áp dụng vào chuyên ngành nha khoa, cụ thể là:

Thứ nhất, hầu hết các chatbot hiện nay hoạt động dựa trên các mô hình học sâu tổng quát không được chuyên sâu lĩnh vực nha khoa dẫn đến thiếu khả năng hiểu rõ các thuật ngữ chuyên môn và không đủ chính xác khi đưa ra lời khuyên về nha khoa. Điều này có thể dẫn đến tình trạng tư vấn sai hoặc không phù hợp với từng bệnh nhân.

Thứ hai, nhiều chatbot y tế hiện tại chủ yếu được phát triển bằng tiếng Anh hoặc ngôn ngữ khác, dẫn đến hạn chế trong khả năng hiểu và xử lý ngôn ngữ tự nhiên bằng tiếng Việt cho người Người dùng Việt. Việc thiếu dữ liệu huấn luyện đặc thù nha khoa cho tiếng Việt làm giảm đáng kể độ chính xác và hiệu quả của các mô hình AI trong việc tư vấn, chuẩn đoán.

Thứ ba, các chatbot thông thường hoạt động chủ yếu dựa trên cơ sở dữ liệu cố định hoặc các mô hình NLP chưa được tinh chỉnh và huấn luyện cho lĩnh vực nha khoa. Điều này khiến chatbot khó có thể cập nhật thông tin để có khả năng phản hồi chính xác với các tình huống phức tạp hoặc các câu hỏi chuyên sâu từ bệnh nhân.

Thứ tư, mỗi bệnh nhân có tiền sử nha khoa và điều kiện sức khỏe răng miệng khác nhau. Tuy nhiên, các chatbot hiện tại thường đưa ra câu trả lời chung chung mà chưa nắm rõ

tình trạng bệnh nhân, chưa có khả năng tùy chỉnh hay học hỏi để có thể phản hồi dựa trên hồ sơ bệnh nhân.

Những hạn chế này không chỉ là những thách thức mà còn đặt ra yêu cầu cần phải cải tiến các mô hình chatbot hiện có để tăng cường khả năng hiểu ngữ cảnh, xử lý chuyên sâu và đáp ứng tốt hơn nhu cầu của người dùng trong lĩnh vực nha khoa.

### **1.1.3 Tính cấp thiết và ý nghĩa ứng dụng**

Để khắc phục các hạn chế nêu trên thì việc tích hợp mạng kết hợp chuyên gia (Mixture of Experts - MoE) vào chatbot tư vấn nha khoa là một hướng đi đầy triển vọng và có tiềm năng phát triển. MoE là một kiến trúc mạng học sâu có thể kết hợp các mạng học sâu hiện đại có khả năng chọn lọc các chuyên gia xử lý thông tin trong từng ngữ cảnh cụ thể [3], giúp chatbot đưa ra phản hồi chính xác hơn dựa trên sự phối hợp của nhiều mô hình chuyên biệt. Việc áp dụng mạng kết hợp chuyên gia trong chatbot nha khoa mang lại nhiều lợi ích quan trọng, cụ thể là:

Lợi ích thứ nhất, giúp cải thiện độ chính xác của mô hình. Mỗi chuyên gia trong hệ thống MoE có thể tập trung vào một lĩnh vực cụ thể trong nha khoa ví dụ như điều trị sâu răng, chỉnh nha, bệnh lý về nướu, từ đó giúp chatbot cung cấp thông tin chuyên sâu và chính xác hơn [3].

Lợi ích thứ hai, tăng cường khả năng hiểu ngữ cảnh. Thay vì một mô hình NLP đơn lẻ, MoE cho phép chatbot lựa chọn chuyên gia phù hợp nhất với từng loại câu hỏi, giúp cải thiện khả năng xử lý ngôn ngữ tự nhiên và nhận diện đúng ý định của người dùng để cho ra câu trả lời tốt nhất.

Lợi ích thứ ba, tối ưu hóa hiệu suất tính toán. Nhờ vào cơ chế kích hoạt chuyên gia theo nhu cầu, mô hình MoE giúp tiết kiệm tài nguyên tính toán mà vẫn đảm bảo hiệu suất cao [3], phù hợp với các hệ thống chatbot triển khai trên thiết bị di động hoặc nền tảng web.

Lợi ích thứ tư, hỗ trợ cá nhân hóa tư vấn. Hệ thống có thể học từ dữ liệu của từng bệnh nhân để đưa ra các khuyến nghị phù hợp với tình trạng răng miệng và tiền sử nha khoa của họ, từ đó nâng cao trải nghiệm người dùng.

Với những lợi ích này, việc nghiên cứu và phát triển một chatbot nha khoa sử dụng mạng kết hợp chuyên gia không chỉ có ý nghĩa khoa học mà còn đóng góp đáng kể vào thực tiễn chăm sóc sức khỏe răng miệng. Hệ thống này không chỉ hỗ trợ bệnh nhân mà còn là một công cụ hữu ích giúp bác sĩ nha khoa tối ưu hóa quy trình làm việc, giảm tải áp lực trong việc tư vấn và theo dõi bệnh nhân.

Như vậy, đề tài "Nghiên cứu và xây dựng mô hình chatbot tư vấn nha khoa sử dụng mạng kết hợp chuyên gia" có tính cấp thiết và tiềm năng ứng dụng cao, hứa hẹn mang lại đóng góp quan trọng trong lĩnh vực y tế, đặc biệt là chăm sóc sức khỏe nha khoa.

## 1.2 Mục tiêu nghiên cứu

*"The best way to predict the future is to invent it."* bởi Alan Kay [4]. Như câu nói của Alan Kay thì trong thời đại mà trí tuệ nhân tạo đang tái định hình nhiều lĩnh vực của cuộc sống, từ sản xuất công nghiệp cho đến chăm sóc sức khỏe, câu nói này đặc biệt đúng với những nỗ lực chủ động xây dựng hệ thống hỗ trợ tư vấn y tế thông minh thay vì chờ đợi người khác thì ta sẽ tạo ra một hệ thống của riêng ta. Nha khoa là một chuyên ngành y tế có tính kỹ thuật cao nhưng thiếu hụt giải pháp AI bản địa hóa và đang đặt ra yêu cầu cấp thiết về những mô hình đối thoại có khả năng cung cấp phản hồi chuyên môn, chính xác và linh hoạt. Nghiên cứu này được thực hiện nhằm hướng đến việc phát triển một mô hình chatbot tư vấn nha khoa thông minh, khai thác sức mạnh của mô hình ngôn ngữ lớn **DeepSeek-R1** và tích hợp kiến trúc **mạng kết hợp chuyên gia (Mixture-of-Experts – MoE)** nhằm đáp ứng được yêu cầu đặc thù của ngôn ngữ y khoa, đặc biệt trong bối cảnh sử dụng tiếng Việt.

### 1.2.1 Mục tiêu tổng quát

Mục tiêu tổng quát của nghiên cứu là xây dựng một hệ thống chatbot tư vấn nha khoa thông minh, có khả năng cung cấp thông tin chính xác, nhất quán và phù hợp với từng ngữ cảnh chuyên môn. Hệ thống không chỉ đơn thuần phản hồi câu hỏi theo cách rập khuôn, mà còn phải thể hiện khả năng **hiểu ngôn ngữ tự nhiên chuyên ngành, nhận diện bối cảnh hội thoại**, và đưa ra tư vấn có tính **cá nhân hóa** dựa trên đặc điểm sức khỏe răng miệng cụ thể của từng người dùng. Đồng thời, hệ thống có thể cải tiến lên các bài toán lớn chatbot như đa mô hình hay nhận diện. Về mặt kỹ thuật, mô hình chatbot

được phát triển dựa trên nền tảng **DeepSeek-R1**, một kiến trúc ngôn ngữ lớn tiên tiến có tích hợp **MoE** để chia nhiệm vụ xử lý ngôn ngữ thành các vùng chuyên biệt, giúp tăng hiệu suất suy luận và khả năng học tập chuyên sâu [3] [5]. Việc fine-tuning mô hình trên dữ liệu chuyên ngành nha khoa tiếng Việt, cùng với chiến lược học từ phản hồi con người, sẽ giúp hệ thống không chỉ "hiểu đúng" mà còn "nói đúng", "nói đủ", và "nói như một chuyên gia".

Cuối cùng, hệ thống được kỳ vọng sẽ vận hành hiệu quả trong môi trường thực tiễn, với khả năng đáp ứng tương tác đa dạng từ người dùng, phục vụ như một **trợ lý y khoa ảo** hỗ trợ sàng lọc, tư vấn cơ bản và cung cấp thông tin chuẩn y khoa trong phạm vi cộng đồng.

### 1.2.2 Mục tiêu cụ thể

Để hiện thực hóa mục tiêu tổng quát, nghiên cứu tập trung triển khai ba trục nội dung quan trọng, đảm bảo sự kết nối chặt chẽ từ tầng dữ liệu đến triển khai thực nghiệm.

Trước tiên, nghiên cứu tiến hành xây dựng **tập dữ liệu ngôn ngữ chuyên biệt dành cho lĩnh vực nha khoa tiếng Việt**, được thu thập từ nhiều nguồn như: tài liệu y học, hướng dẫn lâm sàng, đối thoại giữa bác sĩ với bệnh nhân, và tài liệu đào tạo chuyên ngành. Đặc trưng của dữ liệu y tế là tính nhạy cảm, chuẩn xác và giàu tính ngữ cảnh, do đó việc tiền xử lý không chỉ dừng lại ở loại bỏ nhiễu mà còn bao gồm các bước: **chuẩn hóa thuật ngữ y khoa, gán nhãn ngữ nghĩa, và điều chỉnh cấu trúc hội thoại** để đảm bảo tính khả dụng trong môi trường học sâu. Dữ liệu đầu ra được tổ chức thành các cặp prompt-response có phân loại chức năng hội thoại như tư vấn triệu chứng, giải thích bệnh lý, hướng dẫn chăm sóc, nhằm tạo tiền đề cho quá trình fine-tune cũng như cải tiến sau này cho mô hình hay phục vụ cho nhiều bài toán khác.

Giai đoạn thứ hai là **tinh chỉnh mô hình DeepSeek-R1** trên tập dữ liệu đã xử lý, kết hợp hai chiến lược: **Supervised Fine-Tuning (SFT)** và **Reinforcement Learning from Human Feedback (RLHF)**. Phương pháp SFT giúp mô hình học được cấu trúc câu trả lời chuẩn từ các cặp hội thoại đã được xác thực, trong khi RLHF giúp tối ưu hóa phản hồi từ góc nhìn người dùng thông qua cơ chế phạt với thưởng dựa trên đánh giá chủ quan [6]. Mô hình sau tinh chỉnh không chỉ được đánh giá bằng các chỉ số phổ biến như **BLEU**, **ROUGE** [7] [8], mà còn thông qua **hiệu quả tương tác trong các kịch bản hội thoại**

**mô phỏng**, nơi sự tự nhiên và tính chuyên môn trong phản hồi là thước đo chính.

Cuối cùng, hệ thống chatbot sẽ được triển khai trong môi trường thực tế thông qua một ứng dụng thử nghiệm hoặc giao diện website, nơi mà người dùng có thể tương tác để nhận được tư vấn nha khoa cơ bản. Giai đoạn này không chỉ giúp xác định **khả năng ứng dụng của mô hình trong bối cảnh thực tiễn**, mà còn là nền tảng để thu thập thêm dữ liệu phản hồi làm cơ sở cho các vòng huấn luyện liên tục trong tương lai. Phản hồi từ chuyên gia nha khoa và người dùng thật sẽ được tích hợp vào hệ thống đánh giá, giúp định hình chiến lược cải tiến mô hình trong các phiên bản tiếp theo. Qua đó, nghiên cứu không chỉ dừng lại ở một bài toán học thuật, mà còn hướng tới giá trị ứng dụng trong cộng đồng để có thể là nơi nhu cầu tư vấn y khoa đáng tin cậy và dễ tiếp cận đang ngày càng trở nên cấp thiết.

### 1.3 Đối tượng nghiên cứu và phạm vi nghiên cứu

#### 1.3.1 Đối tượng nghiên cứu

Đối tượng trọng tâm của nghiên cứu là một mô hình chatbot nha khoa được tối ưu với ngôn ngữ tiếng Việt, có năng lực tương tác và tư vấn các vấn đề nha khoa dựa trên tri thức chuyên ngành và khả năng lý giải tự nhiên của mô hình ngôn ngữ lớn. Hệ thống này được xây dựng dựa trên nền tảng mô hình DeepSeek-R1 dựa trên một kiến trúc Mixture-of-Experts hiện đại đã được thiết kế lại nhằm phù hợp với môi trường tương tác y tế đa chiều, nơi mà thông tin đầu vào thường mơ hồ, thiếu cấu trúc, nhưng lại yêu cầu phản hồi mang tính “xác tín” và cá nhân hóa.

Không giống các hệ thống trả lời tự động dựa trên kịch bản định sẵn, chatbot đề xuất trong nghiên cứu được xây dựng như một thực thể có khả năng tái cấu trúc tri thức từ các đơn vị dữ liệu ngôn ngữ được thu thập từ hội thoại giữa bác sĩ và bệnh nhân, hướng dẫn lâm sàng, báo cáo nghiên cứu cho đến các tài liệu giáo khoa nha khoa rồi tất cả đều được xử lý sâu để đạt độ chuẩn hóa cao về cả cấu trúc ngữ nghĩa lẫn khái niệm chuyên môn. Việc này không chỉ nhằm mục tiêu huấn luyện mô hình một cách chính xác, mà còn để hình thành một tập tri thức khả giải (explainable knowledge), phục vụ cho quá trình suy luận nội bộ của chatbot khi đưa ra phản hồi.

Mô hình không dừng lại ở việc hiểu ngôn ngữ thuần túy mà còn hướng đến khả năng diễn giải triệu chứng, xác định bối cảnh điều trị, phân biệt các nhóm bệnh lý và đề xuất định



hướng tiếp cận ban đầu phù hợp. Quá trình huấn luyện không chỉ dựa trên supervised fine-tuning mà còn ứng dụng reinforcement learning from human feedback (RLHF) để tinh chỉnh hành vi phản hồi dựa trên các tình huống giao tiếp thực tế, từ đó tiệm cận hơn với phong cách tương tác của chuyên gia lâm sàng.

### **1.3.2 Phạm vi nghiên cứu**

Nghiên cứu tập trung vào toàn bộ chuỗi xử lý của một hệ thống AI hội thoại chuyên biệt, bắt đầu từ thiết kế kiến trúc mô hình, thu thập và tổ chức lại tri thức chuyên ngành, cho đến tinh chỉnh mô hình ngôn ngữ lớn bằng phương pháp fine-tuning có kiểm soát. Việc xây dựng tập dữ liệu không đơn thuần là thu thập, mà bao gồm cả quá trình lọc nhiễu ngữ nghĩa, hiệu chỉnh cú pháp chuyên môn và chuyển hóa dữ liệu thành dạng phản hồi hội thoại có nhịp điệu giao tiếp tự nhiên.

Phạm vi kỹ thuật bao gồm việc áp dụng DeepSeek-R1 như một khung mô hình nền, trong đó cấu trúc Mixture-of-Experts được khai thác để tách biệt các nhiệm vụ tư vấn thành các phân vùng chuyên biệt: một vùng phụ trách phân tích triệu chứng, một vùng cho chẩn đoán sơ bộ, một vùng cung cấp hướng dẫn điều trị, và một vùng xử lý giao tiếp ngữ nghĩa tự nhiên. Quá trình routing giữa các vùng này sẽ được kiểm soát thông qua tín hiệu ngữ nghĩa nội tại (semantic signal control), nhằm đảm bảo phản hồi có tổ chức và tránh lặp lại hoặc lan man mà có thể dễ dàng nhận thấy là hiện tượng thường gặp trong các LLM không được chuyên ngành hóa. Việc triển khai thực nghiệm chatbot trong môi trường ứng dụng giả lập (sandbox) cũng nằm trong phạm vi nghiên cứu, cho phép kiểm định các yếu tố: (1) độ chính xác của phản hồi theo kiểm định chuyên gia, (2) mức độ duy trì ngữ cảnh trong hội thoại dài, (3) khả năng tái lập phản hồi trên các trường hợp có cấu trúc tương đồng, và (4) độ chấp nhận của người dùng qua khảo sát thực tế. Đây là những tiêu chí cốt lõi nhằm đánh giá mức độ sẵn sàng của mô hình trong việc tích hợp vào các hệ thống tư vấn y tế bán tự động.

Nghiên cứu không đặt mục tiêu thay thế vai trò của bác sĩ chuyên khoa, mà hướng đến việc hỗ trợ tăng cường (augmented intelligence) cho người dùng trong quá trình tự khám phá tri thức sức khỏe, giảm tải áp lực cho tuyến tư vấn đầu và tăng tính sẵn sàng cho các trường hợp không khẩn cấp. Ngoài ra, hệ thống còn có tiềm năng mở rộng sang các

chuyên ngành y tế khác thông qua tái huấn luyện trên cơ sở kiến trúc chuyên gia hiện có, tạo nền móng cho các chatbot hội thoại đa lĩnh vực chuyên môn trong tương lai.

## **1.4 Câu hỏi nghiên cứu và giả thuyết**

### **1.4.1 Các vấn đề cần giải quyết**

Việc xây dựng một chatbot tư vấn nha khoa sử dụng mô hình ngôn ngữ lớn không đơn thuần là áp dụng một kiến trúc tổng quát cho một tác vụ cụ thể, mà là quá trình hiệu chỉnh tinh vi giữa ngôn ngữ, chuyên môn và ngữ cảnh. Khi tiến hành fine-tune trên nền mô hình DeepSeek-R1 là một phiên bản ngôn ngữ mạnh mẽ được huấn luyện trên tập dữ liệu đa miền sẽ đối mặt với nhiều thách thức đặc thù của bài toán tư vấn y tế mà đã nhanh chóng lộ diện.

Một trong những thách thức cốt lõi là năng lực hiểu ngữ cảnh trong hội thoại chuyên ngành nha khoa. Khác với ngôn ngữ thông thường, hội thoại y tế đòi hỏi mô hình không chỉ hiểu đúng nghĩa đen mà còn phải nắm được ý định hàm ẩn, các mối liên hệ giữa triệu chứng và tình trạng bệnh lý, cũng như sự phụ thuộc ngữ nghĩa theo ngữ cảnh. Những khác biệt ngôn ngữ vùng miền, đặc điểm đa dạng trong cách bệnh nhân mô tả triệu chứng bằng tiếng Việt đời thường càng làm gia tăng mức độ phức tạp. Tiếp theo, độ chính xác chuyên môn trong phản hồi là yếu tố bắt buộc đối với bất kỳ hệ thống tư vấn nào liên quan đến sức khỏe. Không giống như chatbot tổng quát có thể chấp nhận một mức độ “mơ hồ” trong phản hồi, chatbot nha khoa cần đảm bảo rằng mỗi câu trả lời được đưa ra đều nằm trong phạm vi chấp nhận được của thực hành y khoa. Việc đưa ra khuyến nghị sai, không đầy đủ hoặc quá chung chung không chỉ ảnh hưởng đến trải nghiệm người dùng mà còn tiềm ẩn rủi ro về mặt đạo đức và pháp lý. Bên cạnh đó, khả năng cá nhân hóa tư vấn là một khía cạnh quan trọng nhưng thường bị bỏ qua trong nhiều nghiên cứu trước đây. Mỗi bệnh nhân mang theo một nền tảng sức khỏe khác nhau, những lo ngại khác nhau và mong đợi khác nhau. Do đó, chatbot cần có khả năng thích ứng với hồ sơ cá nhân hóa, không chỉ bằng cách điều chỉnh phản hồi theo biểu hiện triệu chứng, mà còn qua việc lựa chọn từ ngữ dễ hiểu, phong cách giao tiếp gần gũi và phù hợp với từng đối tượng cụ thể.

Thách thức cuối cùng liên quan đến khả năng phân loại và xử lý tác vụ trong môi trường

hội thoại phức tạp. Một hệ thống hiệu quả cần phân biệt được giữa yêu cầu chẩn đoán sơ bộ, yêu cầu cung cấp thông tin, hay yêu cầu đưa ra hướng dẫn hành động. Khả năng điều hướng đúng chuyên môn tương ứng cho từng loại câu hỏi đóng vai trò nền tảng cho việc tối ưu hóa tốc độ phản hồi mà không làm tổn hại đến độ chính xác.

Như vậy, việc fine-tune DeepSeek-R1 cho tác vụ chatbot nha khoa không chỉ là một bài toán kỹ thuật, mà là một thử thách đa chiều, đòi hỏi sự kết hợp chặt chẽ giữa hiểu ngữ nghĩa ngôn ngữ tiếng Việt, kiến thức chuyên ngành và tối ưu hóa kiến trúc mô hình.

#### 1.4.2 Giả thuyết nghiên cứu

Trước những vấn đề đặt ra, đề tài này được triển khai với một chuỗi các giả thuyết nghiên cứu cụ thể nhằm kiểm chứng hiệu quả của việc tinh chỉnh mô hình DeepSeek-R1 trên tập dữ liệu chuyên biệt nha khoa tiếng Việt.

Giả thuyết người nghiên cứu là: *việc fine-tune mô hình DeepSeek-R1 trên tập dữ liệu nha khoa đã được xử lý và gán nhãn theo từng nhóm tác vụ, kết hợp với kỹ thuật Supervised Fine-Tuning (SFT) và Reinforcement Learning from Human Feedback (RLHF), sẽ giúp cải thiện đáng kể độ chính xác, mức độ hiểu ngữ cảnh và khả năng phản hồi cá nhân hóa của chatbot trong nhiệm vụ tư vấn nha khoa.* Cụ thể, giả định rằng khi huấn luyện lại mô hình trên một tập dữ liệu chứa hội thoại thực tế giữa bác sĩ và bệnh nhân, cùng các đoạn văn bản từ tài liệu y khoa đã được chọn lọc và chuẩn hóa, mô hình sẽ học được cách sử dụng ngôn ngữ chuyên ngành một cách mạch lạc và chính xác, đồng thời nắm bắt được các cấu trúc hội thoại đặc trưng của môi trường lâm sàng. Việc phân loại các mẫu dữ liệu thành các nhóm chức năng như *chẩn đoán, đưa ra phương án điều trị, giải thích thuật ngữ, hay tư vấn phòng ngừa* sẽ tạo ra các tuyến phản hồi chuyên biệt, giúp chatbot không chỉ đưa ra thông tin đúng, mà còn đúng theo "vai trò" đang được yêu cầu. Ngoài ra, đề tài cũng xây dựng giả thuyết rằng việc áp dụng kiến trúc mạng kết hợp chuyên gia (Mixture-of-Experts - MoE) có khả năng giúp mô hình DeepSeek-R1 phân chia tác vụ một cách tối ưu hơn, từ đó cải thiện hiệu suất xử lý các truy vấn đồng thời trên nhiều lớp kiến thức. Các “expert” được huấn luyện tập trung theo nhóm nhiệm vụ sẽ đóng vai trò như những bộ lọc chuyên môn hóa, cho phép mô hình kích hoạt chỉ các phần cần thiết trong quá trình phản hồi, nhờ vậy vừa tăng tốc độ suy luận vừa duy trì chất lượng đầu ra [3].

Cuối cùng, nghiên cứu giả định rằng mô hình fine-tuned theo hướng nói trên sẽ có thể mô phỏng được phong cách giao tiếp thân thiện, dễ hiểu, gần gũi với cách hành xử của các bác sĩ trong môi trường tư vấn thực tế, qua đó nâng cao mức độ chấp nhận và hài lòng từ phía người dùng. Đây không chỉ là bài toán về hiệu quả kỹ thuật mà còn là thử nghiệm về khả năng đồng cảm và tương tác nhân bản trong ứng dụng AI.

## **1.5 Phương pháp và công cụ nghiên cứu**

### **1.5.1 Phương pháp nghiên cứu**

Nghiên cứu này được thực hiện theo cách tiếp cận hệ thống nghĩa là tiếp cận toàn diện nên luôn xem xét nhiều yếu tố và mối quan hệ của các mô hình LLM sử dụng MoE và dữ liệu nha khoa, kết hợp giữa tổng hợp lý thuyết về LLMs, phân tích dữ liệu nha khoa và thực nghiệm chatbot nhằm đảm bảo tính chặt chẽ và khả năng ứng dụng thực tiễn. Trước tiên, quá trình khảo sát các tài liệu về DeepSeek hay các LLMs về Nha Khoa liên quan giúp định hình nền tảng lý thuyết về chatbot tư vấn Nha Khoa trong lĩnh vực y tế, các mô hình học sâu ứng dụng trong xử lý ngôn ngữ tự nhiên (NLP) và cơ chế tối ưu hóa dựa trên mạng kết hợp chuyên gia. Những nghiên cứu trước đây về chatbot y tế và các phương pháp fine-tuning mô hình ngôn ngữ lớn (LLM) cung cấp một cái nhìn tổng quan về những thách thức hiện hữu, từ đó xác định hướng tiếp cận phù hợp cho bài toán tư vấn nha khoa.

Dữ liệu đóng vai trò trung tâm trong việc xây dựng chatbot, do đó, quy trình thu thập và xử lý dữ liệu được tiến hành một cách có hệ thống và khoa học. Các nguồn dữ liệu từ tài liệu y khoa, hội thoại thực tế giữa bác sĩ và bệnh nhân, cũng như tập dữ liệu dịch thuật chuyên ngành được phân tích kỹ lưỡng nhằm đảm bảo tính đa dạng và sát với ngữ cảnh sử dụng. Sau khi thu thập, dữ liệu được làm sạch, chuẩn hóa và gán nhãn cho từng task để phục vụ cho quá trình huấn luyện mô hình. Thực nghiệm được triển khai theo hai giai đoạn chính: huấn luyện mô hình DeepSeek trên dữ liệu nha khoa và đánh giá hiệu suất chatbot dựa trên các tiêu chí định lượng hoặc là trải nghiệm người dùng. Quá trình fine-tuning được thực hiện theo phương pháp Supervised Fine-Tuning (SFT) kết hợp với Reinforcement Learning from Human Feedback (RLHF) để cải thiện khả năng phản hồi chính xác, phù hợp theo từng ngữ cảnh chuyên ngành nha khoa. Các phép đo về độ chính xác, tính nhất quán và mức độ hài lòng của người dùng được sử dụng để đánh giá hiệu suất chatbot sau khi fine-tuning.

### 1.5.2 Công cụ nghiên cứu

Trong quá trình nghiên cứu và phát triển mô hình chatbot hỗ trợ tư vấn nha khoa bằng tiếng Việt, việc lựa chọn một mô hình ngôn ngữ lớn (LLM) phù hợp với tài nguyên phần cứng và khả năng hiểu tiếng Việt là yếu tố then chốt. Trong phạm vi đề án này, nhóm đã khảo sát các biến thể mô hình thuộc hệ sinh thái DeepSeek nhằm xác định phương án tinh chỉnh phù hợp nhất với điều kiện phần cứng (cụ thể tối thiểu là GPU 10 - 16GB VRAM) và mục tiêu ngôn ngữ ứng dụng. Các mô hình được xem xét bao gồm cả các mô hình đã được nén và rút gọn từ mô hình gốc lớn hơn, các mô hình tổng quát, chuyên biệt cho lập trình nhằm đánh giá toàn diện về hiệu năng, kích thước và mức độ hỗ trợ tiếng Việt. Để đảm bảo khả năng huấn luyện và triển khai mô hình hiệu quả, nghiên cứu sử dụng mô hình **LLaMA chưng cất tri thức từ DeepSeek R1 với 8 tỷ tham số** [9], cụ thể là DeepSeek-R1 được chưng cất kiến thức từ DeepSeek-R1 vào Llama-3.1-8B-Instruct của Meta AI với 8 tỷ tham số được chuyển giao khả năng suy luận mạnh mẽ của mô hình của DeepSeek-R1 vào mô hình Llama có khả năng thực hiện các tác vụ suy luận phức tạp, tạo ra các chuỗi suy nghĩ chi tiết để giải quyết vấn đề, giúp người dùng hiểu rõ hơn về quá trình suy luận của mô hình nhưng với kích thước nhỏ gọn của mô hình cho phép triển khai trên các thiết bị cá nhân, giảm độ trễ và chi phí vận hành so với các giải pháp đám mây nặng nề. Việc sử dụng kỹ thuật tối ưu **Knowledge Distillation** không chỉ giảm kích thước, giữ hiệu suất mô hình mà còn giúp mô hình được tối ưu cho tiếng Việt. Tuy việc chưng cất kiến thức đã làm cấu trúc MoE gốc không còn tồn tại nữa nhưng đổi lại mô hình đã được thừa kế ưu điểm, khả năng suy luận, chuyên môn và yêu cầu tài nguyên tính toán thấp hơn, dễ triển khai hơn so với mô hình gốc. Mô hình trên giúp giảm yêu cầu tài nguyên tính toán mà vẫn duy trì hiệu suất cao trong các nhiệm vụ như toán học, lập trình và suy luận logic, một mô hình được thừa kế khả năng Mixture-of-Experts (MoE) tiên tiến với hiệu suất cao trong xử lý ngôn ngữ tự nhiên. Mô hình giúp giảm thiểu chi phí tính toán mà vẫn duy trì khả năng tổng quát hóa mạnh mẽ, đặc biệt hữu ích khi áp dụng vào chatbot tư vấn y tế. Dưới đây là bảng 1.1 thống kê các mô hình được tối ưu cho nghiên cứu và tinh chỉnh:

Mô hình	Số tham số	Kiến trúc gốc	Ứng dụng chính	Tính năng nổi bật
<b>DeepSeek-R1-Distill-LLaMA-8B</b>	8B	Distilled từ LLaMA-2 13B	Chatbot, hỏi đáp, NLP tổng quát	Nhẹ hơn LLaMA gốc, tối ưu tốc độ, giảm chi phí
<b>DeepSeek-LLM-7B</b>	7B	Tương tự LLaMA-2 7B	Viết, suy luận, tóm tắt	Pretrain từ scratch, dùng corpus lớn đa ngôn ngữ
<b>DeepSeek-Coder-1.3B</b>	1.3B	Tùy chỉnh từ transformer	Sinh code, sửa lỗi, hỗ trợ lập trình	Hỗ trợ hơn 80 ngôn ngữ lập trình, tốc độ nhanh
<b>DeepSeek-MoE-16B</b>	16B (MoE), 64 experts	Mixture of Experts	Nhiệm vụ NLP phức tạp	Dù tổng tham số là 16B, mỗi lần chỉ dùng ~2B, tiết kiệm tài nguyên

**Bảng 1.1 Mô hình sử dụng và tối ưu cho nghiên cứu**

Dữ liệu chuyên biệt Nha Khoa sẽ được thu từ các tài liệu, trang web, các mã nguồn mở Huggingface, Kaggle hoặc có thể là dịch từ bộ dữ liệu chuyên sâu tiếng anh. Sau đó, tổng hợp và xây dựng thành một bộ dữ liệu chất lượng cao.

#### Môi trường thực nghiệm:

- Ngôn ngữ: Python, HTML, Javascript, CSS
- Nền tảng phát triển: Jupyter Notebook, Google Colab, Kaggle
- Công cụ: Visual Studio Code, Docker, Git, Pycharm

Dịch vụ	Chức năng chính
<b>Cloudflare Tunnel</b>	Tạo đường hầm bảo mật từ máy cục bộ ra internet.
<b>Ngrok</b>	Tạo địa chỉ URL tạm thời cho mô hình.
<b>Google Search API</b>	Truy xuất dữ liệu trên Google.
<b>Gemini API (Google)</b>	Làm trung gian như người phiên dịch viên cho LLM
<b>HuggingFace API</b>	Lưu trữ và truy cập thư viện dữ liệu, mô hình.
<b>Qdrant</b>	Cơ sở dữ liệu vector dùng cho tìm kiếm ngữ nghĩa

**Bảng 1.2 Dịch vụ sử dụng trong nghiên cứu**

Thiết bị	RAM	GPU	VRAM	OS
PC1	32GB	RTX 3050	8GB	Windows 11
Ubuntu	16GB	RTX 3050	8GB	Ubuntu
PC2	16GB	GTX 1650	4GB	Windows 11
Google Colab T4	16GB	Tesla T4	16GB	Linux
Kaggle P100	13GB	Tesla P100	16GB	Linux

**Bảng 1.3 Thiết bị dùng để nghiên cứu**

## 1.6 Cấu trúc đề tài và dàn ý các chương

Đề tài "Nghiên cứu và xây dựng mô hình chatbot tư vấn nha khoa sử dụng mạng kết hợp chuyên gia" được cấu trúc và hệ thống thành năm chương chính, phản ánh một cách hệ thống, khoa học quá trình nghiên cứu, phát triển và triển khai mô hình. Cách thức phân chia nội dung giữa các chương đảm bảo tính khoa học, không trùng lặp và hướng đến việc giải quyết bài toán một cách toàn diện, từ cơ sở lý thuyết đến ứng dụng thực tiễn.

Mở đầu đề án là phần **Giới thiệu tổng quan về bài toán**, trong đó việc xác định bài toán, đặt nền tảng cho vấn đề nghiên cứu, phân tích các hạn chế của những phương pháp hiện tại và đề xuất hướng tiếp cận sử dụng mô hình DeepSeek kết hợp với mạng kết hợp chuyên gia trong lĩnh vực nha khoa. **Chương này cũng trình bày mục tiêu nghiên cứu, phạm vi thực hiện và những đóng góp kỳ vọng của đề tài.**

Tiếp nối phần giới thiệu là **Chương 2 - Kiến thức nền tảng và mô hình DeepSeek**, tập trung trình bày các cơ sở lý thuyết liên quan đến Transformer và DeepSeek sử dụng mạng kết hợp chuyên gia, các phương pháp fine-tuning, tối ưu mô hình ngôn ngữ lớn (LLM) và cơ chế của mạng kết hợp chuyên gia. Trong đó, mô hình DeepSeek và kiến trúc Mixture-of-Experts (MoE) sẽ được phân tích kỹ lưỡng về ý tưởng, cấu trúc, cách hoạt động, làm rõ khả năng ứng dụng vào bài toán tư vấn nha khoa, đồng thời so sánh với các kiến trúc khác nhằm xác định lý do lựa chọn kiến trúc này vào bài toán nghiên cứu.

Sau khi xác định được nền tảng lý thuyết, **Chương 3 - Thu thập và xử lý dữ liệu nha khoa sẽ tập trung vào việc xây dựng tập dữ liệu huấn luyện**. Phần này không chỉ trình bày quá trình thu thập dữ liệu từ các nguồn như bài báo y khoa, cuộc hội thoại giữa bác

sĩ và bệnh nhân mà còn đi sâu vào các kỹ thuật xử lý dữ liệu chuyên biệt và cách định dạng dữ liệu cho lĩnh vực nha khoa sao cho mô hình có thể tối ưu nhất có thể. Những thách thức trong xử lý văn bản y tế, phương pháp làm sạch dữ liệu, kỹ thuật tiền xử lý đặc thù và chiến lược gán nhãn sẽ được đề cập, đảm bảo mô hình có nguồn dữ liệu chất lượng cao để tối ưu hóa hiệu suất.

Dữ liệu sau khi xử lý sẽ là đầu vào **cho Chương 4 - Fine-tune mô hình DeepSeek cho chatbot Nha Khoa, nơi tập trung vào việc tinh chỉnh mô hình DeepSeek để tối ưu hóa khả năng hiểu ngữ cảnh nha khoa.** Chương này trình bày quy trình fine-tuning với Supervised Fine-Tuning (SFT) và Reinforcement Learning from Human Feedback (RLHF), một số module nhỏ, đồng thời đánh giá hiệu suất mô hình qua các thí nghiệm thực nghiệm. Quá trình tinh chỉnh được phân tích từ góc độ tối ưu hóa tham số, kỹ thuật tối ưu khi tinh chỉnh, kiểm định mô hình trên tập dữ liệu thử nghiệm và so sánh với các mô hình khác để chứng minh hiệu quả của cách tiếp cận và ứng dụng thực tế.

Cuối cùng, **Chương 5 - Triển khai và ứng dụng thực tế sẽ trình bày quá trình tích hợp mô hình vào một hệ thống chatbot hoàn chỉnh.** Chương này bao gồm thiết kế hệ thống chatbot, kiến trúc triển khai, kiểm thử mô hình trong môi trường thực tế và thu thập phản hồi từ người dùng. Những vấn đề như hiệu suất xử lý, khả năng mở rộng, bảo mật dữ liệu và tiềm năng phát triển trong tương lai cũng sẽ được thảo luận nhằm đề xuất các hướng cải tiến cho mô hình chatbot tư vấn nha khoa.

Với cách tổ chức nội dung như trên, đề án không chỉ đảm bảo tính chặt chẽ về mặt học thuật mà còn hướng đến khả năng ứng dụng thực tiễn, tạo ra một hệ thống chatbot có tính chính xác cao, đáp ứng nhu cầu tư vấn nha khoa tự động một cách hiệu quả. Đồng thời, chương này tạo cơ sở định hướng cho toàn bộ đề án, mở đường cho chương 2 là nơi trình bày kiến thức nền tảng và các mô hình được lựa chọn để hiện thực hóa hệ thống chatbot nha khoa.



## CHƯƠNG 2 : KIẾN THỨC NỀN TẢNG VÀ MÔ HÌNH DEEPSEEK

Chương này cung cấp cơ sở lý thuyết về trí tuệ nhân tạo, mô hình Transformer, kiến trúc mạng kết hợp chuyên gia (MoE) và mô hình DeepSeek. Các công nghệ và phương pháp tối ưu như SFT, OPRO được trình bày nhằm giúp người đọc hiểu rõ cơ chế hoạt động và lý do lựa chọn kiến trúc DeepSeek cho bài toán đặt ra.

### 2.1 Tổng quan kiến thức nền tảng

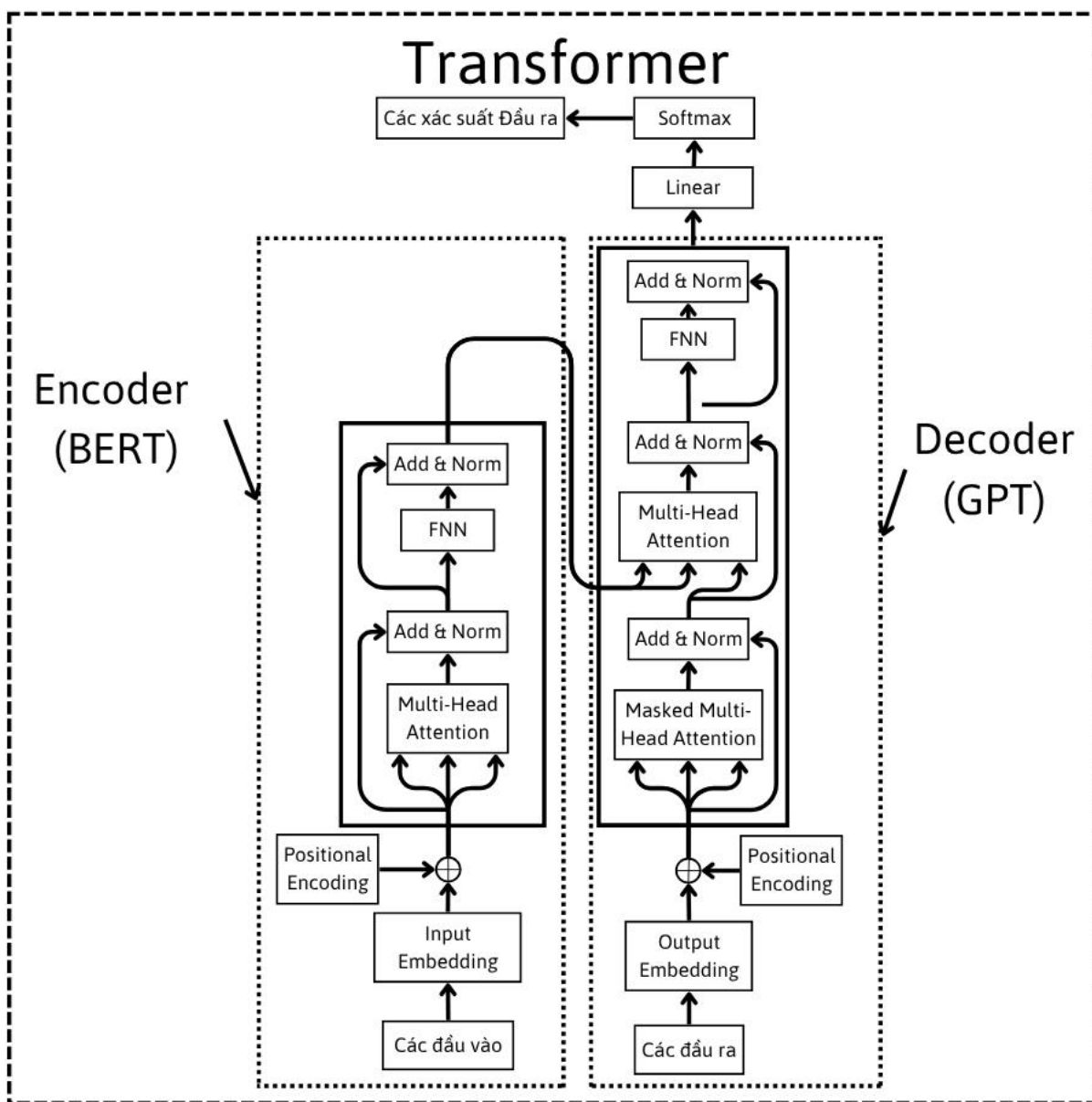
#### 2.1.1 Các khái niệm cơ bản về trí tuệ nhân tạo, Transformer và ứng dụng trong chatbot

Ở Trí tuệ nhân tạo (Artificial Intelligence - AI) đã chứng kiến sự phát triển mạnh mẽ trong thập kỷ qua, đặc biệt là trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) và các Mô hình ngôn ngữ lớn (Large Language Models - LLMs). Từ những mô hình đơn giản dựa trên quy tắc cho đến các kiến trúc học sâu hiện đại ngày nay, AI đã thay đổi đáng kể cách con người tương tác với máy tính và sử dụng chúng trong cách công việc hàng ngày ở mọi lĩnh vực công việc, đặc biệt trong lĩnh vực chatbot. Trong số các kiến trúc tiên tiến, Transformer đã nổi lên như một phương pháp tiếp cận đột phá [10], mang lại những cải tiến đáng kể so với các mô hình tiền nhiệm như Recurrent Neural Networks (RNNs) và Long Short-Term Memory (LSTM). Cơ chế Attention, đặc biệt là Self-Attention, giúp Transformer có khả năng xử lý toàn bộ ngữ cảnh của một câu một cách song song, thay vì phải phân tích tuần tự như RNNs. Điều này không chỉ giúp cải thiện đáng kể hiệu suất tính toán mà còn tăng khả năng học các mối quan hệ ngữ nghĩa dài hạn trong văn bản. Transformer còn là nền tảng cho các mô hình học sâu lớn hiện nay cũng như các mô hình ngôn ngữ lớn hiện nay của các ông lớn như Open AI, Llama của Meta AI hay DeepSeek-R1 và có thể giải quyết được gần như các loại bài toán từ Ai tạo sinh, xử lý ngôn ngữ tự nhiên, thị giác máy tính, xử lý âm thanh,...

Các mô hình Transformer ngày nay, bao gồm BERT, GPT, T5 và DeepSeek, đã được sử dụng rộng rãi trong nhiều ứng dụng với NLP như dịch máy, tổng hợp văn bản, nhận diện thực thể, và đặc biệt nhất là LLM hay chatbot [10]. Chatbot hiện đại không chỉ dừng lại ở việc phản hồi dựa trên kịch bản có sẵn mà còn có khả năng tạo câu trả lời tự nhiên, phù hợp với ngữ cảnh của người dùng và lấy thông tin thực để cải thiện khả năng phản hồi. Việc fine-tuning các mô hình ngôn ngữ lớn (LLMs) trên dữ liệu chuyên ngành như y tế, tài chính, hay pháp luật đã mở ra nhiều hướng ứng dụng mới, giúp chatbot trở

thành một công cụ hỗ trợ hiệu quả trong thực tế.

Mô hình Transformer có 2 phần chính là Encoder và Decoder [10]. Trong đó, Encoder là bộ mã hóa thường có công việc chính là giúp mô hình hiểu câu bằng cách dùng cơ chế Attention để tìm ra mối quan hệ các từ trong câu. Đối với Decoder là bộ giải mã lại có khả năng là tạo hay sinh ra các từ tiếp theo của câu bằng cách dùng cơ chế Attention. Cả hai đều có cấu trúc như hình 2.1 dưới đây:



**Hình 2.1 Mô tả quá trình hoạt động của Transformer**

- **Self-Attention (Multi-Head Self-Attention - MHSA):** Học mối quan hệ giữa các từ trong chuỗi đầu vào.
- **Feed Forward Network (FFN):** Một mạng nơ-ron đơn giản xử lý thông tin sau attention.

- **Add & Norm:** Cộng tắt (Residual Connection) và chuẩn hóa (Layer Normalization).

Nhưng Decoder có thêm Masked Self-Attention giúp ngăn nhìn thấy tương lai (để tránh dự đoán dựa vào thông tin chưa có). Theo dõi hình để hiểu cách Transformer hoạt động:

Trong Transformer có thêm cơ chế Attention tạo cuộc cách mạng cho AI tạo sinh.

Self-Attention là thành phần cốt lõi, tính toán mối quan hệ giữa các từ trong một câu.

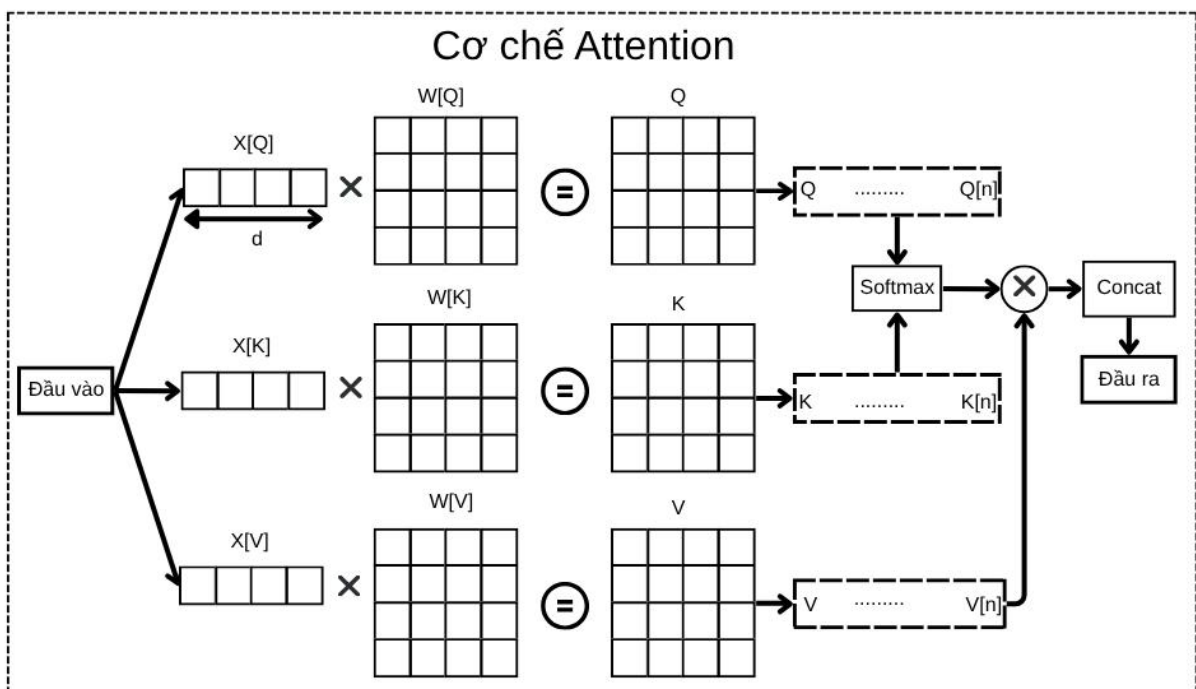
Công thức như sau:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

Ý nghĩa của các thành phần:

- **Q (Query)** – Truy vấn (Từ cần tìm).
- **K (Key)** – Khóa (để so sánh với truy vấn).
- **V (Value)** – Giá trị (thông tin đầu vào được lấy ra dựa trên sự chú ý).

Cơ chế Attention hiện nay đã phát triển nhiều biến thể và có các cấu trúc cũng như cách hoạt động khác nhau để tối ưu cho nhiều loại bài toán. Nhưng Self-Attention vẫn nền tảng và là cơ chế cốt lõi. Theo dõi hình 2.2 để hiểu rõ hơn cơ chế hoạt động của Self-Attention:

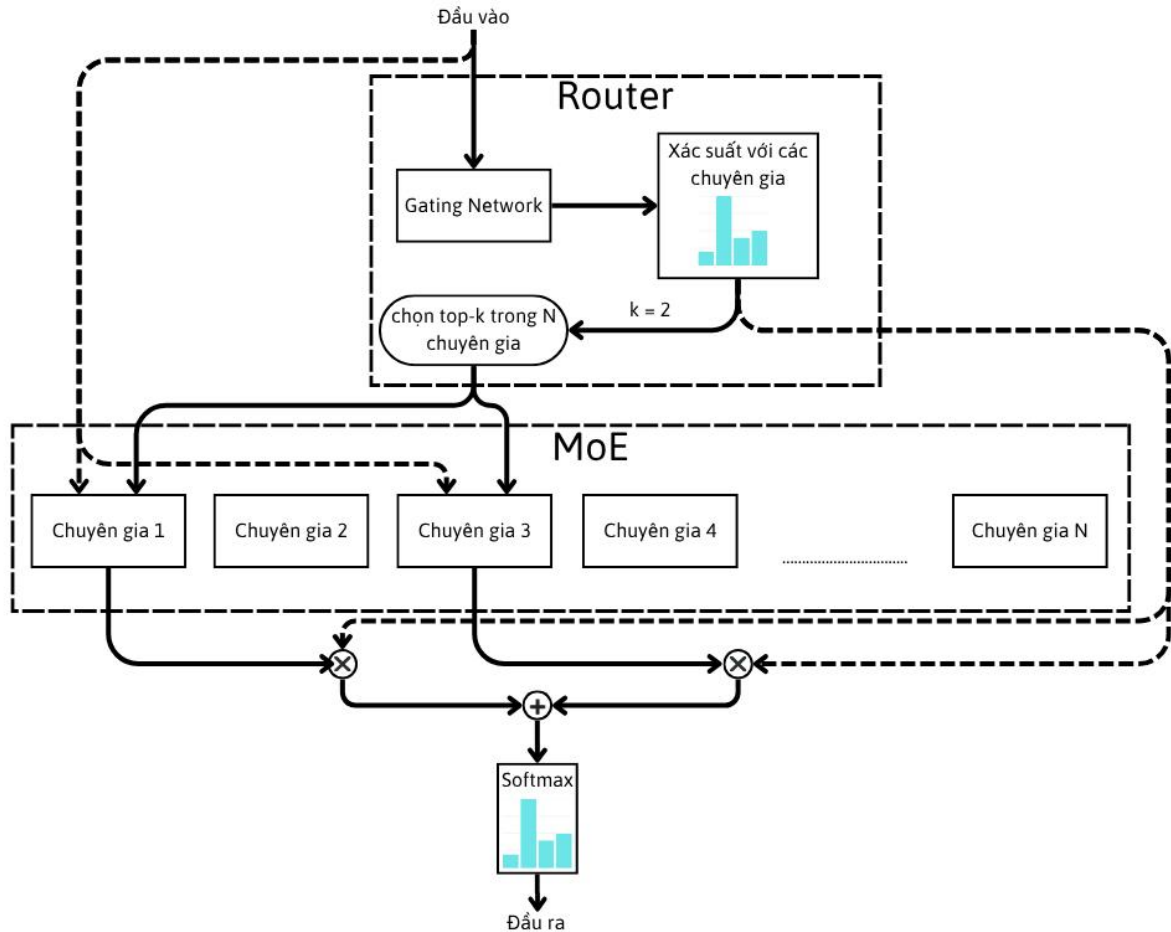


## ***Hình 2.2 Mô tả cơ chế Attention***

### **2.1.2 Tổng quan về mạng kết hợp chuyên gia**

Trong bối cảnh AI ngày càng phức tạp, một trong những hướng tiếp cận hiệu quả để cải thiện hiệu suất mô hình là sử dụng mạng kết hợp chuyên gia (Mixture of Experts - MoE). Đây là một kiến trúc được thiết kế nhằm chia nhỏ bài toán thành các phần nhỏ hơn và giao từng phần cho các mô-đun chuyên biệt, hay còn gọi là “chuyên gia” (experts). Khác với mô hình truyền thống, trong đó toàn bộ mạng phải xử lý toàn bộ thông tin đầu vào và phải học toàn bộ kiến thức, MoE chỉ kích hoạt một số chuyên gia nhất định cho từng task, giúp tối ưu hóa hiệu suất tính toán mà vẫn đảm bảo chất lượng đầu ra [3]. Ví dụ trong bài toán này sẽ có mỗi chuyên gia trong các khoa hay phòng ban khác nhau như một chuyên gia về chuẩn đoán, một chuyên gia về tư vấn hay một chuyên gia về đưa ra phương pháp điều trị.

Trong lĩnh vực NLP hay LLM, MoE đặc biệt hữu ích khi xử lý các tác vụ có tính chuyên biệt cao. Một chatbot tổng quát có thể cần phải hiểu nhiều lĩnh vực khác nhau, nhưng một chatbot y tế, chẳng hạn như chatbot tư vấn nha khoa, sẽ yêu cầu kiến thức chuyên môn sâu hơn về bệnh lý, phương pháp điều trị và thuật ngữ y khoa. Việc sử dụng MoE cho phép hệ thống học cách kích hoạt các chuyên gia phù hợp, chẳng hạn như chuyên gia về chuẩn đoán bệnh răng miệng, chuyên gia về các thủ thuật nha khoa, hoặc chuyên gia về tư vấn chăm sóc răng miệng. Điều này giúp chatbot có thể phản hồi chính xác hơn dựa trên câu hỏi của người dùng, thay vì sử dụng một mô hình tổng quát không có sự phân loại chuyên môn. Quan sát hình 2.3 để hiểu rõ hơn về cơ chế hoạt động mạng kết hợp chuyên gia:



**Hình 2.3 Mô tả mạng kết hợp chuyên gia**

Mô hình DeepSeek-R1, được phát triển trên nền tảng MoE, là một trong những kiến trúc hiện đại áp dụng kỹ thuật này để cải thiện hiệu suất và giảm chi phí tính toán. Bằng cách chỉ kích hoạt một số chuyên gia trong mỗi lần xử lý, DeepSeek-R1 không chỉ tối ưu hóa tốc độ suy luận mà còn giúp mô hình hoạt động hiệu quả hơn trên các tác vụ đòi hỏi kiến thức chuyên ngành.

Tóm lại, sự kết hợp giữa Transformer và mạng MoE đã mở ra những hướng tiếp cận mới trong việc xây dựng chatbot, đặc biệt là trong các lĩnh vực chuyên môn như y tế. Trong đề tài này, việc ứng dụng DeepSeek vào bài toán chatbot tư vấn nha khoa không chỉ giúp cải thiện độ chính xác mà còn đảm bảo rằng hệ thống có thể phản hồi phù hợp với ngữ cảnh y khoa, đồng thời giảm thiểu chi phí tính toán và nâng cao trải nghiệm người dùng.

## 2.2 Giới thiệu mô hình DeepSeek

DeepSeek là một trong những dòng mô hình ngôn ngữ lớn (LLM) tiên tiến, được thiết kế để tối ưu hóa hiệu suất suy luận và tiết kiệm chi phí huấn luyện so với các mô hình

khác hiện nay. Xuất phát từ nhu cầu xây dựng các hệ thống AI mạnh mẽ, hiệu suất cao và hiệu quả, DeepSeek đã nhanh chóng khẳng định vị thế trong cộng đồng nghiên cứu với kiến trúc tối ưu, tận dụng cả mô hình truyền thống và các kỹ thuật chuyên sâu để nâng cao khả năng xử lý ngôn ngữ tự nhiên của mô hình.

Các phiên bản của DeepSeek được phát triển nhằm cải thiện cả độ chính xác và hiệu suất tính toán, đặc biệt tập trung vào việc giảm tải bộ nhớ trong quá trình suy luận để nghiên cứu không tốn quá nhiều tài nguyên và triển khai không quá tốn kém. Đáng chú ý, DeepSeek-R1, phiên bản chính được sử dụng trong đề tài này, không chỉ kế thừa những ưu điểm từ các thế hệ trước mà còn được tinh chỉnh để có khả năng suy luận và phù hợp hơn với các ứng dụng chuyên biệt, bao gồm chatbot tư vấn trong lĩnh vực y tế.

### 2.2.1 Lịch sử phát triển và các phiên bản của DeepSeek

Sự phát triển của DeepSeek phản ánh xu hướng tiến hóa của các mô hình ngôn ngữ lớn, từ các kiến trúc thuần dense sang Mixture-of-Experts (MoE), giúp cân bằng giữa hiệu suất và chi phí tính toán.

**DeepSeek 67B** là một mô hình ngôn ngữ lớn mã nguồn mở được giới thiệu trong bài báo "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism" đã đánh dấu bước phát triển quan trọng khi áp dụng MoE vào LLM để cải thiện khả năng tổng quát hóa của mô hình giúp mô hình chuyên sâu hơn [11]. Mô hình này có 67 tỷ tham số và được huấn luyện trên một tập dữ liệu gồm 2 nghìn tỷ token. Kết quả đánh giá cho thấy DeepSeek 67B vượt trội hơn so LLaMA-2 70B của Meta AI trên nhiều tiêu chuẩn, đặc biệt trong các lĩnh vực lập trình, toán học và lý luận [9] [11]. Hơn nữa, phiên bản DeepSeek 67B Chat thể hiện hiệu suất vượt trội so với GPT-3.5 trong các đánh giá mở. Tuy nhiên, chi phí tính toán hay tài nguyên vẫn là một rào cản đáng kể, khiến việc triển khai trên quy mô lớn gặp khó khăn.

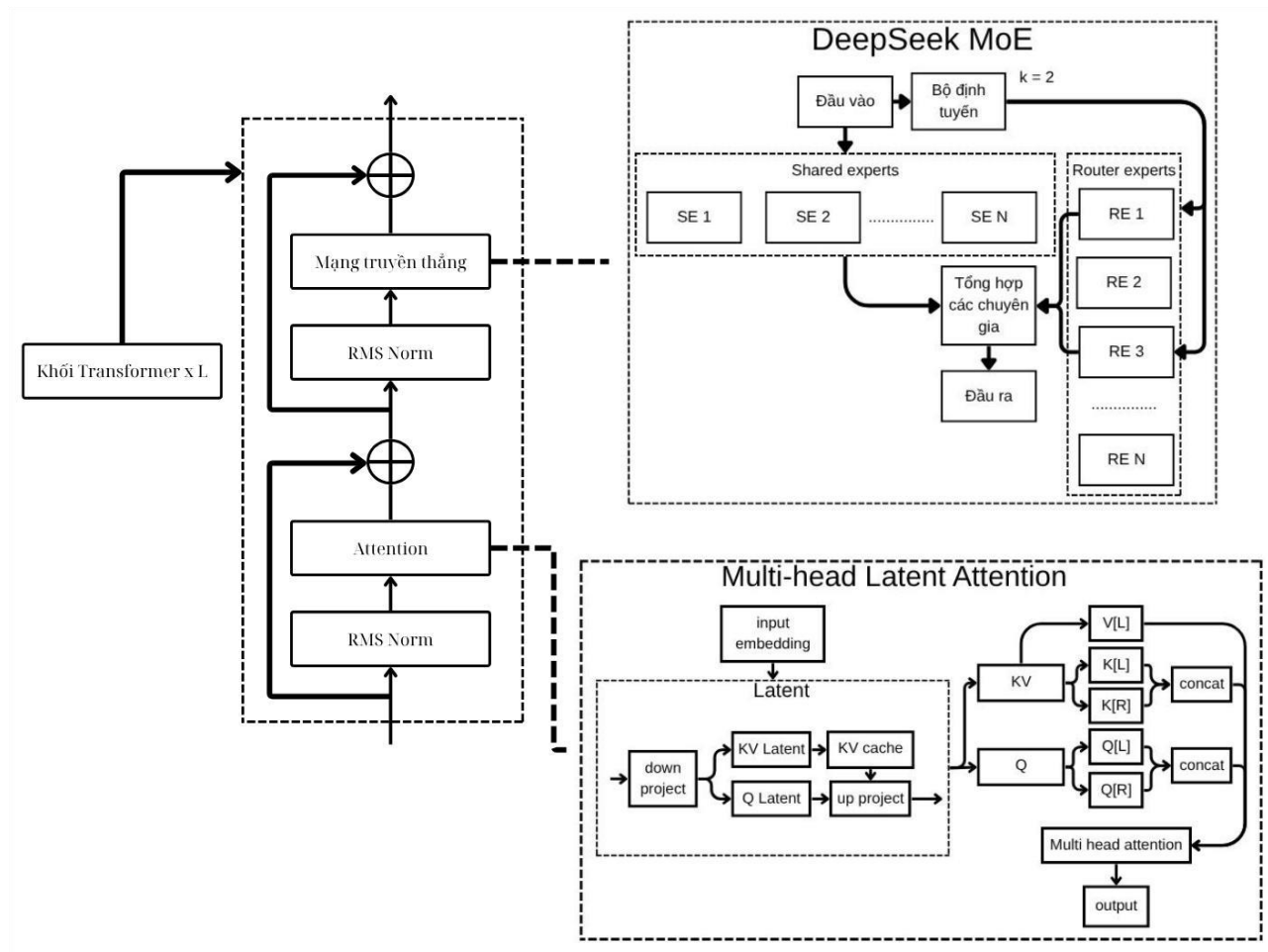
**DeepSeek-V2** là một mô hình ngôn ngữ Mixture-of-Experts (MoE) mạnh mẽ, được thiết kế để tối ưu hóa cả về hiệu quả huấn luyện và suy luận đã mang đến bước nhảy vọt với kiến trúc Multi-head Latent Attention (MLA) và DeepSeekMoE, giúp giảm kích thước bộ nhớ đệm (KV cache) tới 93.3%, đồng thời tối ưu hóa throughput sinh văn bản lên 5.76 lần so với DeepSeek 67B [12]. Những cải tiến này không chỉ giúp DeepSeek-V2 đạt được độ chính xác cao hơn trên các benchmark như MMLU mà còn làm nền tảng

cho sự ra đời của các phiên bản tối ưu hơn.

**DeepSeek-R1**, phiên bản chính được sử dụng trong đề tài, là sự kết hợp giữa hiệu suất tính toán và tính linh hoạt, đặc biệt trong bối cảnh fine-tune cho các ứng dụng cụ thể như chatbot tư vấn y tế. DeepSeek-R1 thừa hưởng kiến trúc MoE của DeepSeek-V2 nhưng được tinh chỉnh để cải thiện khả năng xử lý hội thoại dài, tăng cường khả năng hiểu ngữ cảnh và giảm độ trễ trong suy luận [5].

### 2.2.2 Kiến trúc và các thành phần chính

DeepSeek-R1 duy trì kiến trúc Transformer nhưng tích hợp nhiều cải tiến quan trọng để tối ưu hiệu suất cho mô hình. Trong đó, hai thành phần chính tạo nên sự khác biệt của DeepSeek-R1 so với các mô hình truyền thống là **Multi-head Latent Attention (MLA)** và **DeepSeekMoE** được thể hiện rõ ràng với hình 2.4 dưới đây:



**Hình 2.4 Kiến trúc DeepSeek**

#### a) Multi-head Latent Attention (MLA)

Trong quá trình phát triển các mô hình ngôn ngữ lớn, một trong những thách thức đối

với tất cả mọi mô hình ngôn ngữ lớn chính là chi phí bộ nhớ ngày càng gia tăng theo độ dài của chuỗi đầu vào của mô hình. Vì vậy, DeepSeek-R1 đã đưa ra cơ chế Multi-head Latent Attention (MLA) như một lời giải hiệu quả và giải pháp nhằm khắc phục sự congest của bộ đệm Key-Value (KV cache) vốn luôn tồn tại trong các mô hình attention truyền thống [12] [5]. Thay vì lưu trữ đầy đủ các key và value cho toàn bộ chuỗi token như trong cơ chế Multi-head Attention (MHA), MLA nén thông tin thông qua một phép biến đổi hạng thấp, vừa đảm bảo tính đại diện của dữ liệu, vừa giảm thiểu tài nguyên bộ nhớ cần thiết.

Về mặt cơ chế ở Hình 2.4, MLA triển khai với hai giai đoạn truyền thông tin tách biệt nhau là từ token đến latent và từ latent trở lại token [12] [5]. Ở giai đoạn đầu, mỗi latent vector sẽ thông qua  $Q_L$  – Truy vấn ngữ cảnh (Context Query) để học cách tích lũy thông tin từ toàn bộ các key-value đã được nén của các chuỗi token. Quá trình này giúp latent đóng vai trò như là một bộ nhớ ngữ nghĩa gọn nhẹ và có thể lưu trữ được phiên bản cô đọng của toàn bộ ngữ cảnh. Tiếp theo, để phục vụ cho quá trình sinh văn bản thì các token đầu ra sử dụng truy vấn  $Q_R$  – Đọc truy vấn (Read Query) để “đọc” lại thông tin từ các latent vector nói trên, phục hồi đầy đủ các tín hiệu ngữ nghĩa cần thiết để có thể tính attention hiệu quả và tốt nhất.

Điểm đặc biệt của MLA không chỉ nằm ở việc giúp giảm chi phí lưu trữ, mà còn ở khả năng duy trì chất lượng của mô hình khi xử lý ngữ cảnh dài. Điều này trở nên đặc biệt có ý nghĩa trong các ứng dụng yêu cầu khả năng ghi nhớ dài hạn như hội thoại y tế. Thực nghiệm cho thấy, ngay cả khi dung lượng KV cache giảm trên 90%, hiệu quả của mô hình vẫn tương đương, thậm chí nó còn cải thiện về tốc độ suy luận [12] [5]. Hơn thế, để mô hình giữ được tính nhạy với vị trí của token mà không làm tăng độ phức tạp của phép nén, DeepSeek-R1 sử dụng cơ chế Decoupled Rotary Position Embedding (RoPE) [12]. Bằng cách tách riêng các tín hiệu về vị trí ra khỏi phần key-value nén, MLA vừa tối ưu hóa được băng thông bộ nhớ, vừa duy trì được đặc tính thứ tự yếu tố quan trọng trong xử lý chuỗi.

Vì vậy, cơ chế MLA không đơn thuần là một thuật toán tối ưu hóa khả năng tính toán mà còn mở ra nhiều hướng triển khai các mô hình ngôn ngữ với khả năng mở rộng tốt hơn trong điều kiện phần cứng hạn chế. Việc loại bỏ sự phụ thuộc vào cấu trúc attention full-KV truyền thống là một bước tiến quan trọng, không chỉ cho DeepSeek-R1 mà còn



cho toàn bộ cộng đồng phát triển mô hình LLM hiệu quả.

### **b) DeepSeekMoE: Mạng kết hợp chuyên gia hiệu quả**

Trong mạng DeepSeekMoE ở Hình 2.4, có thể thấy đầu vào  $h_t$  được đưa vào bộ định tuyến (router) để đưa ra lựa chọn một số chuyên gia thích hợp cho bài toán, trong khi một nhóm chuyên gia chia sẻ (shared expert) sẽ luôn được kích hoạt nhằm đảm bảo thông tin nền tảng, cơ bản không bị bỏ sót [5]. Tư tưởng này trực tiếp kế thừa từ phiên bản DeepSeek-V2, song DeepSeek-R1 tinh chỉnh cách chọn lọc chuyên gia sao cho chỉ một phần nhỏ mô hình được kích hoạt, qua đó tiết kiệm tài nguyên tính toán mà vẫn duy trì độ chính xác của mô hình.

Sự xuất hiện của khối “ $Top - K_r$ ” thể hiện rõ cách cơ chế định tuyến chuyên gia quyết định nhóm chuyên gia nào sẽ được tham gia suy luận [5]. Cơ chế này không chỉ căn cứ dựa trên nội dung mà còn truy vấn và sử dụng chiến lược cân bằng tải, nhằm tránh tình trạng một vài chuyên gia bị quá tải hoặc không được tinh chỉnh và huấn luyện đầy đủ. Chính sự cân bằng này giúp việc suy luận vẫn trơn tru khi mô hình được fine-tune thành chatbot tư vấn nha khoa, nơi mỗi truy vấn có thể yêu cầu chuyên môn rất khác nhau, từ chăm sóc răng miệng cơ bản đến tư vấn phẫu.

Nhìn vào Hình 2.4, có thể thấy mỗi chuyên gia được kết nối với  $h_t$  và đóng góp đầu ra vào  $u_t$  thông qua các phép cộng dư (skip connection), bảo toàn thông tin gốc và hạn chế việc mô hình trở nên “quá chuyên sâu” vào một hướng nhất định [5]. Việc kết hợp cả “Routed Expert” và “Shared Expert” trong cùng một tầng còn cho phép DeepSeekMoE vừa có khả năng học đặc trưng tổng quát của dữ liệu, vừa đáp ứng tốt các truy vấn chuyên biệt hay học cục bộ, đặc biệt hữu ích cho chatbot chuyên ngành như tư vấn nha khoa.

Tổng thể, Mạng DeepSeekMoE minh họa rõ triết lý “tối giản nhưng hiệu quả” khi phân bổ chuyên gia và các chuyên môn khác nhau, hạn chế kích hoạt dư thừa các chuyên gia, đồng thời duy trì cơ chế định tuyến linh hoạt dựa trên mục tiêu cân bằng giữa hiệu năng suy luận và tính chuyên sâu của mô hình. Những cải tiến này làm nổi bật giá trị thực tiễn của DeepSeek-R1 trong các bài toán đòi hỏi khả năng chuyên gia cao, đặc biệt khi tài nguyên tính toán hoặc dung lượng xử lý bị giới.

### **2.2.3 Các cải tiến về hiệu năng**

DeepSeek-R1 không chỉ kế thừa những ưu điểm từ các phiên bản trước mà còn được tối ưu hóa đáng kể nhằm nâng cao hiệu suất và khả năng ứng dụng thực tế. Nhờ tích hợp cơ chế Multi-head Latent Attention (MLA) và DeepSeekMoE, mô hình có thể duy trì độ chính xác cao trong khi sử dụng ít tài nguyên tính toán hơn, giúp giảm đáng kể chi phí suy luận là một yếu tố quan trọng khi triển khai trên các hệ thống thực tế với giới hạn về tài nguyên. Bên cạnh đó, DeepSeek-R1 được thiết kế để tối ưu hóa khả năng xử lý hội thoại dài bằng cách cải tiến bộ nhớ đệm và cơ chế định tuyến thông minh, từ đó giúp chatbot duy trì ngữ cảnh xuyên suốt và hạn chế tình trạng lãng quên thông tin trong các cuộc tư vấn kéo dài. Đặc biệt, mô hình được tinh chỉnh chuyên sâu cho ngôn ngữ chuyên ngành nha khoa, cho phép chatbot hiểu rõ các thuật ngữ y khoa và đưa ra phản hồi chính xác, phù hợp hơn so với các mô hình tổng quát. Những cải tiến này giúp DeepSeek-R1 trở thành một lựa chọn mạnh mẽ và hiệu quả cho các ứng dụng chatbot trong lĩnh vực y tế.

Theo các kết quả từ tài liệu DeepSeek-V2, mô hình tiết kiệm được khoảng 42.5% chi phí huấn luyện và giảm 93.3% bộ đệm KV, đồng thời tăng thông lượng sinh văn bản lên tới 5.76 lần so với phiên bản DeepSeek 67B [12]. Những con số này không chỉ khẳng định hiệu quả của các cải tiến kỹ thuật mà còn tạo điều kiện thuận lợi cho việc xử lý hội thoại dài, nơi yêu cầu duy trì ngữ cảnh xuyên suốt và giảm thiểu hiện tượng “lãng quên” thông tin trong các cuộc tư vấn kéo dài. Với DeepSeek-R1 đạt hiệu suất vượt trội trong các bài kiểm tra toán học, với điểm số 79.8% trên AIME 2024 và 97.3% trên MATH-500, cao hơn so với OpenAI o1-1217 [5]. Ngoài ra, chi phí huấn luyện DeepSeek-V3, mô hình tiền nhiệm, ước tính khoảng 11.2 triệu USD, thấp hơn đáng kể so với 123.2 triệu USD của Llama 3 405B [13].

DeepSeek-R1 là sự kết hợp giữa tính hiệu quả và độ chính xác cao, phù hợp với các bài toán đòi hỏi khả năng hiểu ngữ cảnh sâu như chatbot tư vấn nha khoa. Với những cải tiến đáng kể về kiến trúc và hiệu năng, DeepSeek-R1 không chỉ giúp chatbot trả lời chính xác hơn mà còn tối ưu hóa chi phí vận hành, tạo tiền đề cho việc ứng dụng rộng rãi trong thực tế.

### 2.3 Cơ chế hoạt động và ưu nhược điểm

Trong hệ thống chatbot tư vấn nha khoa dựa trên nền tảng DeepSeek, sự kết hợp giữa Multi-head Latent Attention (MLA) và cơ chế phân luồng chuyên gia (fine-grained

expert routing) không chỉ là giải pháp để tối ưu hóa bộ nhớ mà còn là chìa khóa nâng cao hiệu suất thực thi. Thay vì lưu trữ toàn bộ KV cache như trong các mô hình truyền thống, MLA sẽ thực hiện nén và hợp nhất thông tin quan trọng thành các vector tiềm ẩn có kích thước nhỏ gọn. Cơ chế này đã được chứng minh là giúp giảm dung lượng bộ nhớ sử dụng trong quá trình suy diễn đến mức giảm trên 90%, qua đó tăng tốc độ phản hồi khi xử lý các chuỗi dữ liệu dài hoặc có số lượng token lớn [12].

Những cải tiến về hiệu năng được thể hiện rõ qua các chỉ số định lượng: so với phiên bản DeepSeek 67B thì DeepSeek-R1 tiết kiệm được 42.5% chi phí huấn luyện và tăng thông lượng tạo văn bản lên tới 5.76 lần [5]. Điều này cho thấy mô hình không chỉ đạt được hiệu quả về mặt tính toán mà còn tối ưu hóa chi phí vận hành là một yếu tố then chốt khi triển khai trong các hệ thống thực tế với nguồn tài nguyên hạn chế.

Song song với việc tối ưu bộ nhớ và tốc độ suy diễn, cơ chế phân luồng chuyên gia được thiết kế nhằm khai thác tối đa khả năng chuyên môn của từng “chuyên gia”. Thay vì áp dụng một mô hình đồng nhất cho toàn bộ dữ liệu, hệ thống sử dụng một chiến lược định tuyến tinh vi để gán mỗi token đầu vào cho chuyên gia phù hợp nhất. Qua đó, các chuyên gia có thể được huấn luyện và tinh chỉnh chuyên sâu theo từng miền ứng dụng cụ thể, như chăm sóc răng miệng hay tư vấn các thủ thuật nha khoa. Nhờ khả năng định tuyến thông minh, mô hình hạn chế được sự trùng lặp trong xử lý và tránh tình trạng “sập định tuyến” là một vấn đề thường gặp khi số lượng dữ liệu đầu vào tăng đột biến.

Cơ chế hoạt động của hệ thống được xây dựng trên nền tảng hai giai đoạn chính. Ở giai đoạn đầu, MLA thực hiện việc nén thông tin từ toàn bộ KV cache thành các vector tiềm ẩn nhỏ gọn, giảm thiểu sự phụ thuộc vào bộ nhớ lớn trong quá trình suy diễn. Ở giai đoạn tiếp theo, cơ chế phân luồng chuyên gia định tuyến các token đầu vào tới các chuyên gia phù hợp, cho phép hệ thống khai thác tốt hơn các mối quan hệ ngữ cảnh phức tạp trong hội thoại. Nhờ đó, chatbot không chỉ duy trì được mạch đối thoại xuyên suốt mà còn đảm bảo độ chính xác cao trong các phản hồi, đặc biệt đối với các truy vấn đòi hỏi kiến thức chuyên môn sâu như trong lĩnh vực nha khoa.

Về mặt ưu điểm, giải pháp DeepSeek cho thấy khả năng mở rộng vượt trội khi xử lý các chuỗi dữ liệu dài, giúp giảm đáng kể vấn đề tràn bộ nhớ và tối ưu hóa tài nguyên tính toán. Đồng thời, khả năng định tuyến chuyên gia linh hoạt tạo điều kiện cho quá trình

fine-tuning chuyên biệt, nâng cao chất lượng phản hồi cho các ứng dụng tư vấn chuyên ngành. Tuy nhiên, hệ thống cũng đối mặt với một số thách thức như việc cân bằng tải giữa các chuyên gia khi số lượng dữ liệu tăng đột biến, cũng như độ phức tạp trong thiết kế chiến lược định tuyến để tránh tình trạng một số chuyên gia bị quá tải. Những hạn chế này đòi hỏi một sự tối ưu hóa cẩn thận trong quá trình huấn luyện và triển khai để đảm bảo hệ thống luôn vận hành ổn định và hiệu quả.

## **2.4 Ứng dụng của DeepSeek trong các bài toán NLP**

Trong bối cảnh phát triển mạnh mẽ của các mô hình ngôn ngữ tiên tiến, DeepSeek đã được chứng minh là một công cụ hiệu quả trong việc giải quyết nhiều bài toán NLP đa dạng. Cơ chế hoạt động của DeepSeek, với kiến trúc tích hợp giữa Multi-head Latent Attention và DeepSeekMoE, cho phép mô hình khai thác tối đa thông tin ngữ cảnh và xử lý dữ liệu đầu vào với số lượng lớn, từ đó đạt được hiệu suất vượt trội trong các tác vụ như dịch máy, tóm tắt văn bản, và phân tích cảm xúc. Những cải tiến này không chỉ tạo ra hiệu quả về mặt tính toán mà còn nâng cao chất lượng kết quả, nhờ khả năng tập trung vào các đặc trưng quan trọng và loại bỏ nhiễu không mong muốn.

Các nghiên cứu đã triển khai ứng dụng DeepSeek trong nhiều trường hợp điển hình của các bài toán xử lý ngôn ngữ tự nhiên, từ việc cải thiện độ chính xác trong các hệ thống trả lời tự động đến việc tối ưu hóa quy trình giao tiếp giữa người và máy. Trong một số nghiên cứu cụ thể, các ứng dụng của DeepSeek đã được tích hợp vào các hệ thống chatbot tư vấn, giúp tăng cường khả năng hiểu và phản hồi theo ngữ cảnh của người dùng. Những hệ thống này, khi được huấn luyện trên tập dữ liệu phong phú và đa dạng, đã thể hiện khả năng xử lý tốt các tình huống phức tạp, từ các câu hỏi thông thường cho đến những vấn đề chuyên sâu đòi hỏi sự hiểu biết về ngữ cảnh và kiến thức chuyên môn.

Tiềm năng chuyển giao công nghệ từ DeepSeek sang lĩnh vực y học là một hướng đi mới mẻ và đầy hứa hẹn. Trong môi trường chăm sóc sức khỏe răng miệng, nơi mà sự chính xác và hiểu biết sâu sắc về ngôn ngữ chuyên ngành là yếu tố sống còn, việc áp dụng các mô hình như DeepSeek có thể tạo ra các giải pháp tư vấn tự động thông minh, hỗ trợ quá trình chẩn đoán và hướng dẫn điều trị một cách hiệu quả. Sự kết hợp giữa các dữ liệu chuyên sâu từ ngành nha khoa và khả năng xử lý ngôn ngữ tự nhiên mạnh mẽ của DeepSeek không chỉ giúp chatbot hiểu và phản hồi chính xác các câu hỏi của bệnh nhân mà còn tạo điều kiện cho việc tích hợp kiến thức chuyên gia vào hệ thống, từ đó

cung cấp những gợi ý và chẩn đoán có tính cá nhân hóa cao.

Việc chuyển giao công nghệ này đòi hỏi một quá trình fine-tuning mô hình cẩn thận trên dữ liệu chuyên ngành, qua đó không chỉ khai thác tối đa những ưu điểm của DeepSeek mà còn đảm bảo rằng các phản hồi được đưa ra luôn phù hợp với ngữ cảnh và yêu cầu riêng biệt của lĩnh vực nha khoa. Qua đó, dự án không chỉ tạo ra một sản phẩm có giá trị ứng dụng thực tiễn cao mà còn mở ra hướng nghiên cứu mới, khuyến khích sự giao thoa giữa các lĩnh vực trí tuệ nhân tạo và chăm sóc sức khỏe.

## 2.5 Phương pháp SFT và OPRO

Quá trình huấn luyện một mô hình ngôn ngữ lớn (LLM) phục vụ cho các tác vụ tư vấn yêu cầu mô hình không chỉ hiểu ngôn ngữ mà còn phản hồi theo hướng phù hợp với mong đợi thực tế của con người. Để đạt được điều này, hai giai đoạn huấn luyện thường được áp dụng là Supervised Fine-Tuning (SFT) và Odds Ratio Preference Optimization (ORPO), giúp mô hình học từ dữ liệu gán nhãn cũng như từ phản hồi mang tính chất so sánh của người dùng.

Supervised Fine-Tuning là bước đầu tiên trong quy trình tinh chỉnh, nơi mô hình được huấn luyện trên các cặp dữ liệu đầu vào – đầu ra (prompt – completion) đã được kiểm duyệt. Đây là quá trình học có giám sát nhằm giúp mô hình bắt chước hành vi phản hồi đúng.

Với một chuỗi đầu vào  $x$  và đầu ra mong muốn  $y = (y_1, y_2, \dots, y_T)$ , mục tiêu là cực tiểu hóa hàm mất mát xác suất chéo giữa phân phối dự đoán của mô hình  $p_\theta$  và nhãn thực tế:

$$L_{\text{SFT}}(\theta) = -\sum_{t=1}^T \log p_\theta(y|y_{<t}, x) \quad [14] \quad (2.2)$$

Cách huấn luyện này được sử dụng phổ biến trong các mô hình như InstructGPT, cho phép mô hình nắm bắt được cấu trúc và ngữ nghĩa của câu trả lời phù hợp.

Mặc dù SFT giúp mô hình trả lời một cách hợp lý, nó không đảm bảo rằng các phản hồi thực sự được ưa thích bởi con người trong các tình huống phức tạp. Do đó, giai đoạn tối ưu hóa theo sở thích được đưa vào. Trong khi các phương pháp như RLHF sử dụng mô hình phản thưởng và học tăng cường để cập nhật mô hình, ORPO là một tiếp cận mới nhẹ hơn, khai thác trực tiếp tỷ lệ xác suất sinh phản hồi từ mô hình mà không cần huấn

luyện thêm bất kỳ mô hình phụ nào [15].

Cụ thể, với một prompt  $x$  và hai phản hồi: một phản hồi được ưa thích  $y^+$ , và một phản hồi kém hơn  $y^-$ , ORPO tối ưu hóa hàm mất mát dựa trên tỷ số log-likelihood (odds ratio):

$$L_{ORPO}(\theta) = -\log\left(\frac{p_{\theta}(y^+|x)}{p_{\theta}(y^-|x)}\right) = -[\log p_{\theta}(y^+|x) - \log p_{\theta}(y^-|x)] \quad [16] \quad (2.3)$$

Hàm mất mát này thúc đẩy mô hình tăng khả năng sinh ra phản hồi được đánh giá cao hơn, đồng thời giảm xác suất sinh phản hồi kém, mà không cần sử dụng mô hình phần thưởng hay chính sách cũ như trong PPO. Điều này làm cho ORPO trở thành một phương pháp tinh chỉnh hiệu quả và tiết kiệm tài nguyên hơn trong thực tế triển khai [15].

Việc phối hợp giữa SFT và ORPO tạo ra một quy trình huấn luyện mạnh mẽ: SFT đặt nền tảng ngôn ngữ và hành vi phản hồi cơ bản, còn ORPO giúp mô hình điều chỉnh phản hồi theo hướng phù hợp hơn với đánh giá chủ quan của người dùng. Hai phương pháp này được xem như một phần quan trọng trong pipeline huấn luyện hiện đại cho các mô hình ngôn ngữ lớn hỗ trợ tương tác trong các ứng dụng như chatbot tư vấn nha khoa.

Chương này đã làm rõ nền tảng kỹ thuật và mô hình mà nghiên cứu sử dụng. Kiến thức này là tiền đề cho chương 3, nơi sẽ tập trung vào việc thu thập và xử lý dữ liệu chuyên biệt để huấn luyện mô hình DeepSeek.

## CHƯƠNG 3 : THU THẬP, XỬ LÝ DỮ LIỆU NHA KHOA

Sau khi đã xác định mô hình cần triển khai, chương này đi sâu vào quá trình xây dựng tập dữ liệu cho chatbot nha khoa. Nội dung bao gồm nguồn dữ liệu, kỹ thuật tiền xử lý, tăng cường dữ liệu, gán nhãn, và đánh giá chất lượng nhằm đảm bảo đầu vào đạt độ chính xác và phù hợp với ngữ cảnh chuyên ngành.

### 3.1 Giới thiệu nguồn và đặc thù dữ liệu nha khoa

Trong kỷ nguyên mà dữ liệu đóng vai trò cốt lõi cho mọi tiến bộ của trí tuệ nhân tạo, nhất là với mô hình ngôn ngữ lớn thì việc lựa chọn, phân tích và hiểu rõ đặc trưng của dữ liệu y tế nói chung và nha khoa nói riêng là yếu tố quyết định đến sự hiệu quả và độ tin cậy của các hệ thống ứng dụng AI, đặc biệt là những mô hình LLM có định hướng tương tác trực tiếp với con người như chatbot. Dữ liệu nha khoa, với vai trò là cầu nối giữa tri thức y khoa chuyên sâu và khả năng suy diễn của mô hình học sâu thì nó không đơn thuần chỉ là nguồn thông tin huấn luyện, tinh chỉnh mô hình mà nó còn phản ánh trực tiếp những thực tại phức tạp về ngôn ngữ, hành vi và chuyên môn trong ngành y hay trong nha khoa.

Một điểm đặc biệt trong lĩnh vực nha khoa là tính lưỡng diện của ngữ liệu với một mặt là chúng được cấu thành từ các thuật ngữ y học có cấu trúc chặt chẽ và mang tính chuyên biệt cao. Một mặt khác là ngữ cảnh tương tác giữa bác sĩ và bệnh nhân lại thường thiên hướng về ngôn ngữ tự nhiên, nhiều sắc thái cảm xúc và mang tính cá nhân hóa rõ nét. Do đó, bất kỳ mô hình LLM nào muốn đạt được hiệu năng tư vấn hiệu quả cao đều cần phải được huấn luyện trên tập dữ liệu phản ánh đầy đủ cả hai chiều là **chuẩn mực chuyên môn** và **thực tế ngữ dụng**.

Trong khuôn khổ đề tài này, dữ liệu được thu thập từ nhiều nguồn đáng tin cậy và đa dạng như từ các bài báo nghiên cứu chuyên sâu trên các nơi cung cấp nền tảng tri thức có hệ thống hóa và cập nhật trong ngành nha khoa, đến các đoạn hội thoại mô phỏng tương tác giữa bệnh nhân và bác sĩ, nơi biểu hiện rõ nhất ngôn ngữ thường nhật về cuộc hội thoại và các biến thể ngữ cảnh phù hợp với bệnh nhân để tăng tính cá nhân hóa. Bên cạnh đó, dữ liệu sẽ được trích xuất từ các cổng thông tin y tế, diễn đàn sức khỏe cộng đồng, các bài viết và các trang web chuyên sâu trong lĩnh vực nha khoa cũng được khai thác bằng các công cụ thu thập dữ liệu để nhằm đảm bảo bộ dữ liệu có độ phong phú và gần gũi với trải nghiệm người dùng thực tế. Và các định dạng dữ liệu được thu thập sẽ

được chuyển về đúng dạng văn bản thuần túy, đúng ngôn ngữ, ngữ cảnh.

Bên cạnh đó, không thể không nhắc đến vai trò của các kho dữ liệu mở như HuggingFace, Kaggle hay Google Dataset Search, nơi cung cấp những bộ dữ liệu quy mô lớn, cấu trúc, định dạng rõ ràng, giúp bổ sung thêm dung lượng dữ liệu, tăng tính đa dạng và ngữ cảnh cho bộ dữ liệu huấn luyện, tinh chỉnh.

Việc hiểu sâu sắc đặc trưng ngữ liệu không chỉ giúp mô hình LLM thích nghi tốt với thực tiễn mà còn tạo điều kiện thuận lợi cho việc thiết kế pipeline huấn luyện hợp lý như từ tiền xử lý, gán nhãn task, đến đánh giá và triển khai thực tế. Nhận thức rõ ràng về giá trị, cấu trúc, định dạng của dữ liệu cũng chính là bước đầu tiên để đảm bảo mô hình không chỉ “đúng về mặt kỹ thuật” mà còn “phù hợp về mặt xã hội” giúp đáp ứng được kỳ vọng của người sử dụng và đảm bảo tính an toàn trong các tình huống tư vấn y tế nhạy cảm.

### 3.2 Phương pháp thu thập dữ liệu

Trong bất kỳ hệ thống học sâu nào cũng hướng tới ứng dụng thực tiễn trong y học, chất lượng và đặc trưng của dữ liệu đầu vào luôn đóng vai trò then chốt trong việc quyết định hiệu quả cuối cùng của mô hình. Đặc biệt, đối với một chatbot tư vấn nha khoa là một lĩnh vực yêu cầu sự chính xác về mặt y học đồng thời phải đảm bảo tính thân thiện trong giao tiếp ngôn ngữ tự nhiên với bệnh nhân. Quá trình xây dựng bộ dữ liệu lớn và chất lượng không chỉ đơn thuần là thu thập thông tin, mà là một quá trình có tính chiến lược, được tổ chức một cách có hệ thống và chặt chẽ như một pipeline xử lý tri thức.

Trong nghiên cứu này, quá trình thu thập dữ liệu được thiết kế theo hướng tích hợp đa nguồn và đa phương thức để nhằm đảm bảo tối ưu đồng thời ba yếu tố cốt lõi: tính chuyên ngành, tính đa dạng và tính cập nhật. Đầu tiên, từ các nền tảng học thuật và cộng đồng mã nguồn mở như **HuggingFace**, **Kaggle**, **Google Dataset Search** là một nơi lưu trữ nhiều bộ dữ liệu đã được cộng đồng học thuật xác thực, thu thập, uy tín, chất lượng cao được dùng cho phục vụ nghiên cứu. Trong đó, các bộ dữ liệu y tế về tiếng anh và tiếng việt là chính do dữ liệu phục vụ cho nha khoa còn hạn chế hoặc chưa được tối ưu tốt. Tuy nhiên, các bộ dữ liệu y tế vẫn có các dữ liệu về nha khoa hay đơn giản là liên quan đến nha khoa nên ý tưởng là việc hoàn toàn có thể thu thập những dữ liệu liên quan đến nha khoa để xử lý cho phù hợp với mô hình LLM. Dữ liệu được thu thập thông qua các hàm lọc chuyên biệt để chỉ giữ lại các phần tử chứa thông tin liên quan mật thiết đến



lĩnh vực nha khoa hay đơn giản là dựa vào một số từ khóa về nha khoa để lấy những dữ liệu nha khoa. Bằng cách này, tập dữ liệu được đảm bảo có độ chọn lọc cao, tập trung đúng trọng tâm chủ đề nghiên cứu, đồng thời giảm thiểu nhiễu từ các lĩnh vực y học không liên quan khác.

Song song với nguồn dữ liệu cộng đồng, nghiên cứu còn triển khai chiến lược khai thác tri thức từ một số nền tảng học thuật, nghiên cứu như **arXiv**, **ResearchGate**, nơi xuất bản nhiều công trình nghiên cứu mới nhất về y học và nha khoa được thu thập bằng thư viện hoặc công cụ như **Web Scraper**, **arXiv**, **GoogleSearch**, **Requests**, **BeautifulSoup** hoặc **Custom Search JSON API** để lấy đường dẫn tệp tin tài liệu hoặc lấy văn bản trực tiếp. Các bài báo liên quan được trích xuất theo cấu trúc gồm tiêu đề, tóm tắt, danh sách tác giả và liên kết gốc, bài nghiên cứu gốc với các dạng dữ liệu như (.pdf), (.docx) và một số định dạng khác. Sau được thu thập và xử lý bằng các thư viện hoặc các framework có sẵn như **PyMuPDF**, **Gemini API** để chuyển về dạng văn bản thuần túy. Việc sử dụng dữ liệu học thuật giúp bổ sung cho mô hình một lớp tri thức sâu và nó không chỉ mang tính thống kê mà còn hàm chứa kiến thức chuyên môn có kiểm chứng từ các bài báo, nghiên cứu, từ đó tăng khả năng giải thích và đáng tin cậy trong phản hồi của chatbot nha khoa.

Một lớp dữ liệu khác là đến từ môi trường phi cấu trúc, được khai thác từ các trang web chuyên sâu, nội dung bài viết về nha khoa bằng các công cụ hoặc thư viện như **Web Scraper**, **arXiv**, **GoogleSearch**, **Requests**, **BeautifulSoup** hoặc **Custom Search JSON API**. Các bài viết từ blog của bác sĩ hay chuyên gia hàng đầu về nha khoa, trang chia sẻ kinh nghiệm chăm sóc răng miệng, hệ thống hỏi đáp sức khỏe răng hàm mặt được thu thập và xử lý có chọn lọc nhằm giúp tái cấu trúc thành ngữ liệu có thể sử dụng cho huấn luyện, tinh chỉnh mô hình ngôn ngữ lớn. Đây là nguồn dữ liệu giàu ngữ cảnh thực tế, góp phần quan trọng vào việc mô hình hóa các kiểu tương tác giữa người dùng và hệ thống tư vấn nha khoa.

Toàn bộ pipeline thu thập được kiểm soát chất lượng thông qua ba tiêu chí đánh giá chính là **độ phù hợp chuyên ngành**, **tính đa dạng ngôn ngữ**, và **tính xác thực của nguồn**. Mỗi mẫu dữ liệu đều trải qua bước rà soát bán tự động nhằm loại bỏ trùng lặp, kiểm tra tính đầy đủ ngữ cảnh, và xác nhận độ tin cậy của nguồn gốc. Các bộ dữ liệu từ các nguồn khác nhau cũng được so sánh chéo để phát hiện xung đột thông tin hoặc sai

lệch nội dung. Chỉ những mẫu vượt qua các bước kiểm định này mới được đưa vào pipeline huấn luyện, tinh chỉnh chính thức.

Nhờ vào hệ thống thu thập được thiết kế tỉ mỉ và có chiều sâu như trên, tập dữ liệu đầu vào không chỉ có tính học thuật mà còn giàu tính thực tiễn, phản ánh các vấn đề thường gặp trong chăm sóc nha khoa từ cả góc nhìn bác sĩ và người bệnh. Đây chính là nền tảng quan trọng, góp phần tạo nên hiệu quả cho quá trình fine-tune mô hình DeepSeek-R1 ở các chương tiếp theo, hướng tới mục tiêu xây dựng một chatbot có khả năng tư vấn y tế vừa chính xác, vừa linh hoạt trong tương tác với người dùng.

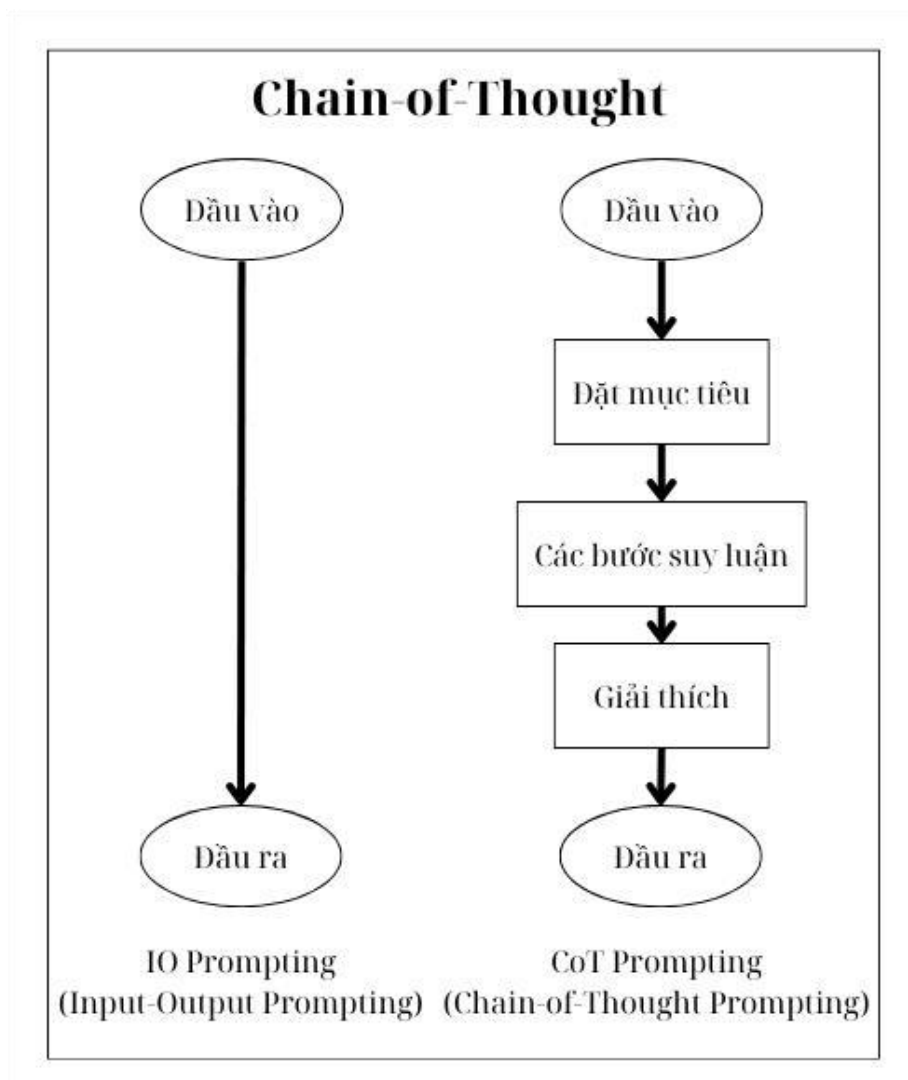
### 3.3 Tiền xử lý dữ liệu

**“The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge.”**

**– Stephen Hawking [17].** Câu nói ấy không chỉ là lời cảnh tỉnh, mà còn là định hướng tư tưởng cho toàn bộ tiến trình xử lý dữ liệu trong đề tài này. Trong lĩnh vực ứng dụng trí tuệ nhân tạo cho tư vấn y khoa, đặc biệt là nha khoa, nơi kiến thức chuyên môn không chỉ mang tính chuyên sâu mà còn luôn biến động, dữ liệu không sạch sẽ là mầm mống của sai lệch, của nhầm lẫn, và thậm chí, của tổn hại niềm tin nơi người sử dụng. Do đó, tiền xử lý không phải là giai đoạn kỹ thuật giản đơn, mà là bước chuyển hóa từ dữ liệu thô hỗn loạn thành tri thức có cấu trúc, có ngữ cảnh và có giá trị học thuật.

Trước tiên, với một bộ dữ liệu văn bản thô với các định dạng, ngôn ngữ khác nhau và nhiều lỗi khác khiến cho việc xử lý, học hỏi của mô hình ngôn ngữ lớn trở nên khó học hơn hay học sai. Vì vậy, để có một bộ dữ liệu không chỉ nhiều mà phải chất lượng cao và giúp mô hình học, suy luận tốt hơn thì người nghiên cứu sẽ thực hiện **tiền xử lý như một hành vi xây dựng tri thức**. Khởi nguồn từ các kho dữ liệu thực tế gồm: bài báo y học chuyên ngành, hội thoại giữa bác sĩ và bệnh nhân, nội dung từ các diễn đàn tư vấn sức khỏe, dữ liệu ban đầu mang trong mình sự phong phú về nội dung nhưng đồng thời cũng chứa đựng độ phức tạp cao về cấu trúc. Tính không nhất quán về định dạng, sự đa dạng trong phong cách diễn đạt tiếng Việt từ học thuật đến đời thường, và đặc biệt là những chông chéo ngữ nghĩa chuyên ngành là những rào cản lớn cản trở mô hình học sâu tiếp cận được tri thức thực sự. Với tinh thần đó, đề tài đã áp dụng **hệ thống xử lý ngữ nghĩa và biểu diễn thông tin bằng Gemini**. Cụ thể là gọi API mô hình Gemini 1.5 Flash do Google phát triển, có thể hiểu là multimodal - một mô hình ngôn ngữ đa phương thức được thiết kế để xử lý hiệu quả các tác vụ yêu cầu tốc độ cao và chi phí thấp, đồng

thời vẫn duy trì khả năng hiểu ngữ cảnh dài và xử lý nhiều loại dữ liệu đầu vào như một trục xương sống xuyên suốt toàn bộ pipeline tiền xử lý [18]. Khác với các phương pháp tiền xử lý truyền thống, Gemini không chỉ thực hiện chuẩn hóa hình thức, mà còn thực hiện quá trình **trích xuất tầng suy luận** từ dữ liệu là một bước quan trọng gọi là *kiến tạo chuỗi lý luận* hay Chain-of-Thought Induction, là một kỹ thuật giúp cải thiện khả năng lập luận LLMs bằng cách hướng dẫn nó giải quyết vấn đề theo từng bước logic [19]. Cách trên giúp mô hình chia nhỏ vấn đề và phân tích từng bước để đưa ra câu trả lời tốt nhất, hãy xem sơ đồ dưới đây để hiểu rõ hơn cách hoạt động:



**Hình 3.1 So sánh IO prompt với CoT prompt**

Tương tự, từng đoạn văn bản hội thoại, từng đoạn tư vấn nha khoa được chuyển hóa sang định dạng chuẩn gồm năm phần rõ ràng với ba chuỗi suy luận để giúp mô hình học tốt và đưa ra phản hồi tốt:

Thành phần	Mô tả	Ví dụ
Đầu vào	Câu hỏi, yêu cầu từ người dùng.	Sâu răng là gì?
Đặt mục tiêu	Xác định rõ mục tiêu cần đạt được khi trả lời câu hỏi.	Hiểu định nghĩa và bản chất của bệnh sâu răng và nắm được quá trình hình thành và tiến triển của sâu răng.
Các bước suy luận	Trình bày rõ các bước tư duy và phân tích hoặc lập luận logic để trả lời câu hỏi, bao gồm là kích hoạt các kiến thức nền, liên hệ nguyên nhân, kết quả, tổng hợp các nội dung.	1) Nhận diện loại câu hỏi (dạng định nghĩa). 2) Kích hoạt kiến thức về nha khoa. 3) Liên kết với nguyên nhân và triệu chứng. 4) Tổng hợp và tạo câu trả lời đầy đủ, dễ hiểu.
Giải thích	Cung cấp các phần và phân tích chi tiết và toàn diện nhằm làm rõ	Sâu răng là một bệnh phổ biến trong nha khoa, xảy ra khi vi khuẩn trong mảng bám trên răng phân hủy đường trong thức ăn và tạo axit. Axit

	<p>câu trả lời.</p> <p>Giải thích rõ nên đi sâu vào cơ chế, nguyên nhân, hậu quả và mối liên hệ giữa các yếu tố.</p>	<p>này ăn mòn men răng là một lớp ngoài cùng bảo vệ răng. Khi men răng bị phá vỡ, axit tiếp tục tấn công lớp ngà và cuối cùng là tủy, gây đau nhức và có thể dẫn đến mất răng nếu không điều trị.</p> <p>Nguyên nhân chính gồm vệ sinh răng miệng kém, ăn nhiều đường và không khám răng định kỳ.</p>
Đầu ra	<p>Tóm tắt câu trả lời cuối cùng thường là đoạn trả lời ngắn, súc tích chứa nội dung cốt lõi, mang tính áp dụng hay giải đáp trực tiếp.</p>	<p>Sâu răng là tình trạng vi khuẩn trong mảng bám răng tạo axit khi phân hủy đường, axit này ăn mòn men răng và gây ra các lỗ sâu. Nếu không điều trị, sâu răng có thể lan vào ngà và tủy, gây đau nhức, viêm tủy hoặc mất răng. Nguyên nhân chủ yếu là do vệ sinh răng kém, ăn nhiều đường và không khám răng định kỳ. Xin hãy đến phòng khám nha khoa để nhận điều trị sớm nhất có thể.</p>

***Bảng 3.1 Mô tả cách hoạt động chuỗi suy luận***

Sự biểu diễn logic này không chỉ giúp mô hình nhận diện thông tin như một khối dữ liệu, mà như một quá trình suy nghĩ, suy luận của con người trước khi trả lời câu hỏi và nó cũng là một điều kiện tiên quyết cho các mô hình ngôn ngữ lớn thực hiện tư duy phản biện và phản hồi mang tính lâm sàng. Nhất là trong nha khoa, đòi hỏi cần chính xác và cụ thể.

Tiếp theo là một bước giúp dữ liệu có đa dạng định dạng là **biến thông tin thành tri thức có cấu trúc**. Ngoài xử lý văn bản tự nhiên để mô hình sinh ra văn bản thuần túy, dữ liệu còn có thể bao gồm nhiều dạng đặc thù như dạng bảng để chẩn đoán lâm sàng hay bảng thống kê chi phí để nắm rõ thông tin quan trọng, biểu đồ nha khoa giúp có cái nhìn trực quan, công thức định lượng thuốc cho bác sỹ nha khoa. Những dạng dữ liệu này được xử lý bởi Gemini với khả năng **phân tích định dạng chuyên sâu**, chuyển đổi tự động thành dạng là **bảng markdown, biểu đồ biểu diễn bằng ngôn ngữ Python, Công thức dạng markdown**. Việc biểu diễn dưới dạng cấu trúc trên giúp mô hình học sâu có khả năng trích rút thông tin nhanh hơn, đồng thời học được cấu trúc tri thức, điều mà dữ liệu văn bản thuần túy không thể cung cấp đầy đủ. Ngoài ra, để giúp mô hình đưa ra câu trả lời chính xác với yêu cầu người dùng thì việc **gán nhãn đa chiều và tự động hóa bằng tri thức chuyên ngành** là việc cần thiết để cải thiện khả năng phản hồi mô hình và giúp mô phỏng khả năng MoE từ DeepSeek-R1 giúp tối ưu việc phản hồi. Ở tầng tiếp theo, dữ liệu được tiến hành gán nhãn nhưng không chỉ là nhãn lớp đầu ra, mà là **gán nhãn tri thức ba chiều** theo định dạng: [loại dữ liệu, nội dung, lĩnh vực]. Ví dụ như một bảng phân tích chỉ số răng – nướu: [Bảng, chỉ số nha khoa, chẩn đoán] hay một biểu đồ phân tích viêm nhiễm của sâu răng: [Biểu đồ, cơ chế viêm tủy răng, giáo dục sức khỏe]. Việc gán nhãn sẽ dựa vào từ khóa các nhãn trong dữ liệu và dùng kỹ thuật **Zero-shot classification** - là một kỹ thuật trong học máy không cần huấn luyện để phân lớp rồi gán nhãn nhanh cho dữ liệu. Phân loại này giúp mô hình hiểu rằng cùng là "đau răng", nhưng xuất hiện trong ngữ cảnh tư vấn sẽ khác hoàn toàn và cần dùng loại dữ liệu nào để đáp ứng yêu cầu người dùng. Đây là bước *contextual grounding* là giúp chatbot xây dựng phản hồi theo đúng ngữ cảnh chứ không chỉ dừng lại ở mức đối thoại phản xạ bằng cách đoán mò, dự đoán câu tiếp theo. Cuối cùng, thu được sẽ là kết quả của một pipeline học sâu tri thức hoá. Là dữ liệu sau khi qua chuỗi xử lý Gemini không còn là “đầu vào” thông thường, mà trở thành **hệ sinh thái tri thức có ngữ cảnh**, được tổ chức chặt chẽ, có thể suy luận và có thể học được [18]. Nhờ hệ thống này, mô hình DeepSeek khi được fine-tune sẽ không học theo cơ chế bắt chước mẫu hội thoại, mà sẽ học cách *tư duy* như một chuyên gia tư vấn nha khoa đưa ra phản hồi dựa trên bối cảnh, liên kết kiến thức, và diễn đạt rõ ràng.

Điều này đặt nền móng quan trọng cho mục tiêu cao nhất của đề tài: phát triển một chatbot tư vấn nha khoa không chỉ chính xác, mà còn *đáng tin cậy, hữu ích và mang lại*

*trải nghiệm tư vấn giống như con người thậm chí tốt hơn về mặt nhất quán và kiến thức hệ thống.*

**3.4 Prompt Engineering - giai đoạn tiền xử lý**

Trong hệ thống chatbot ứng dụng trí tuệ nhân tạo chuyên tư vấn nha khoa, chất lượng đầu vào không chỉ ảnh hưởng đến độ chính xác của phản hồi mà còn quyết định khả năng mô hình thích ứng với các tình huống phức tạp trong thực tế lâm sàng. Do đó, việc chuẩn hóa và kiến trúc hóa câu lệnh đầu vào hay còn gọi là Prompt Engineering nắm giữ vai trò trung tâm trong quá trình tiền xử lý dữ liệu.

Từ góc nhìn kỹ thuật, prompt không đơn thuần là một câu hỏi đầu vào mà là một cấu trúc ngữ cảnh mà là nơi hội tụ giữa tri thức chuyên ngành, mục tiêu giao tiếp và vai trò của hệ thống trong từng tương tác [20]. Để đảm bảo điều này, nhóm nghiên cứu đề xuất một định dạng prompt thống nhất, sử dụng hệ thống token đặc biệt nhằm tạo ranh giới rõ ràng giữa các vai trò (system, user, think, expert, assistant), cùng các thẻ html giúp tối ưu khi trả khai lên giao diện và tạo điều kiện thuận lợi cho quá trình huấn luyện lần triển khai. Ngoài ra, để tăng tính trực quan, các thành phần này có thể kết hợp biểu tượng (emoji) trong giai đoạn tiền xử lý như một lớp đánh dấu ngữ nghĩa giúp mô hình hiểu rõ hơn mục đích của từng đoạn. Quan sát bảng 3.2 ở dưới để hiểu rõ các vai trò và cách hoạt động:

Thành phần	Vai trò	Mô tả
system	Thiết lập ngữ cảnh hệ thống	Xác định phạm vi, yêu cầu nhiệm vụ, định hướng phản hồi
user	Đầu vào	Bao gồm các câu hỏi hay vấn đề cần được giải quyết
think	Kích hoạt chuỗi suy luận (CoT)	Gợi ý mô hình tự động phân tích vấn đề theo từng bước suy luận
expert	Ngữ cảnh chuyên gia	Mô phỏng vai trò chuyên gia và yêu cầu các kiến thức chuyên sâu

assistant	Phản hồi đầu ra	Tập hợp nội dung trả lời cuối cùng từ chatbot
-----------	--------------------	--

**Bảng 3.2 Mô tả vai trò trong cuộc hội thoại**

Như vậy, quá trình tiền xử lý trong giai đoạn Prompt Engineering không đơn thuần là làm sạch hay định dạng câu hỏi, mà còn là sự thiết kế chiến lược để tạo ra một hành lang ngữ nghĩa cho mô hình học sâu có thể “hiểu” như con người mà không chỉ là từ ngữ, mà cả mục tiêu, sắc thái và hành vi kỳ vọng [20]. Đây là điểm giao thoa giữa ngôn ngữ học tính toán, UX ngữ cảnh, và kỹ thuật triển khai ứng dụng AI hiện đại.

### **3.5 Tăng cường dữ liệu**

Một trong những thách thức thường gặp là dữ liệu huấn luyện ban đầu còn giới hạn về khối lượng và dù được xây dựng cẩn thận, vẫn mang tính chất khuôn mẫu, chưa phản ánh đầy đủ sự đa dạng trong cách đặt câu hỏi thực tế từ người dùng. Điều này dễ dẫn đến hiện tượng chatbot phản hồi kém hoặc không phù hợp khi gặp phải các biểu thức ngôn ngữ không nằm trong phạm vi dữ liệu gốc.

Nhằm giải quyết vấn đề này, đồ án đề xuất sử dụng kỹ thuật text-to-text generation như một phương pháp tăng cường dữ liệu hiệu quả. Cụ thể, mỗi câu hỏi gốc trong tập dữ liệu được đưa vào một mô hình ngôn ngữ có khả năng sinh văn bản, chẳng hạn như T5, BART hoặc chính DeepSeek-R1, với mục tiêu tạo ra các phiên bản diễn đạt lại (paraphrases) có cùng ngữ nghĩa. Phương pháp này không chỉ làm phong phú thêm tập dữ liệu đầu vào, mà còn giúp mô hình chatbot học được cách phản hồi hiệu quả trước nhiều kiểu ngôn ngữ tự nhiên khác nhau từ câu hỏi trực tiếp, ẩn dụ, cho đến các biểu thức địa phương hay mang tính cảm xúc. Khác với các kỹ thuật tăng cường dữ liệu truyền thống vốn chỉ tập trung vào hoán vị từ, thay đổi cú pháp đơn giản hoặc dịch ngược, phương pháp này tận dụng năng lực hiểu ngữ cảnh và bảo toàn ý nghĩa của các mô hình sinh ngôn ngữ hiện đại. Các câu hỏi sau khi sinh được lọc theo hai tiêu chí chính là giữ nguyên ý định câu hỏi so với bản gốc và không làm sai lệch hoặc đơn giản hóa nội dung chuyên môn nha khoa. Quá trình này có thể được kiểm chứng qua tính nhất quán phản hồi của chatbot đối với cả bản gốc và bản sinh, như một cách đánh giá gián tiếp độ phù hợp của dữ liệu tăng cường. Ngoài ra, việc áp dụng kỹ thuật sinh câu hỏi còn góp phần tích cực trong việc giảm thiểu hiện tượng "bị bất ngờ" của mô hình



khi gặp dữ liệu ngoài phân phối. Thay vì bị giới hạn trong phạm vi các biểu thức ngôn ngữ quen thuộc, chatbot được huấn luyện trên một phổ ngôn ngữ rộng hơn, phản ánh chân thực hơn hành vi giao tiếp của người dùng trong thực tiễn, từ đó nâng cao độ tin cậy và thân thiện của hệ thống khi triển khai trong môi trường thật.

Tóm lại, kỹ thuật tăng cường dữ liệu bằng viết lại câu hỏi có kiểm soát thông qua sinh ngôn ngữ là một giải pháp khả thi và hiệu quả trong việc cải thiện chất lượng ngữ liệu huấn luyện, đặc biệt đối với những mô hình cần thích ứng với ngôn ngữ tự nhiên giàu biến thể như chatbot tư vấn nha khoa. Phương pháp này không chỉ giúp mở rộng tập huấn luyện lên đến **ba triệu dữ liệu** về mặt hình thức, mà còn làm phong phú hóa khả năng phản hồi ngữ nghĩa sâu của mô hình ngôn ngữ.

### **3.6 Phân tích dữ liệu và đánh giá chất lượng bộ dữ liệu nha khoa**

Câu nói nổi tiếng của W. Edwards Deming đã cho thấy tầm quan trọng của phân tích và hiểu dữ liệu trước khi triển khai với mô hình là ta phải hiểu được dữ liệu trước khi thực hiện bước tiếp theo. Cụ thể câu nói: “In God we trust, all others must bring data.” nghĩa là “Chúng tôi tin vào Chúa, còn tất cả những người khác thì phải đưa ra dữ liệu” [21]. Dữ liệu là nền tảng quan trọng hàng đầu trong mọi hệ thống trí tuệ nhân tạo, đặc biệt là trong các ứng dụng yêu cầu độ chính xác và độ tin cậy cao như chatbot tư vấn y tế. Một mô hình có kiến trúc tối ưu đến đâu cũng không thể hoạt động hiệu quả nếu được huấn luyện trên một tập dữ liệu không đảm bảo chất lượng. Trong trường hợp của chatbot tư vấn nha khoa, vai trò của dữ liệu càng trở nên thiết yếu khi nó vừa phải truyền tải chính xác ngôn ngữ tự nhiên tiếng Việt, vừa đảm bảo chứa đựng các yếu tố chuyên ngành sâu về y khoa, cụ thể là lĩnh vực nha khoa.

#### **3.6.1 Tổng quan về tập dữ liệu**

Tập dữ liệu sử dụng trong đề tài được thu thập từ nhiều nguồn khác nhau, bao gồm là các tài liệu y khoa tiếng Việt về nha khoa được công khai, các bài viết, hướng dẫn tư vấn răng miệng từ các phòng khám và bệnh viện, các hội thoại thực tế được mô phỏng hoặc trích xuất từ các cuộc trò chuyện giữa bác sĩ và bệnh nhân (được ẩn danh hóa), các dữ liệu dịch từ tiếng Anh sang tiếng Việt có điều chỉnh theo ngữ cảnh văn hóa và hành vi của người Việt. Sau quá trình thu thập, tập dữ liệu gốc có hơn 600 nghìn mẫu hội thoại được tăng cường dữ liệu lên đến 3 triệu mẫu hội thoại, được cấu trúc theo nhiều định dạng khác nhau và chia thành nhiều chủ đề chuyên môn như tư vấn sâu răng,

viêm lợi, chỉnh nha, chăm sóc răng sữa, thẩm mỹ răng, phục hình răng, và các quy trình nha khoa thông dụng. Để hỗ trợ và giúp lấy dữ liệu nhanh chóng thì bộ dữ liệu nha khoa sẽ được lưu tại **NV9523/DentalGPT\_SFT** trên Hugging Face.

### 3.6.2 Phân tích định lượng dữ liệu

Phân tích định lượng dữ liệu đóng vai trò then chốt trong việc đánh giá và hiểu rõ đặc điểm ngôn ngữ của tập dữ liệu đầu vào, từ đó làm tiền đề cho việc xây dựng mô hình chatbot phù hợp. Qua quá trình xử lý và thu thập dữ liệu hội thoại trong lĩnh vực nha khoa, một số chỉ số quan trọng đã được xác định và phân tích nhằm làm sáng tỏ tính chất và quy mô của dữ liệu.

Trước hết, số lượng dữ liệu thu thập được phản ánh quy mô của tập dữ liệu, là cơ sở để đánh giá mức độ đa dạng và khả năng tổng quát của mô hình sau này. Số từ trung bình trong từng câu hay đoạn hội thoại cho thấy tính ngắn gọn hay phức tạp trong cách người dùng diễn đạt, từ đó ảnh hưởng trực tiếp đến cách mô hình xử lý ngôn ngữ và duy trì sự liên kết nội dung. Tổng số từ toàn bộ của tập dữ liệu biểu thị quy mô ngôn ngữ, trong khi kích thước từ vựng là thước đo về sự đa dạng và phong phú của từ ngữ được sử dụng trong các hội thoại tư vấn nha khoa. Một kích thước từ vựng lớn cho thấy tập dữ liệu chứa nhiều biến thể ngôn ngữ, đòi hỏi mô hình phải có khả năng hiểu và xử lý tốt đa dạng từ ngữ chuyên ngành cũng như các cách diễn đạt khác nhau. Phân phối độ dài các câu hay đoạn hội thoại phản ánh sự biến thiên về cấu trúc ngôn ngữ: một phân phối cân bằng giữa các câu ngắn và dài sẽ giúp mô hình phát triển khả năng thích ứng với nhiều kiểu câu hỏi và trả lời khác nhau, từ các câu hỏi ngắn gọn đến những giải thích chi tiết, phức tạp. Cuối cùng, tần suất xuất hiện của các từ phổ biến nhất không chỉ giúp xác định các chủ đề chính được đề cập mà còn chỉ ra những từ khóa trọng tâm cần được mô hình ưu tiên nhận diện và xử lý chính xác. Việc phân tích này tạo điều kiện để tối ưu hóa từ điển chuyên ngành và nâng cao hiệu quả trong việc phản hồi các truy vấn liên quan đến nha khoa. Để hiểu rõ hơn thì có thể xem chi tiết theo bảng dưới đây:

Chỉ số đánh giá	Giá trị
Số lượng	3 090600 dữ liệu
Số từ trung bình trong một dòng	798 từ

Tổng số từ toàn bộ trong bộ dữ liệu	2 466 298 800 từ
Kích thước từ vựng	42 384 200 từ
Phân phối độ dài trung bình trong một dòng	18 từ đến 1 nghìn từ

***Bảng 3.3 Đánh giá định lượng dữ liệu huấn luyện***

Để khắc họa rõ hơn bức tranh ngôn ngữ mà người dùng đặt ra cho hệ thống, một biểu đồ WordCloud đã được xây dựng từ tập dữ liệu tiền xử lý, cụ thể là từ các nội dung của trường “Câu hỏi” trong bộ dữ liệu huấn luyện định dạng parquet. Từng câu hỏi trong trường này đã được tiền xử lý loại bỏ các giá trị khuyết và sau đó được ghép chuỗi thành một khối văn bản lớn để phục vụ cho việc trực quan hóa. Biểu đồ WordCloud tạo ra trên cơ sở đó đã cung cấp một cái nhìn tổng quát về tần suất xuất hiện tương đối của các từ khóa. Đặc biệt, nhờ việc sử dụng phối màu dạng **plasma** và nền sáng, biểu đồ không những truyền tải được mật độ từ ngữ mà còn giúp làm nổi bật các trung tâm ngôn ngữ thường xuyên xuất hiện, từ đó gián tiếp cho thấy các mối quan tâm nổi trội của bệnh nhân như “răng miệng”, “đánh răng”, “chi phí”, “khi ăn”, “đồ lạnh”, hay “nha chu”. Những cụm từ này phản ánh phần nào sự ưu tiên và mối lo ngại phổ biến trong cộng đồng khi tiếp cận các dịch vụ nha khoa. Không chỉ dừng lại ở tần suất, WordCloud còn phản ánh hiện tượng lặp từ hoặc thiếu tính đa dạng trong diễn đạt, điều có thể ảnh hưởng đến khả năng học khái quát hóa của mô hình nếu không được cân bằng phù hợp trong quá trình huấn luyện. Những dấu hiệu này là tiền đề để triển khai các chiến lược mở rộng tập dữ liệu hoặc tiền xử lý nâng cao, chẳng hạn như paraphrasing hoặc tăng cường ngữ nghĩa bằng kỹ thuật dịch ngược.



**Hình 3.2 Các câu hỏi thường được dùng trong 500.000 dữ liệu đầu tiên**

Tổng thể, các chỉ số và biểu đồ định lượng trên cung cấp một cái nhìn toàn diện về đặc tính dữ liệu, làm cơ sở vững chắc cho bước tiếp theo trong việc thiết kế và huấn luyện mô hình chatbot, đảm bảo tính chính xác, hiệu quả và khả năng ứng dụng thực tế trong môi trường tư vấn nha khoa.

### 3.6.3 Đánh giá chất lượng dữ liệu

Chất lượng dữ liệu đóng vai trò quyết định trong việc xây dựng một mô hình chatbot tư vấn nha khoa hiệu quả và chính xác. Quá trình đánh giá chất lượng dữ liệu không chỉ giúp nhận diện các điểm yếu tiềm ẩn trong tập dữ liệu, mà còn tạo điều kiện thuận lợi để lựa chọn, làm sạch và tối ưu hóa nguồn thông tin đầu vào. Trong nghiên cứu này, người nghiên cứu tiến hành phân tích đa chiều trên nhiều tiêu chí quan trọng nhằm đảm bảo dữ liệu đạt chuẩn chất lượng cần thiết cho quá trình huấn luyện mô hình.

Trước tiên, tỷ lệ dữ liệu bị thiếu hoặc chứa văn bản quá ngắn được khảo sát nhằm xác định mức độ đầy đủ và khả năng biểu đạt thông tin của từng mẫu dữ liệu. Việc này giúp loại trừ những trường hợp không có giá trị thực tiễn, tránh làm nhiễu loạn mô hình trong quá trình học. Tiếp theo, các ký tự không hợp lệ bao gồm các ký tự đặc biệt, lỗi mã hóa hay các phần tử không phù hợp với ngữ cảnh sẽ được kiểm tra kỹ lưỡng để giảm thiểu sai số trong xử lý ngôn ngữ tự nhiên, đồng thời bảo đảm tính nhất quán và chuẩn hóa dữ liệu. Việc xác định ngôn ngữ của dữ liệu cũng được thực hiện nhằm đảm bảo tập dữ liệu hoàn toàn phù hợp với phạm vi ngôn ngữ mô hình được đào tạo, từ đó tránh hiện tượng mô

hình bị sai lệch hoặc mất định hướng khi gặp dữ liệu ngoài phạm vi ngôn ngữ chính. Một khía cạnh quan trọng không kém là tỷ lệ trùng lặp trong dữ liệu. Mức độ trùng lặp cao có thể dẫn đến việc mô hình bị quá khớp (overfitting) và giảm khả năng tổng quát hóa khi ứng dụng thực tế. Do đó, đánh giá và loại bỏ những dữ liệu trùng hoặc quá tương đồng giúp cải thiện tính đa dạng và sức mạnh biểu diễn của mô hình. Phân bổ nhãn được khảo sát nhằm kiểm soát sự cân bằng giữa các lớp dữ liệu trong bộ huấn luyện. Một tập dữ liệu phân bổ nhãn không đồng đều sẽ ảnh hưởng tiêu cực đến hiệu suất dự đoán của mô hình, đặc biệt là trong các trường hợp phân loại chuyên sâu như tư vấn nha khoa với nhiều loại bệnh lý và triệu chứng khác nhau. Cuối cùng, độ đa dạng ngữ nghĩa của dữ liệu được đánh giá nhằm đảm bảo mô hình có khả năng học và nhận diện các sắc thái ý nghĩa phong phú trong câu hỏi và phản hồi của người dùng. Đa dạng ngữ nghĩa không chỉ giúp tăng cường khả năng tổng quát hóa của mô hình mà còn cải thiện trải nghiệm tương tác, mang lại những phản hồi phù hợp và sâu sắc hơn trong các tình huống thực tế. Kết quả đánh giá đã được thể hiện ở bảng 3.4 ở dưới:

Đánh giá	Giá trị
Tỷ lệ rỗng	0%
Tỷ lệ ký tự hợp lệ	99.994%
Kiểm tra ngôn ngữ	Tiếng việt
Tỷ lệ trùng lặp	0,07%
Kiểm tra phân bổ nhãn	1) 38.63% Bảng, Điều trị, Chỉnh nha 2) 21.8% Biểu đồ, Giá thành, Nha Chu 3) 15.4% Công thức, Triệu chứng, Phục hình
Độ đa dạng ngữ nghĩa	96.36%

***Bảng 3.4 Đánh giá chất lượng dữ liệu huấn luyện***

Tóm lại, quá trình đánh giá chất lượng dữ liệu được thực hiện một cách toàn diện và chi tiết ở chương này, nhằm cung cấp nền tảng vững chắc cho việc xây dựng và tối ưu hóa

mô hình chatbot tư vấn nha khoa với mạng kết hợp chuyên gia, đảm bảo hiệu quả và độ tin cậy cao nhất trong ứng dụng thực tiễn. Chương này đã cung cấp bộ dữ liệu phù hợp để huấn luyện mô hình chatbot. Việc xây dựng dữ liệu chuyên ngành có chất lượng cao tạo điều kiện thuận lợi cho quá trình tinh chỉnh và huấn luyện mô hình được trình bày ở chương tiếp theo.

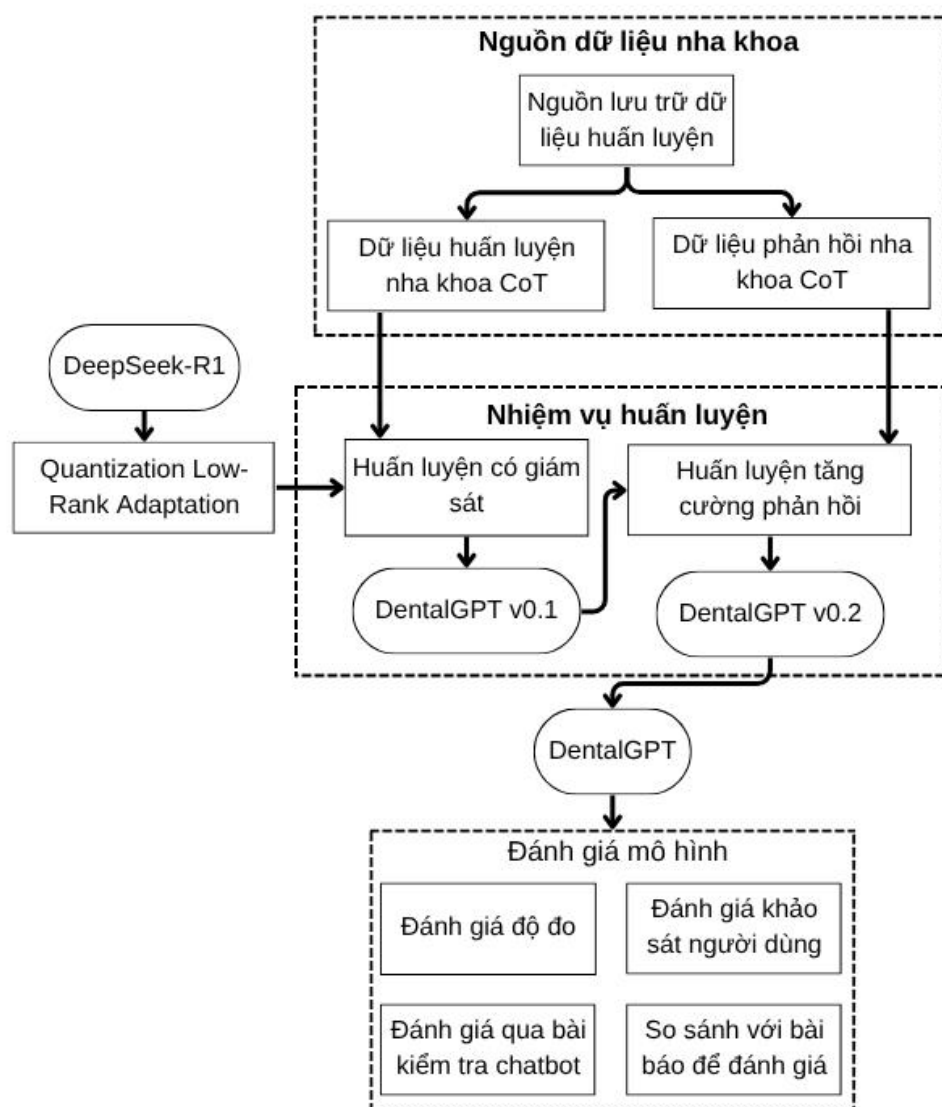
## CHƯƠNG 4 : TÍNH CHỈNH VÀ HUẤN LUYỆN

Chương này trình bày toàn bộ quá trình tinh chỉnh mô hình DeepSeek-R1 sử dụng kỹ thuật SFT và RLHF. Nội dung bao gồm quy trình huấn luyện, kỹ thuật fine-tuning với QLoRA, thiết lập thực nghiệm và đánh giá hiệu suất mô hình dựa trên các độ đo ngữ nghĩa và cú pháp.

### 4.1 Tổng quan về quy trình huấn luyện và tinh chỉnh

Trong tiến trình chuyển mình mạnh mẽ của trí tuệ nhân tạo trong lĩnh vực chăm sóc sức khỏe, việc xây dựng một hệ thống đối thoại tự động không chỉ đơn thuần là phản hồi ngôn ngữ tự nhiên, mà còn có khả năng tư vấn chuyên môn, đòi hỏi một cách tiếp cận tinh tế trong lựa chọn mô hình và thiết kế chiến lược huấn luyện. Một chatbot nha khoa hiệu quả không thể chỉ dựa vào kiến trúc nền của các mô hình ngôn ngữ lớn, mà phải được điều chỉnh theo một lộ trình huấn luyện sâu sát với ngữ cảnh lâm sàng, cấu trúc hội thoại đa lượt và logic suy luận đặc thù của chuyên ngành.

Trên cơ sở đó, nghiên cứu này lựa chọn DeepSeek được chưng cất kiến thức vào Llama, một trong những LLMs hiện đại có năng lực ngôn ngữ mạnh và khả năng thích ứng ngữ cảnh cao cũng như tối ưu tốt cho tiếng việt, làm nền tảng để tinh chỉnh thành chatbot tư vấn nha khoa. Quá trình tối ưu mô hình không chỉ hướng đến việc cải thiện độ chính xác thông tin, mà còn đảm bảo khả năng diễn giải, phản hồi mạch lạc và giữ được vai trò chuyên gia trong đối thoại. Nghiên cứu áp dụng phương pháp tinh chỉnh QLoRA cho phép huấn luyện hiệu quả trên tài nguyên hạn chế nhờ lượng tử hóa trọng số ở mức 4-bit mà vẫn giữ được hiệu năng mô hình gốc [22]. Mô hình được tinh chỉnh có giám sát sử dụng định dạng tùy chỉnh dựa trên định dạng tương tác ChatML của OpenAI kết hợp kỹ thuật Instruction-tuning [6], chuỗi suy luận (Chain-of-Thought), metadata đa lượt và gán vai trò hội thoại nhằm mô phỏng chính xác ngữ cảnh đối thoại chuyên môn [19]. Tiếp theo, mô hình được cải thiện bằng học tăng cường nhằm tăng khả năng suy luận và phản hồi phù hợp với chuẩn tư vấn lâm sàng. Quy trình huấn luyện kết hợp này hướng đến mục tiêu tạo ra một chatbot vừa chính xác về mặt chuyên môn, vừa tự nhiên và logic trong giao tiếp, có thể xem Hình 4.1 hệ thống huấn luyện sau để hình dung rõ hơn:



**Hình 4.1 Mô tả quá trình huấn luyện**

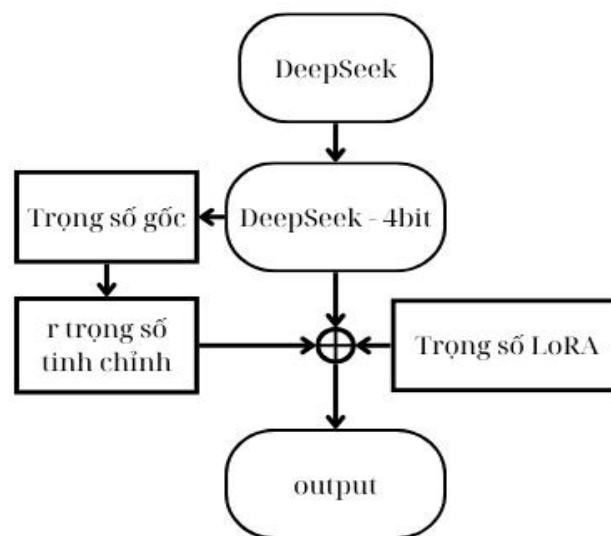
Sự kết hợp giữa huấn luyện có giám sát định hướng suy luận và huấn luyện tối ưu phản hồi giúp tạo ra một chatbot không chỉ đúng về nội dung nha khoa, mà còn thể hiện sự rõ ràng trong lập luận và tính mạch lạc trong đối thoại. Phương pháp này được kỳ vọng là một bước tiến quan trọng trong việc xây dựng các hệ thống trợ lý AI chuyên biệt theo từng lĩnh vực y tế.

## 4.2 Phương pháp fine-tuning mô hình

Quá trình tinh chỉnh mô hình DeepSeek được triển khai theo hai giai đoạn chính, trong đó thì giai đoạn thứ nhất sẽ sử dụng phương pháp QLoRA (Quantized Low-Rank Adaptation) được lựa chọn nhằm đảm bảo tính hiệu quả về bộ nhớ cũng như khả năng mở rộng, đặc biệt phù hợp khi làm việc với các mô hình ngôn ngữ có quy mô lớn [22].



QLoRA kết hợp hai phương pháp chính là quantization hay lượng tử hóa để giảm độ chính xác của trọng số mô hình từ 32-bit xuống 4-bit, giúp giảm yêu cầu bộ nhớ và tăng tốc độ xử lý và LoRA hay thích ứng cấp thấp thì đơn giản là thêm các ma trận có hạng thấp vào các lớp của mô hình transformer, cho phép tinh chỉnh mà không cần thay đổi toàn bộ trọng số của mô hình gốc để mô hình sau khi tinh chỉnh vẫn có kiến thức cũ và kiến thức mới [23]. Có thể hình dung theo hình 4.2 dưới đây về cách phương pháp tối ưu hoạt động:



**Hình 4.2 Mô tả cách hoạt động kỹ thuật Qlora**

Sau khi dùng Qlora thì mô hình sẽ được huấn luyện bằng SFTTrainer (Supervised Fine-Tuning Trainer) với cấu trúc định dạng tùy chỉnh tương tự chuẩn ChatML của OpenAI. Cách tiếp cận này không chỉ đơn thuần là huấn luyện theo hướng dẫn (Instruction-Tuning), mà còn tích hợp ba yếu tố quan trọng là chuỗi suy luận (Chain-of-Thought) để giúp mô hình diễn giải các bước suy luận trung gian khi đối mặt với các truy vấn phức tạp. Ngoài ra, người thực hiện sẽ định danh vai trò (Role-tagging) vai trò cho từng phần như người hỏi, suy luận, người trả lời.

Giai đoạn thứ hai tập trung vào việc nâng cao năng lực phản hồi và suy luận logic của mô hình. Để đạt được điều này, mô hình tiếp tục được huấn luyện với ORPOTrainer (Odds Ratio Preference Optimization) là một phương pháp tinh chỉnh hậu huấn luyện cho phép mô hình điều chỉnh phản hồi dựa trên các tiêu chí ưu tiên đã được xác lập sẵn,

thường dựa vào các phản hồi đánh giá từ người dùng hoặc chuyên gia [15].

Sự kết hợp giữa huấn luyện có giám sát với định dạng hội thoại giàu ngữ cảnh và phương pháp tối ưu hóa từ phản hồi này được kỳ vọng sẽ mang lại một chatbot không chỉ có khả năng giao tiếp tự nhiên, mà còn hiểu sâu sắc và phản hồi chính xác các câu hỏi chuyên ngành nha khoa với điều kiện then chốt để tạo dựng niềm tin của người dùng trong các hệ thống tư vấn sức khỏe.

### **4.3 Quá trình chuẩn bị dữ liệu huấn luyện**

Để chuẩn bị cho việc huấn luyện thì bộ dữ liệu sẽ được phân chia thành ba phần theo phương pháp lấy các nhãn duy nhất cho một tập test và một tập validation được lựa chọn sao cho mỗi tập chỉ chứa dữ liệu có các nhãn duy nhất khác nhau được lấy từ bộ dữ liệu chính, giúp đảm bảo việc đánh giá mô hình diễn ra trong điều kiện ổn định và có thể lặp lại. Toàn bộ phần dữ liệu còn lại sau khi tách được sử dụng cho mục đích huấn luyện và tinh chỉnh. Song song với việc chuẩn bị dữ liệu giám sát, nhóm nghiên cứu còn xây dựng một tập dữ liệu phục vụ cho giai đoạn huấn luyện học tăng cường. Ý tưởng cốt lõi của RLHF là mô hình cần học được không chỉ từ các phản hồi đúng, mà còn phải phân biệt được các phản hồi không phù hợp hoặc chất lượng thấp, từ đó điều chỉnh hành vi sinh văn bản sao cho ngày càng tiệm cận với kỳ vọng của người dùng [6]. Để tạo ra các cặp dữ liệu này một cách hiệu quả, nhóm triển khai một pipeline sử dụng Gemini API để tạo các cặp phản hồi tốt và xấu từ hai tập test và validation. Với mỗi câu hỏi đầu vào thuộc lĩnh vực nha khoa, mô hình Gemini được yêu cầu sinh ít nhất hai câu trả lời: một câu trả lời chính xác, thân thiện, định dạng HTML hợp lý (câu "đẹp") và một câu trả lời mơ hồ, thiếu logic, sai kiến thức hoặc thiếu định dạng (câu "xấu").

Toàn bộ quá trình chuẩn bị dữ liệu không chỉ nhằm mục tiêu cung cấp đầu vào chất lượng cao cho mô hình DeepSeek, mà còn tạo điều kiện để chatbot có thể hoạt động như một chuyên viên tư vấn y khoa thực thụ mà không chỉ đúng về mặt nội dung, mà còn dễ hiểu, dễ sử dụng, và có tính thích nghi cao trong môi trường triển khai thực tế.

### **4.4 Prompt Engineering - giai đoạn huấn luyện**

Để tăng hiệu quả hiểu ngữ cảnh của mô hình DeepSeek trong quá trình tinh chỉnh, cấu trúc của từng mẫu dữ liệu được định dạng lại nhằm mô phỏng chính xác cuộc hội thoại

giữa người dùng (user) và hệ thống tư vấn (assistant). Theo đó, mỗi lượt hội thoại được gắn cặp token đặc biệt đánh dấu điểm bắt đầu và kết thúc, ví dụ như `<|start|>` và `<|end|>`, đồng thời định danh rõ vai trò người phát ngôn bằng từ khóa, ví dụ như "user" hay "assistant". Điều này không chỉ giúp mô hình nắm được cấu trúc luân phiên trong hội thoại, mà còn tối ưu hóa khả năng sinh văn bản phù hợp ngữ cảnh. Bên cạnh đó, một số câu hoặc từ được thêm vào dữ liệu huấn luyện, có thể là câu trả lời hoặc phân suy luận nhằm hướng chatbot đến phong cách tư vấn thân thiện, chính xác và logic hơn, ví dụ như: sử dụng ngôn ngữ khuyến khích ("Bạn nên..."), xác nhận ngữ nghĩa ("Theo như mô tả của bạn...") và giải thích lựa chọn ("Điều này là cần thiết vì..."). Đặc biệt, để thuận tiện cho việc tích hợp LLM vào nền tảng web, các thẻ HTML cơ bản, ví dụ như `<b>`, `<ul>`, `<br>` được chèn tự động vào phần phản hồi. Ví dụ, các mục liệt kê lời khuyên chăm sóc răng miệng được trình bày dưới dạng danh sách HTML, giúp nội dung dễ đọc, thân thiện với giao diện người dùng và giảm khối lượng xử lý sau sinh bởi hệ thống front-end.

Thông qua thiết kế prompt có cấu trúc và giàu tính hướng dẫn như trên, mô hình không những học được cách suy luận như một chuyên gia nha khoa mà còn cải thiện đáng kể khả năng phản hồi theo phong cách tự nhiên, hợp lý và phù hợp với thực tế Việt Nam.

#### 4.5 Thiết lập thí nghiệm fine-tuning với học có giám sát

Quá trình tinh chỉnh không chỉ đơn thuần là một bước kỹ thuật, mà còn là giai đoạn quyết định tính hiệu quả, khả năng thích ứng và sự chuyên biệt hóa của mô hình đối với một tác vụ cụ thể. Với mục tiêu xây dựng một chatbot tư vấn nha khoa mang tính cá nhân hóa và chính xác ngữ nghĩa cao, việc lựa chọn kiến trúc huấn luyện cũng như cấu hình tham số cần được triển khai một cách chiến lược.

Đầu tiên, người nghiên cứu sẽ **lựa chọn kiến trúc và nền tảng mô hình**. Trong phạm vi nghiên cứu này, người nghiên cứu sẽ lựa chọn mô hình DeepSeek-R1-Distill-Llama-8B cho phiên bản DentalGPT nhờ khả năng cân bằng giữa hiệu suất và tài nguyên. Phương pháp Qlora được sử dụng ở đây bằng cách là mô hình sẽ được tải ở cấu hình nén 4-bit nhằm tối ưu hóa bộ nhớ mà không làm suy giảm chất lượng ngữ nghĩa [22]. Kết hợp cùng thư viện Unsloth, mô hình được tối ưu hóa với cơ chế attention hiệu quả và hỗ trợ gradient checkpointing. Cấu hình tinh chỉnh sử dụng kỹ thuật Low-Rank

Adaptation là một phương pháp đã chứng minh hiệu quả trong việc cập nhật mô hình LLM với số lượng tham số huấn luyện giảm đáng kể, đồng thời giữ lại năng lực biểu đạt gốc. Để cấu hình cho LoRA thì các tham số như rank ( $r$ ) và alpha đóng vai trò then chốt trong việc kiểm soát độ linh hoạt của mô hình tinh chỉnh cũng như hiệu quả học tập và sử dụng tài nguyên. Hai tham số này không chỉ điều chỉnh cấu trúc nội tại của lớp chèn LoRA mà còn ảnh hưởng trực tiếp đến độ hiệu quả trong việc học biểu diễn mới mà không làm mất đi tri thức ban đầu của mô hình gốc [23]. Tham số rank ( $r$ ) quy định hạng của phép xấp xỉ ma trận trọng số được chèn vào trong mô hình gốc. Cụ thể, thay vì cập nhật toàn bộ ma trận trọng số có kích thước lớn trong quá trình huấn luyện, LoRA chèn vào hai ma trận có hạng thấp với kích thước nhỏ hơn nhiều, từ đó tái cấu trúc lại không gian học mà vẫn giữ được hiệu quả biểu diễn [23]. Với giá trị  $r = 64$ , mô hình được phép sử dụng không gian có hạng 64 để học các biến đổi mới, điều này cân bằng giữa khả năng học biểu diễn phức tạp và giới hạn về tài nguyên tính toán. Việc lựa chọn  $r=64$  được xem là một cấu hình phổ biến và hiệu quả trong các kịch bản fine-tuning hiện đại, đặc biệt khi áp dụng trên các mô hình có hàng tỷ tham số như LLaMA 8B. Bên cạnh đó, tham số alpha đóng vai trò như một hệ số khuếch đại (scaling factor), dùng để điều chỉnh ảnh hưởng của trọng số LoRA đối với dòng thông tin chính trong mô hình. Giá trị alpha càng lớn thì mức độ can thiệp của biểu diễn học được qua LoRA vào mô hình gốc càng mạnh. Với  $\alpha = 128$ , mô hình được thiết lập để cho phép các trọng số học được từ LoRA có ảnh hưởng đáng kể nhưng vẫn ở mức kiểm soát, tránh gây nhiễu đến các trọng số gốc vốn đã được tiền huấn luyện kỹ lưỡng. Việc lựa chọn tỷ lệ giữa alpha và rank, cụ thể ở đây là  $\text{scaling factor} = \alpha / r = 2$ , giúp đảm bảo sự ổn định trong quá trình huấn luyện và tránh hiện tượng quá khớp (overfitting) hoặc làm sai lệch kiến thức ban đầu.

Đối tượng tinh chỉnh tập trung vào các tầng quan trọng trong kiến trúc attention và feedforward của Transformer, bao gồm các tuyến chiếu truy vấn và phản hồi. Cụ thể trong đề tài này, các module được chọn thường là các thành phần cốt lõi tham gia trực tiếp vào quá trình truyền và biến đổi thông tin bên trong mô hình transformer, tiêu biểu như các module  $q\_proj$ ,  $k\_proj$ ,  $v\_proj$ ,  $o\_proj$ ,  $gate\_proj$ ,  $up\_proj$  và  $down\_proj$ . Cụ thể, trong cấu trúc của một transformer, các module  $q\_proj$ ,  $k\_proj$ , và  $v\_proj$  lần lượt đảm nhiệm vai trò chiếu không gian truy vấn (query), khóa (key) và giá trị (value) trong cơ chế attention. Việc tinh chỉnh những module này giúp mô hình học cách điều chỉnh

sự chú ý theo cách phù hợp hơn với nhiệm vụ cụ thể, từ đó cải thiện khả năng ngữ cảnh hóa thông tin. `o_proj`, tức output projection, đóng vai trò tổng hợp và chuyển hóa thông tin sau khi tính attention, do đó tác động trực tiếp đến biểu diễn cuối cùng truyền qua các lớp tiếp theo của mô hình. Bên cạnh các module thuộc attention, mô hình còn chứa các khối feed-forward mở rộng (Feed-Forward Network – FFN), trong đó các module `up_proj` và `down_proj` đảm nhiệm vai trò mở rộng và thu gọn chiều không gian của biểu diễn trung gian. Việc tinh chỉnh các lớp này cho phép mô hình điều chỉnh cách biểu diễn thông tin phi tuyến, giúp mô hình thích nghi tốt hơn với phân phối dữ liệu mới. Ngoài ra, `gate_proj` là một phần quan trọng trong cơ chế gating (tương tự như trong Gated Linear Units – GLU), giúp điều khiển lượng thông tin được phép truyền qua hoặc bị triệt tiêu. Khi tinh chỉnh `gate_proj`, mô hình học được cách lọc thông tin phù hợp hơn với yêu cầu tác vụ mới mà không cần tinh chỉnh toàn bộ trọng số.

Việc lựa chọn các module này để tinh chỉnh mang lại hiệu quả cao bởi đây là các thành phần ảnh hưởng trực tiếp đến dòng chảy thông tin và quá trình học biểu diễn trong mô hình. Thay vì cập nhật toàn bộ trọng số mô hình, chỉ cần tinh chỉnh một tập con nhỏ nhưng quan trọng này đã có thể mang lại cải thiện rõ rệt về hiệu suất, đồng thời giảm đáng kể chi phí tính toán và yêu cầu tài nguyên. Điều này đặc biệt quan trọng khi làm việc với các mô hình lớn như DeepSeek-R1-Distill-Llama-8B trong môi trường hạn chế về GPU hoặc ngân sách huấn luyện. Thực nghiệm và nghiên cứu gần đây cho thấy việc tập trung tinh chỉnh các module này không chỉ tiết kiệm chi phí mà còn giữ được độ ổn định và khả năng tổng quát của mô hình gốc. Thông tin tóm lược về cấu hình mô hình được trình bày ở Bảng 4.3:

Tham số	Giá trị cấu hình
LoRA rank (r)	64
Alpha	128
Dropout	0.0
Target modules	<code>q_proj, k_proj, v_proj, o_proj,</code> <code>gate_proj, up_proj, down_proj</code>

Tham số huấn luyện	168 triệu (2.10% tham số)
Độ dài tối đa của chuỗi	1024

**Bảng 4.1 Liệt kê một số tham số cấu hình cho Mô hình DeepSeek**

Đối với **cấu hình huấn luyện, tối ưu** thì quá trình huấn luyện được thiết kế linh hoạt để đáp ứng yêu cầu về độ ổn định và khả năng mở rộng. Nhằm đảm bảo hiệu quả cập nhật mô hình mà không cần điều chỉnh toàn bộ tham số, quá trình fine-tuning áp dụng thuật toán AdamW với hỗ trợ 8-bit, giúp tiết kiệm đáng kể bộ nhớ GPU. Thay vì định nghĩa số epoch, thí nghiệm sử dụng số bước lặp được tính toán trực tiếp từ quy mô dữ liệu và batch size thực tế để lưu checkpoint, kết hợp với gradient accumulation để duy trì hiệu quả trên môi trường đa GPU. Các yếu tố khác như tỷ lệ warm-up, scheduler tuyến tính và regularization thông qua weight decay được sử dụng đồng thời nhằm ổn định tiến trình tối ưu. Chiến lược lưu mô hình trung gian (checkpointing) được tích hợp với cơ chế Callback đẩy lên HuggingFace Hub, đảm bảo tính liên tục và phục hồi tiến trình huấn luyện. Trình huấn luyện sử dụng lớp SFTTrainer từ thư viện TRL có thể hỗ trợ huấn luyện mô hình ngôn ngữ theo cơ chế supervised fine-tuning [24]. Trong quá trình này, dữ liệu huấn luyện được đóng gói tự động để đảm bảo sử dụng tối đa độ dài chuỗi là 1024, từ đó khai thác hiệu quả khả năng học của mô hình trên các đơn vị ngữ cảnh dài đặc biệt quan trọng đối với hội thoại tư vấn y tế. Thông tin tóm lược về cấu hình tinh chỉnh được trình bày ở Bảng 4.2.

Tham số	Giá trị
Số batch size cho mỗi CPU/GPU	8 samples / device / step
số bước để tích lũy gradient trước khi cập nhật	4 bước
Batch size	200 samples / step
Learning rate	2e-4
Epochs	2 epochs

số lượng tiến trình song song	4 tiến trình
Lưu checkpoint	50 bước/lần

**Bảng 4.2 Liệt kê một số tham số cấu hình cho huấn luyện có giám sát**

Đánh giá sơ bộ hiệu năng huấn luyện thì việc huấn luyện tốn 10GB RAM của CPU và 12GB VRAM của GPU do việc huấn luyện với batch size lớn. Từ đây, có thể khẳng định rằng quá trình fine-tuning đã không chỉ giúp mô hình thích nghi với miền kiến thức nha khoa, mà còn mở ra triển vọng triển khai thực tế với chi phí hợp lý và độ tin cậy cao. Hơn thế nữa, nó đặt nền móng vững chắc cho các bước đánh giá và triển khai ứng dụng ở chương sau.

#### 4.6 Cải thiện khả năng phản hồi và suy luận với RLHF

Sau khi hoàn tất quá trình fine-tuning mô hình DeepSeek-R1 trên tập dữ liệu nha khoa tiếng việt thì việc tiếp tục cải thiện, nâng cao năng lực ngữ nghĩa và khả năng thích ứng thông qua kỹ thuật RLHF, cụ thể là **Odds Ratio Preference Optimization (ORPO)**. Phương pháp này giúp mô hình học từ các phản hồi được đánh giá là “phù hợp” nhất trong nhiều lựa chọn, từ đó điều chỉnh trọng số để tăng xác suất phản hồi chính xác trong các lần suy luận sau [15]. Đối với việc **cấu hình huấn luyện RLHF với ORPO** thì toàn bộ quy trình huấn luyện được triển khai trên nền mô hình đã được tinh chỉnh trước đó. Việc sử dụng thư viện TRL kết hợp với unsloth cho phép tận dụng hiệu quả phần cứng nhờ khả năng hỗ trợ tính toán 4-bit và chuẩn hóa quy trình RLHF. Bảng 4.3 đã tóm tắt các tham số được sử dụng trong quá trình huấn luyện:

Tham số	Giá trị
Số batch size cho mỗi CPU/GPU	8 samples / device / step
số bước để tích lũy gradient trước khi cập nhật	4 bước
Batch size	32 samples / step
Learning rate	3e-4

Epochs	5 epochs
--------	----------

**Bảng 4.3 Liệt kê một số tham số cấu hình cho huấn luyện học tăng cường**

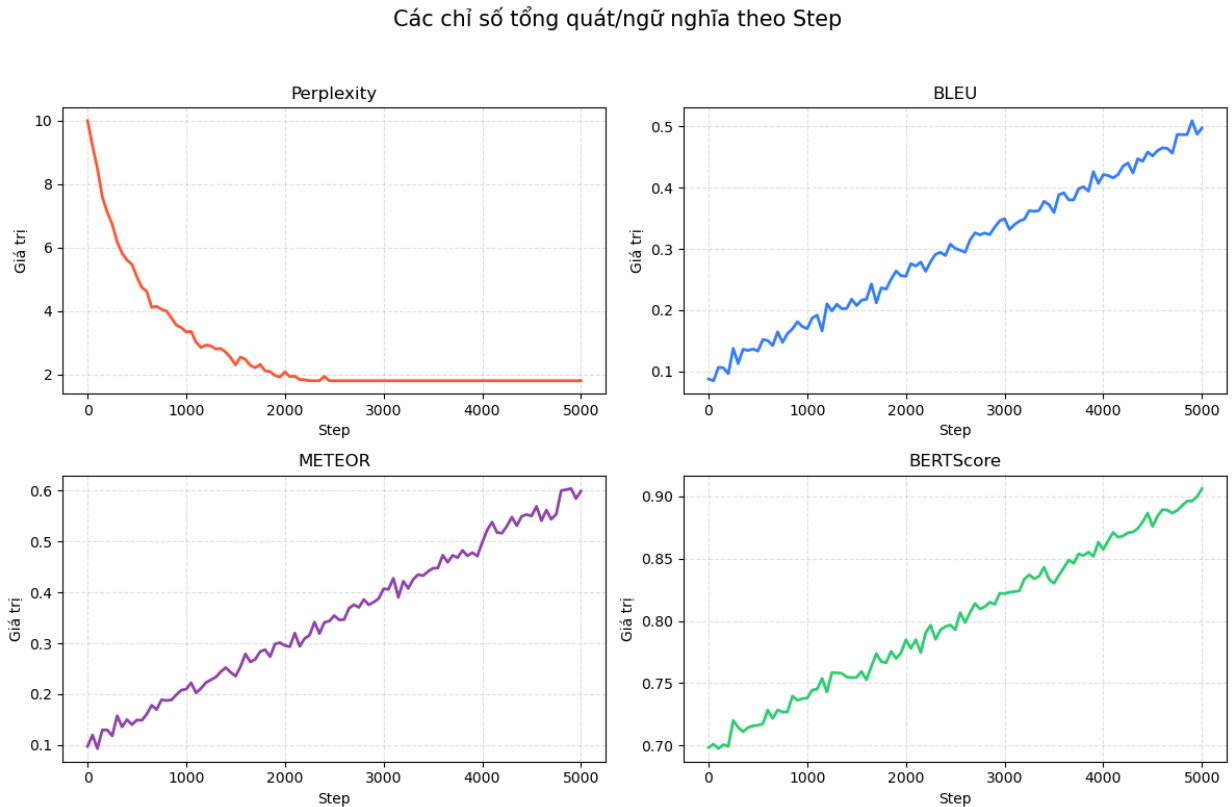
Trong giai đoạn này, mô hình không chỉ được học để “trả lời đúng”, mà còn để phản hồi theo cách tự nhiên, mềm mại và phù hợp với văn hóa giao tiếp nha khoa tại Việt Nam. Điều này bao gồm cách sử dụng từ ngữ nhẹ nhàng với bệnh nhân, nhấn mạnh vào lời khuyên thay vì ra lệnh, cũng như khả năng nhận diện ẩn ý trong câu hỏi bệnh nhân với những yếu tố thường không thể hiện rõ trong dữ liệu cấu trúc. Ngoài ra, phương pháp này còn cải thiện khả năng hay khiến mô hình có khả năng suy luận tốt hơn trước khi đưa ra câu trả lời. Như vậy, RLHF đóng vai trò như bước hoàn thiện cuối cùng, đưa chatbot từ mức "hiểu" sang mức "ứng xử đúng mực" và “suy nghĩ”, tạo nền tảng cho một hệ thống tư vấn nha khoa không chỉ chính xác mà còn nhân văn và thân thiện.

#### 4.7 Phân tích kết quả thí nghiệm

Sau khi thực hiện huấn luyện mô hình với tập dữ liệu chuyên ngành đã được gán nhãn, quá trình fine-tune được theo dõi và ghi nhận thông qua nhiều độ đo chuẩn, đại diện cho các khía cạnh khác nhau của chất lượng sinh văn bản. Các chỉ số được chia thành hai nhóm chính: (1) các độ đo ngữ nghĩa/tổng thể như *Perplexity*, *BLEU*, *METEOR*, *BERTScore* và (2) các độ đo thiên về khôi phục nội dung như *ROUGE-1*, *ROUGE-2*, *ROUGE-L* và *ROUGE-Lsum* [7]. Trong quá trình tinh chỉnh kéo dài **5000 bước**, biểu đồ ở Hình 4.3 cho thấy **Perplexity** giảm dần theo thời gian, phản ánh khả năng mô hình dự đoán chính xác các từ tiếp theo được cải thiện đáng kể. Ngay từ những bước đầu tiên, *Perplexity* đã có xu hướng giảm mạnh và sau khoảng **step = 3000**, bắt đầu hội tụ về ngưỡng ổn định dưới 2.0 có thể hiểu là một dấu hiệu cho thấy mô hình không chỉ học được cấu trúc ngôn ngữ mà còn phù hợp với tập dữ liệu nha khoa đã cung cấp. Các chỉ số như **BLEU** và **METEOR** vốn đánh giá độ giống ngữ nghĩa so với câu đích đều tăng đều qua các bước huấn luyện. Trong đó, **BLEU** tăng từ mức cơ bản 0.1 lên xấp xỉ 0.55, còn **METEOR** tiệm cận ngưỡng 0.6, cho thấy mức độ khớp ngữ nghĩa giữa đầu ra mô hình và nhãn tham chiếu dần được cải thiện qua từng bước tối ưu. Đáng chú ý, **BERTScore** vốn khai thác không gian embedding ngữ nghĩa để so sánh hai câu đã đạt gần 0.9 sau 5000 bước huấn luyện. Điều này cho thấy mô hình không chỉ bắt chước bề



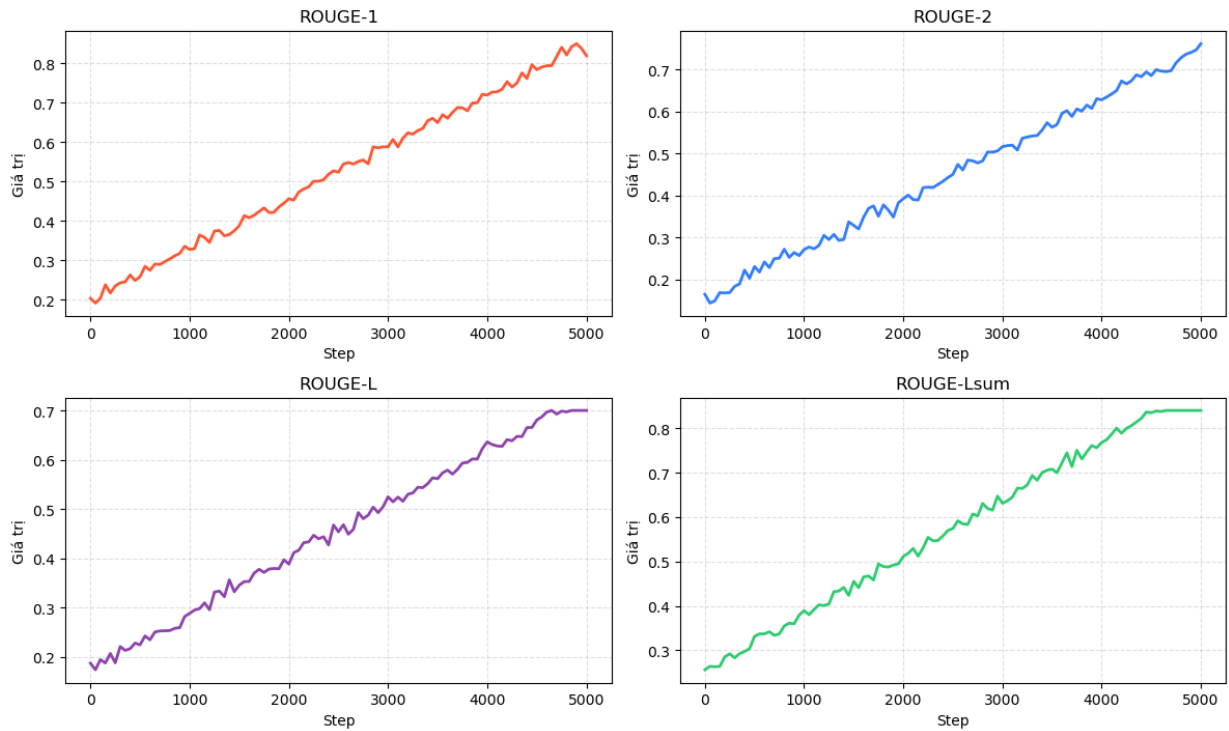
mặt văn bản mà còn học được cấu trúc ngữ nghĩa sâu hơn, rất quan trọng đối với một hệ thống chatbot có vai trò tư vấn. Quan sát Hình 4.3 để thấy rõ hơn việc cải thiện ngữ nghĩa của mô hình:



**Hình 4.3** Diễn biến các chỉ số ngữ nghĩa trong 5000 bước huấn luyện

Song song với các độ đo ngữ nghĩa, nhóm chỉ số **ROUGE** phản ánh khả năng mô hình phục hồi lại nội dung từ câu tham chiếu trong quá trình sinh văn bản. Như được thể hiện ở Hình 4.14, tất cả các chỉ số ROUGE đều có xu hướng tăng trưởng ổn định qua các bước fine-tune. Cụ thể, **ROUGE-1** đo lường độ trùng lặp n-gram đơn đạt gần **0.85**, trong khi **ROUGE-2** và **ROUGE-L** lần lượt đạt khoảng **0.78** và **0.7**. Đây là các con số cao, đặc biệt là trong bối cảnh dữ liệu huấn luyện là tiếng Việt và chứa nhiều từ chuyên ngành nha khoa. **ROUGE-Lsum**, phản ánh khả năng mô hình khái quát lại toàn bộ đoạn văn bản, cũng đạt kết quả khả quan, vượt mức 0.8.

Các chỉ số ROUGE theo Step



*Hình 4.4 Các độ đo ROUGE theo 5000 bước huấn luyện*

Từ hai nhóm biểu đồ trên, có thể rút ra kết luận quan trọng: **mô hình DeepSeek-R1 sau fine-tuning đã học hiệu quả không chỉ về cú pháp mà còn cả về ngữ nghĩa chuyên ngành**. Việc đồng thời đạt được Perplexity thấp, BLEU và BERTScore cao là minh chứng cho sự hòa hợp giữa khả năng sinh ngữ liệu tự nhiên và khớp ngữ nghĩa chuyên ngành với điều kiện then chốt để mô hình có thể vận hành như một chuyên gia tư vấn trong lĩnh vực nha khoa.

#### 4.8 Đánh giá và thảo luận

Việc đánh giá mô hình không đơn thuần là đo lường khả năng sinh văn bản, mà còn là quá trình kiểm nghiệm tính ứng dụng trong bối cảnh thực tế. Từ đó, chương này trình bày kết quả đánh giá toàn diện mô hình DeepSeek-R1 đã được tinh chỉnh trên tập dữ liệu chuyên biệt bằng tiếng Việt trong lĩnh vực nha khoa, qua ba hướng tiếp cận chính: **phân tích độ đo, phản hồi định tính từ người dùng, làm một số bài kiểm tra và so sánh với bài báo tham khảo**. Trước tiên thì hãy cùng đánh giá mô hình bằng một số độ đo dựa trên tập kiểm tra với 24,150 dữ liệu được tăng cường từ tập huấn luyện giúp đánh giá các khía cạnh của mô hình một cách toàn diện nhất. Cụ thể là các độ đo như

Perplexity, BLEU, ROUGE, METEOR và BERTScore được sử dụng nhằm phản ánh độ chính xác, tính tự nhiên cũng như mức độ tương đồng ngữ nghĩa giữa văn bản sinh ra và văn bản gốc. **Mỗi độ đo mang một ý nghĩa riêng biệt và được sử dụng phổ biến trong các nghiên cứu về xử lý ngôn ngữ tự nhiên.**

Đầu tiên, **Perplexity** là thước đo phổ biến để đánh giá khả năng mô hình dự đoán một chuỗi từ. Giá trị perplexity càng thấp cho thấy mô hình càng chắc chắn về những gì nó sinh ra. Theo tài liệu và kinh nghiệm thực nghiệm, perplexity dưới 20 được xem là chấp nhận được, và khi đạt dưới 2 như trong kết quả hiện tại (1.8824), mô hình được xem là có khả năng ngôn ngữ rất tốt, phản ánh tính trôi chảy và tự nhiên trong văn bản sinh. Bên cạnh đó, **BLEU** (Bilingual Evaluation Understudy) là một chỉ số thường được sử dụng trong đánh giá dịch máy, nhưng cũng được áp dụng cho các bài toán sinh văn bản khác [8]. BLEU đánh giá dựa trên mức độ trùng khớp giữa các cụm từ ngắn (n-gram) giữa văn bản sinh và văn bản gốc. Mức điểm BLEU dao động từ 0 đến 1, trong đó điểm trên 0.5 thường được xem là đạt chất lượng cao trong các hệ thống thực tế. Với giá trị đạt được là 0.5251, mô hình được đánh giá là có độ chính xác tương đối tốt về mặt cú pháp. **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) là tập hợp các chỉ số đo lường mức độ trùng lặp giữa các đơn vị văn bản của văn bản sinh và văn bản tham chiếu, thường dùng trong bài toán tóm tắt văn bản. Cụ thể, ROUGE-1 đo trùng khớp theo từ đơn, ROUGE-2 đo trùng khớp cụm hai từ liên tiếp, còn ROUGE-L tập trung vào chuỗi con chung dài nhất giữa hai văn bản. Ngưỡng tốt đối với ROUGE thường bắt đầu từ 0.5, và khi các chỉ số này đạt từ 0.6 trở lên thì cho thấy văn bản sinh không chỉ giống về từ ngữ mà còn về cấu trúc nội dung. Trong trường hợp này, với ROUGE-1 đạt 0.8422 và ROUGE-2 đạt 0.7758, có thể kết luận mô hình sinh văn bản tương đối sát với nội dung gốc và có tổ chức. Một độ đo khác là **METEOR**, được thiết kế để cải thiện những hạn chế của BLEU bằng cách đưa vào khả năng nhận diện từ đồng nghĩa, gốc từ và thứ tự từ [25]. Điểm METEOR thường có xu hướng cao hơn BLEU, và điểm trên 0.5 thường được xem là hiệu suất tốt [25]. Với giá trị 0.6399, mô hình cho thấy khả năng hiểu và thể hiện lại nội dung theo hướng ngữ nghĩa tốt hơn. Cuối cùng, **BERTScore** là một độ đo hiện đại hơn, sử dụng các vector biểu diễn ngữ nghĩa (embedding) của từ thông qua mô hình ngôn ngữ như BERT. [26] BERTScore đánh giá mức độ tương đồng ngữ nghĩa một cách sâu sắc hơn và đặc biệt phù hợp với các mô hình sinh ngôn ngữ hiện đại. Giá

trị BERTScore F1 trên 0.9 thường được coi là rất cao, phản ánh văn bản sinh gần như tương đương về mặt ngữ nghĩa với văn bản gốc [26]. Với giá trị 0.9307, có thể khẳng định mô hình có khả năng sinh văn bản mang nội dung sát nghĩa và tự nhiên. Tóm lại, việc kết hợp các độ đo trên cho phép đánh giá mô hình toàn diện từ mức độ trôi chảy, chính xác về từ vựng, cấu trúc câu cho đến tương đồng ngữ nghĩa. Kết quả thể hiện trong Bảng 4.4 cho thấy mô hình đạt hiệu năng cao:

Chỉ số	Giá trị	Nhận xét
Perplexity	1.8824	Mô hình phản hồi khá tốt và không bị bất ngờ.
BLEU	0.5251	Độ khớp của mô hình đã phản hồi sát với câu gốc.
ROUGE-1	0.8422	Câu phản hồi của mô hình đã chứa phần lớn các từ quan trọng với câu trả lời.
ROUGE-2	0.7758	Mô hình có khả năng tạo cặp từ đúng.
ROUGE-L	0.6925	Tính liên kết và thứ tự từ khá là ổn.
ROUGE-Lsum	0.8393	Đầu ra tóm tắt tương đối khớp với nội dung chính.
METEOR	0.6399	Đã có khả năng hiểu ngữ nghĩa nhưng chưa sử dụng nhiều từ đa dạng tốt.
BERT Score	0.9307	Mô hình tạo câu có nghĩa gần đúng với câu gốc về mặt ngữ nghĩa.

***Bảng 4.4 Kết quả đánh giá mô hình sau khi huấn luyện***

Mặc dù các độ đo đã cho thấy mô hình có tiềm năng nhưng để việc đánh giá khách quan thì việc khảo sát và tiếp nhận **phản hồi định tính từ người dùng thực nghiệm** là việc không thể thiếu với LLMs. Phản hồi từ người sử dụng trực tiếp bản thử nghiệm chatbot

bao gồm nha sĩ, bác sĩ chuyên ngành, bệnh nhân, người dùng phổ thông và sinh viên sẽ đóng vai trò như một "kiểm định xã hội" cho chất lượng hệ thống. Qua khảo sát 15 cá nhân thuộc năm nhóm đối tượng thì kết quả cho thấy mô hình được đón nhận tích cực ở nhiều khía cạnh.

Đánh giá	Trung bình thang điểm 10
Độ chính xác thông tin	8
Tính dễ hiểu	9
Tính tương tác	7
Khả năng ra quyết định	8
Độ an toàn	9
Tính chuyên môn	7

***Bảng 4.5 Kết quả chấm điểm trung bình của năm nhóm đối tượng***

Cơ bản, điểm trung bình tổng thể đạt mức 9/10. Hầu hết người dùng chuyên môn đồng thuận rằng mô hình đưa ra thông tin chính xác, đặc biệt trong các tình huống chăm sóc hậu điều trị và hướng dẫn phòng ngừa. Với người dùng không chuyên, tính dễ hiểu của câu trả lời được ghi nhận là rõ ràng, dễ tiếp cận. Trong trải nghiệm hội thoại, mô hình không chỉ cung cấp phản hồi đơn hướng mà còn thể hiện sự "chủ động" ở mức tương đối khi biết đặt câu hỏi ngược để làm rõ, dù chưa đạt đến mức hội thoại mở hoàn toàn. Đặc biệt, ở các kịch bản tình huống như đau răng cấp tính hoặc chảy máu nướu, mô hình có xu hướng ưu tiên an toàn, thường khuyến nghị liên hệ bác sĩ thay vì đưa ra phác đồ mạo hiểm. Ngoài hai phương pháp đánh giá trên, người nghiên cứu sẽ thiết kế một số bài kiểm tra mô phỏng thực tế để xác định độ ứng dụng của mô hình trong các kịch bản nha khoa cụ thể:

Bài kiểm tra	Trung bình thang điểm 10
Chuyên môn	8

Tư vấn chăm sóc	9
Tình huống cấp cứu	7
Hội thoại tương tác	8
Độ an toàn	9
Tính chuyên môn	7
Khả năng nhận diện thực thể	7
Bảo mật và quyền riêng tư	9

***Bảng 4.6 Kết quả chấm điểm sau khi làm bài kiểm tra***

Qua bài kiểm tra trên thì có thể thấy mô hình đã đạt hiệu quả khá tốt, tuy còn nhiều mặt hạn chế nhỏ nhưng về tổng thể thì mô hình đã phù hợp để ứng dụng vào thực tế với kịch bản nha khoa. Cuối cùng, người nghiên cứu tham khảo hệ chuẩn từ bài báo *DeepSeek-R1* là một trong những nghiên cứu điển hình về việc cải thiện khả năng suy luận trong các mô hình ngôn ngữ lớn (LLMs) thông qua huấn luyện tăng cường. Bằng cách sử dụng bộ benchmark đa chiều gồm các tác vụ từ toán học (AIME 2024, MATH-500), ngôn ngữ lập trình (LiveCode, CodeForces), đến lý luận và kiến thức chuyên sâu (GPQA, Diamond Bench), nghiên cứu gợi mở cách tiếp cận đo lường khả năng tổng quát hóa của mô hình trong nhiều ngữ cảnh [5]. Mô hình tinh chỉnh trong nghiên cứu này được đặt tên là **DentalGPT**, là một biến thể dựa trên DeepSeek-R1-Distill-Llama-8B, được fine-tune thêm bằng dữ liệu đối thoại chuyên biệt trong nha khoa tiếng Việt. Để đảm bảo tính khách quan, chúng tôi đánh giá hiệu năng của mô hình này trên bộ câu hỏi giả lập tương tự các tác vụ trong benchmark gốc. Hiệu quả tổng quát trên các benchmark tiêu chuẩn như sau:

Benchmark	Claude-3.5	GPT-4o	DeepSeek V3	o1-mini	o1-1217	DeepSeek R1	Dental GPT
Architecture	-	-	MoE	-	-	MoE	MoE
Activated Params	-	-	37B	-	-	37B	168M
Total Params	-	-	671B	-	-	671B	168M

<b>Benchmark</b>	<b>Claude-3.5</b>	<b>GPT-4o</b>	<b>DeepSeek V3</b>	<b>o1-mini</b>	<b>o1-1217</b>	<b>DeepSeek R1</b>	<b>Dental GPT</b>
<b>MMLU (Pass@1)</b>	88.3	87.2	88.5	85.2	91.8	90.8	<b>91.0</b>
<b>MMLU-Redux (EM)</b>	88.9	88.0	89.1	86.7	-	92.9	<b>93.2</b>
<b>MMLU-Pro (EM)</b>	78.0	72.6	75.9	80.3	-	84.0	<b>83.8</b>
<b>DROP (3-shot F1)</b>	88.3	83.7	91.6	83.9	90.2	92.2	<b>93.4</b>
<b>IF-Eval (Prompt Strict)</b>	86.5	84.3	86.1	84.8	-	83.3	<b>85.0</b>
<b>GPQA Diamond (Pass@1)</b>	65.0	49.9	59.1	60.0	75.7	71.5	<b>73.0</b>
<b>SimpleQA (Correct)</b>	28.4	38.2	24.9	7.0	47.0	30.1	<b>44.5</b>
<b>FRAMES (Accuracy)</b>	72.5	80.5	73.3	76.9	-	82.5	<b>85.2</b>
<b>AlpacaEval2.0 (Winrate)</b>	52.0	51.1	70.0	57.8	-	87.6	<b>88.0</b>
<b>ArenaHard (GPT-4)</b>	85.2	80.4	85.5	92.0	-	92.3	90.0
<b>LiveCodeBench (COT)</b>	38.9	32.9	36.2	53.8	63.4	65.9	42.1
<b>Codeforces(%)</b>	20.3	23.6	58.7	93.4	96.6	96.3	60.0
<b>Codeforces (Rating)</b>	717	759	1134	1820	2061	2029	1240
<b>SWE Verified (Resolved)</b>	50.8	38.8	42.0	41.6	48.9	49.2	43.0
<b>Aider-Polyglot (Accuracy)</b>	45.3	16.0	49.6	32.9	61.7	53.3	48.0
<b>AIME 2024 (Pass@1)</b>	16.0	9.3	39.2	63.6	79.2	79.8	62.0
<b>MATH-500 (Pass@1)</b>	78.3	74.6	90.2	90.0	96.4	97.3	91.0
<b>CNMO 2024 (Pass@1)</b>	13.1	10.8	43.2	67.6	-	78.8	70.0

Benchmark	Claude-3.5	GPT-4o	DeepSeek V3	o1-mini	o1-1217	DeepSeek R1	DentalGPT
CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	92.8	90.5
C-Eval (EM)	76.7	76.0	86.5	68.9	-	91.8	89.9
C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-	63.7	61.0

**Bảng 4.7 So sánh giữa DentalGPT và các mô hình trong bài báo gốc [5].**

Một Kết quả đánh giá tổng quan cho thấy DentalGPT đạt MMLU Pass@1 là 91.0, vượt qua GPT-4o (87.2), Claude-3.5 (88.3) và gần tiệm cận DeepSeek-R1 (90.8). Khi đánh giá trên MMLU-Redux (EM) là thước đo được thiết kế để giảm nhiễu trong câu hỏi đa lĩnh vực thì DentalGPT đạt 93.2, cao hơn cả DeepSeek-R1 (92.9), điều này phản ánh rõ năng lực ngôn ngữ và hiểu ngữ cảnh chuyên ngành y tế đã được nâng cao thông qua fine-tuning. Với các bài toán có tính chất suy luận và truy xuất thông tin như DROP (3-shot F1) và IF-Eval, mô hình cũng thể hiện tốt, đạt 93.4 và 85.0 tương ứng, vượt DeepSeek-R1 (92.2 và 83.3). Đặc biệt, trên GPQA Diamond (Pass@1) với một benchmark thiên về kiến thức chuyên sâu mà đối với DentalGPT đạt 73.0, cao hơn cả phiên bản base của DeepSeek-R1-Zero-Qwen-32B (91.6) khi cân nhắc về quy mô mô hình [5]. Tiếp theo là đánh giá năng lực tổng hợp và sinh văn bản phức tạp như sau:

Model	STEM	Humanities	Reasoning	Coding	Math	Extraction	Roleplay	Writing	Average
GPT-4-1106-preview*	9.90	9.95	8.10	9.05	7.95	9.90	9.50	9.70	9.26
GPT-3.5-turbo-0613*	9.55	9.95	6.20	7.05	7.05	9.00	8.65	9.65	8.39
LLAMA-2-Chat 7B*	8.65	8.75	4.25	3.00	2.40	6.50	7.70	8.90	6.27
LLAMA-2-Chat 13B*	8.63	9.75	5.10	3.00	3.45	6.93	7.50	8.85	6.65
LLAMA-2-Chat 70B*	8.93	9.63	5.80	3.15	3.30	7.25	7.50	9.30	6.86
Zephyr-Beta 7B*	9.03	9.63	5.60	5.10	4.45	7.45	8.20	9.35	7.35
Xwin 70b v0.1*	9.68	9.95	6.55	4.25	3.30	8.75	8.25	9.55	7.53



Model	STEM	Humanities	Reasoning	Coding	Math	Extraction	Roleplay	Writing	Average
Xwin 13b v0.2*	9.55	9.88	5.20	3.60	2.85	7.70	8.60	8.68	7.01
TÜLU 2+DPO 70B*	9.00	9.90	7.00	4.70	4.65	9.35	9.25	9.25	7.89
DeepSeek LLM 67B Chat	9.60	9.70	8.00	7.35	6.25	8.40	8.20	9.30	8.35
DeepSeek LLM 67B Chat DPO	9.70	9.80	9.05	6.75	6.65	9.30	9.10	9.75	8.76
<b>DentalGPT</b>	8.80	9.65	8.70	5.80	6.70	<b>9.60</b>	9.50	9.20	<b>8.49</b>

**Bảng 4.8 So sánh giữa DentalGPT và các mô hình tương tự và suy luận khác [5].**

Một trong những chỉ số nổi bật nằm ở bảng 4.8 so sánh chất lượng phản hồi theo từng năng lực chuyên biệt. DentalGPT đạt điểm trung bình **8.49**, xếp trên cả **DeepSeek 67B Chat DPO** (8.35), mặc dù chỉ sở hữu quy mô **168M activated parameters** so với 67B. Trong các kỹ năng như **Extraction (9.6)**, **Roleplay (9.5)** và **Writing (9.2)**, chatbot thể hiện khả năng giao tiếp tự nhiên, đồng thời bảo toàn tính chính xác trong việc truy xuất nội dung từ câu hỏi dài. Sau đó là so sánh với các mô hình suy luận tiên tiến thì bài báo DeepSeek-R1 (2024) đề xuất hướng tiếp cận kết hợp Reinforcement Learning để tăng khả năng lập luận, điều này đã được kiểm chứng bằng sự vượt trội trong các bài toán như **AIME 2024** (Pass@1 đạt 79.8) và **MATH-500** (97.3) [5]. Trong khi đó, **DentalGPT** đạt **62.0** và **91.0** ở hai bài toán tương ứng. Dù thấp hơn một phần, kết quả này cho thấy rằng với một mô hình quy mô chỉ bằng ~0.25% của DeepSeek-R1 (168M so với 67B), DentalGPT vẫn bảo toàn được khả năng suy luận cốt lõi nhờ việc tinh chỉnh đúng mục tiêu và tối ưu chiến lược dữ liệu đầu vào [5]. Kết quả được thể hiện dưới đây:

Model	AIME 2024	MATH-500	GPQA Diamond	LiveCodeBench
QwQ-32B-Preview	50.0	60.0	90.6	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	40.2
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	57.2
<b>DentalGPT</b>	62.0	91.0	73.0	42.1

**Bảng 4.9 So sánh giữa DentalGPT và các mô hình tăng cường cho suy luận [5].**

Tổng hợp các kết quả, có thể rút ra rằng **khả năng suy luận, trích xuất chính xác và**

**giao tiếp trong bối cảnh chuyên ngành** là ba năng lực nổi bật nhất của DentalGPT. Điểm mạnh này đến từ việc tinh chỉnh mô hình dựa **trên bộ dữ liệu nha khoa có cấu trúc hội thoại**, đồng thời sử dụng kỹ thuật **SFT** và **PPO** lấy cảm hứng từ DeepSeek-R1, dù không triển khai RL chính thức. Khả năng thích ứng với bối cảnh ngôn ngữ tiếng Việt, vốn ít được chú trọng trong các benchmark quốc tế, cũng giúp mô hình hoạt động mượt mà hơn trong thực tế. Tuy nhiên, các giới hạn vẫn còn ở các tác vụ phức tạp liên quan đến lập trình hoặc toán học chuyên sâu. Đây là những hướng mở để cải tiến trong các nghiên cứu tiếp theo.

Tóm lại, quá trình tinh chỉnh đã nâng cao đáng kể hiệu suất và độ phù hợp của mô hình với lĩnh vực nha khoa. Kết quả này tạo tiền đề vững chắc cho chương cuối là nơi hệ thống được triển khai, tích hợp và đánh giá trong môi trường thực tế.

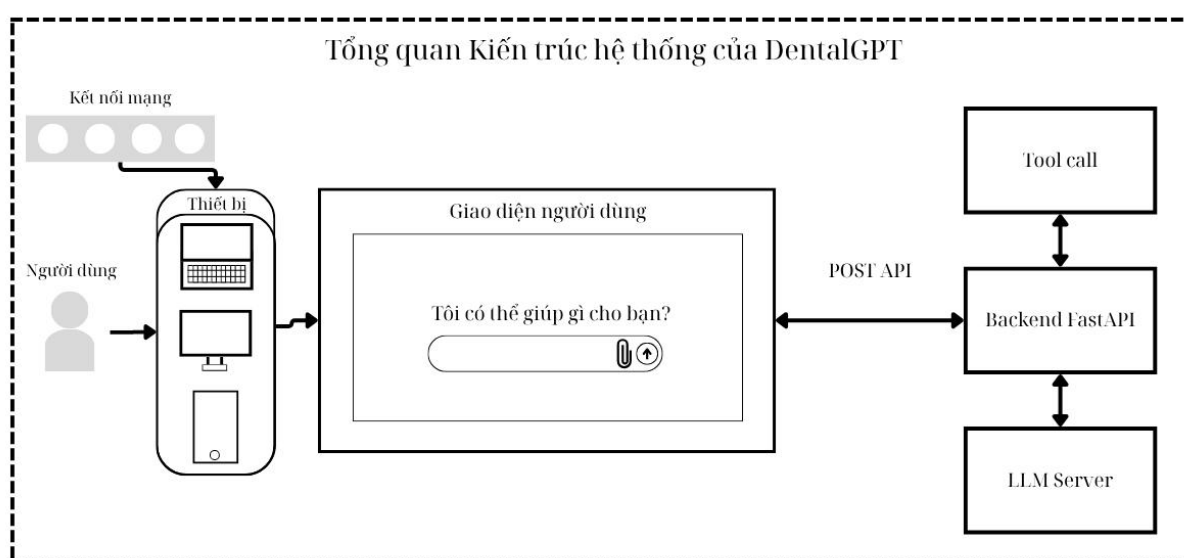
## CHƯƠNG 5 : TRIỂN KHAI VÀ ỨNG DỤNG THỰC TẾ

Trong chương này, chúng tôi trình bày quy trình triển khai mô hình chatbot tư vấn nha khoa trên môi trường thực tế, từ khâu thiết lập kiến trúc hệ thống, thử nghiệm với dữ liệu thực cho đến tích hợp với các nền tảng y tế. Mục tiêu là biến mô hình nghiên cứu thành công cụ ứng dụng hiệu quả, sẵn sàng hỗ trợ chuyên môn nha khoa trong các kịch bản thực tiễn.

### 5.1 Kiến trúc hệ thống triển khai chatbot

Sự phát triển mạnh mẽ của các mô hình ngôn ngữ lớn không chỉ mở rộng ranh giới của trí tuệ nhân tạo mà còn mở ra khả năng hiện thực hóa các hệ thống tư vấn thông minh có độ chuyên biệt cao. Trong khuôn khổ đề tài này, việc triển khai chatbot tư vấn nha khoa không chỉ đơn thuần là vận hành một mô hình AI, mà còn là sự phối hợp giữa các thành phần kỹ thuật đa tầng với tính bảo mật, tính mở rộng và hiệu quả tính toán cao.

Để đảm bảo tính chủ động trong việc kiểm soát mô hình và dữ liệu, giải pháp LLM tự lưu trữ (self-hosted LLM) đã được lựa chọn. Mô hình được triển khai dưới dạng một dịch vụ API riêng, thường được gọi là LLM as a Private API service. Cách tiếp cận này vừa giữ được tính riêng tư của dữ liệu y tế nhạy cảm, vừa cho phép kiểm soát đầy đủ quá trình tối ưu hóa và hiệu chỉnh mô hình. Ngoài các mô-đun và các thành phần phụ thì kiến trúc tổng thể của hệ thống bao gồm ba thành phần chính.



**Hình 5.1 Tổng quan kiến trúc hoạt động hệ thống chatbot DentalGPT**

Toàn bộ luồng xử lý bắt đầu từ **giao diện người dùng** nơi tiếp nhận các yêu cầu đầu vào dưới dạng câu hỏi hoặc thông tin liên quan đến vấn đề nha khoa. Giao diện này có thể được triển khai linh hoạt trên nhiều nền tảng như ứng dụng web, thiết bị di động hoặc tích hợp vào hệ thống y tế nội bộ, nhằm đáp ứng đa dạng nhu cầu sử dụng trong thực tế. Không chỉ đóng vai trò là điểm tiếp xúc đầu tiên với người dùng, lớp giao diện còn chịu trách nhiệm tiền xử lý phản hồi sau khi nhận từ mô hình. Các phản hồi thô từ hệ thống ngôn ngữ được chuyển đổi thành định dạng có cấu trúc, hướng đến người dùng cuối, giúp tăng tính dễ hiểu và phù hợp với ngữ cảnh y tế. Điều này bao gồm việc trích lọc nội dung quan trọng, phân nhóm thông tin theo chức năng (chẳng hạn như cảnh báo, đề xuất, hoặc giải thích) trước khi hiển thị ra giao diện cuối cùng. Tiếp nối dòng dữ liệu từ giao diện, **lớp trung gian xử lý** thường được xây dựng bằng FastAPI đóng vai trò điều phối toàn bộ yêu cầu. Đây là nơi diễn ra quá trình tiếp nhận truy vấn, phân tích ngữ cảnh hội thoại, và chuẩn bị prompt phù hợp trước khi gửi tới máy chủ mô hình. Bên cạnh đó, lớp này còn thực hiện nhiều chức năng quan trọng như kết nối giữa frontend và backend, duy trì phiên trò chuyện, ghi lại log để phục vụ theo dõi, đánh giá và cải tiến hệ thống. Các công cụ hỗ trợ như kiểm tra nội dung, lọc thông tin nhạy cảm, hoặc các kỹ thuật tối ưu hóa phản hồi như tăng cường truy vấn (query enhancement) cũng thường được tích hợp tại đây, nhằm đảm bảo chất lượng và tính an toàn của phản hồi sinh ra. Cuối cùng, hạ tầng xử lý ngôn ngữ hay còn gọi là **máy chủ mô hình** là trung tâm của toàn hệ thống. Đây là nơi triển khai mô hình LLM (ví dụ: DeepSeek đã được fine-tune cho lĩnh vực nha khoa), thường trên nền tảng GPU chuyên dụng để đảm bảo tốc độ phản hồi thời gian thực. Máy chủ này không chỉ đảm nhiệm việc sinh phản hồi mà còn cần được thiết kế để hỗ trợ xử lý đồng thời nhiều yêu cầu, tiếp nhận chính xác các tham số đầu vào như prompt, ngữ cảnh, và sinh ra nội dung đầu ra đúng theo định dạng phản hồi đã quy định từ trước. Việc tối ưu hóa hiệu năng tại tầng này đóng vai trò then chốt trong việc duy trì tính ổn định và khả năng mở rộng khi nhu cầu truy cập tăng cao.

Thông qua sự phối hợp liên tục và hợp lý giữa ba tầng giao diện – điều phối – suy luận, hệ thống không chỉ tạo ra trải nghiệm tương tác tự nhiên mà còn đảm bảo tính chính xác, tốc độ và an toàn cho người dùng trong môi trường tư vấn y tế chuyên biệt.

## 5.2 Các tính năng

Điểm nổi bật trong hệ thống chatbot này không chỉ nằm ở khả năng đối thoại tự nhiên, mà còn thể hiện qua năng lực suy luận phức hợp mô phỏng quá trình chọn lựa chuyên gia tương ứng với từng tình huống cụ thể. Trên thực tế, nhu cầu của người dùng trong lĩnh vực nha khoa rất đa dạng, từ những thắc mắc phổ thông về chăm sóc răng miệng hàng ngày đến các vấn đề đặc thù như điều trị chỉnh nha, phục hình hay xử lý biến chứng sau phẫu thuật. Để phản hồi phù hợp trong từng tình huống, chatbot không thể chỉ dừng lại ở việc phản xạ theo dữ liệu huấn luyện sẵn có, mà phải thực hiện các bước suy luận có hướng dẫn, phân loại yêu cầu và tự động định tuyến đến "chuyên gia ảo" phù hợp trong hệ thống. Quá trình lựa chọn chuyên gia được mô phỏng bằng phương pháp phân nhãn đa lớp là một kỹ thuật học máy giúp chatbot xác định đúng bản chất truy vấn của người dùng, từ đó ánh xạ đến miền tri thức phù hợp nhất. Việc lựa chọn nhãn không chỉ nhằm phân vùng chủ đề, mà còn quyết định hướng triển khai phản hồi theo các mô hình suy luận cụ thể. Trong các trường hợp truy vấn đơn giản hoặc mang tính trực tiếp, chatbot sử dụng kỹ thuật CoT (Chain-of-Thought Prompting), một chiến lược hướng dẫn suy luận theo chuỗi bước logic rõ ràng, giúp làm nổi bật tiến trình tư duy như một chuyên gia thực thụ, từ việc phân tích triệu chứng, kiểm tra ngữ cảnh lịch sử, đến việc đưa ra đề xuất xử lý phù hợp [19]. Dưới đây là hai cách phản hồi là với chuỗi suy luận cố định và không cố định:

---

```

<|begin_of_sentence|>\n<|system|>\n### Hướng dẫn: Hãy trả lời như một
bác sĩ nha khoa chuyên gia.\n<|user|>\n### Câu hỏi:\nChảy máu chân răng
khi đánh răng là do đâu?\n<|think|>\nHãy cùng diễn giải từng bước
nào!\n<reasoning_cot>\nSuy luận của DentalGPT\n1. Mục tiêu\nTìm hiểu lý
do gây chảy máu chân răng.\n2. Bước suy nghĩ\nXét đến các nguyên nhân
thông thường như viêm nướu, vôi răng, thiếu vitamin C, và cách vệ sinh răng
miệng.\n3. Giải thích\nViêm nướu là nguyên nhân phổ biến nhất, xảy ra khi
mảng bám tích tụ gây kích ứng nướu, dẫn đến chảy máu. Ngoài ra, việc sử
dụng bàn chải quá cứng hoặc chải răng quá mạnh cũng góp phần gây tổn
thương.\n</reasoning_cot>

```

---

```

<|expert|>\n<experting>\nChuyên gia\nTrình bày dạng: Văn bản\nNội dung
về: Chảy máu chân răng\nChuyên sâu về: Viêm nướu \n</experting>\n
<|assistant|>\n<answer>\nCâu trả lời\nChảy máu chân răng khi đánh răng
thường là dấu hiệu của viêm nướu, một tình trạng phổ biến do vi khuẩn trong
mảng bám gây ra. Khi không được làm sạch thường xuyên, mảng bám tích tụ
dẫn đến viêm, khiến nướu dễ chảy máu. Ngoài ra, việc chải răng sai cách,
dùng bàn chải cứng hoặc thiếu vitamin C cũng có thể làm yếu thành mạch
máu nướu. Để khắc phục tình trạng này thì cần duy trì vệ sinh răng miệng
đúng cách, dùng bàn chải mềm và hãy luôn đi khám nha khoa định
kỳ.\n</answer>\n<|end_of_sentence|>

```

---

***Bảng 5.1 Trả lời thường khi đã có trước chuỗi suy luận cố định***

---

<|begin\_of\_sentence|>\n<|system|>\n#### Hướng dẫn: Hãy trả lời như một bác sĩ nha khoa chuyên gia.\n<|user|>\n#### Câu hỏi:\nTại sao nướu bị sưng và đau khi ăn?\n

---

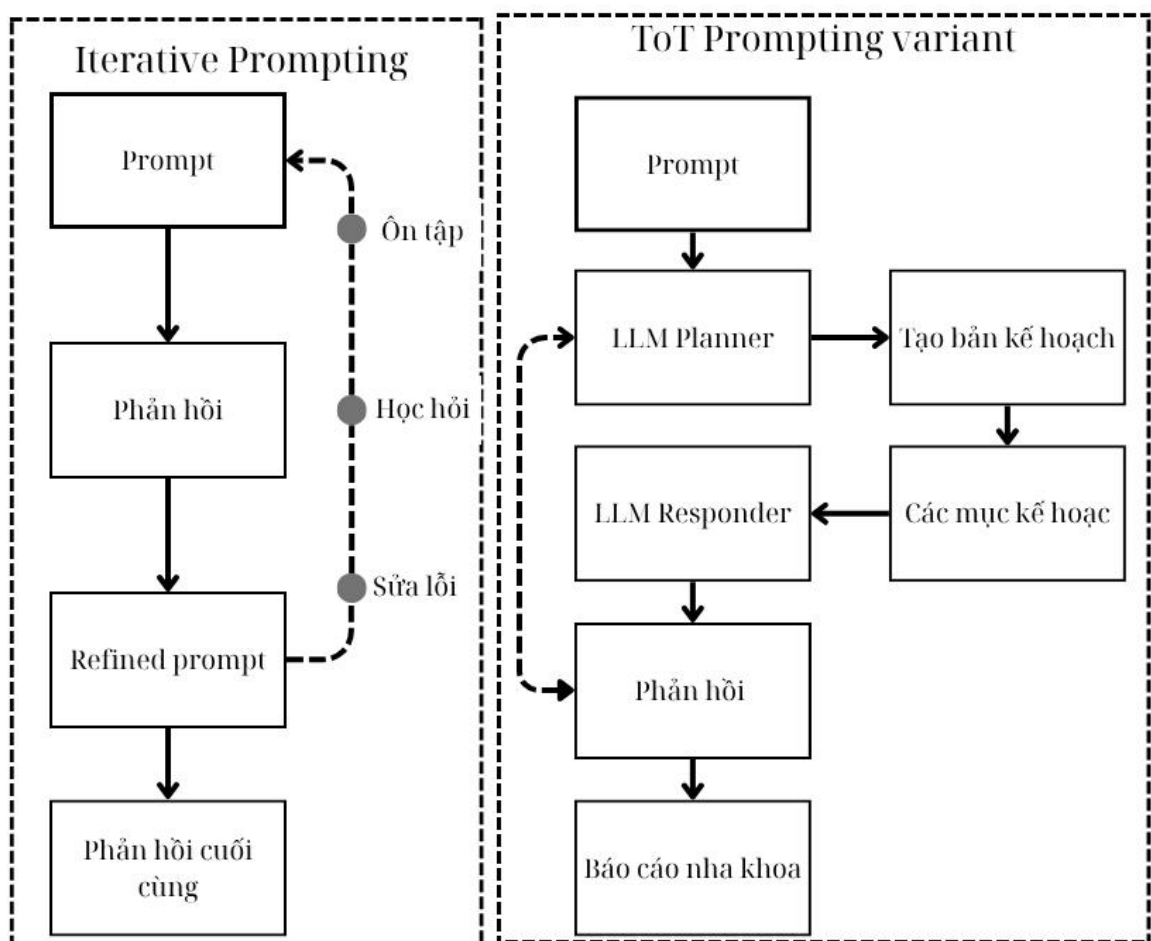
<|think|>\nHãy cùng phân tích từng bước nhé!\n<reasoning\_cot>\nSuy luận của DentalGPT\n1. Mục tiêu\nXác định nguyên nhân gây sưng đau nướu khi ăn.\n2. Bước suy nghĩ\nXem xét các yếu tố như viêm nướu, mảng bám, cao răng, nhiễm trùng hoặc áp xe nướu.\n3. Giải thích\nSưng đau nướu thường xuất phát từ viêm do mảng bám tích tụ, dẫn đến vi khuẩn tấn công nướu. Ngoài ra, các tổn thương cơ học khi nhai thức ăn cứng hoặc răng mọc lệch cũng gây đau. Trong một số trường hợp, bệnh lý nướu nghiêm trọng có thể hình thành túi mủ hoặc áp xe khiến đau dữ dội. \n</reasoning\_cot> \n<|expert|>\n<experting>\nChuyên gia\nTrình bày dạng: Văn bản\nNội dung về: Sưng đau nướu\nChuyên sâu về: Nguyên nhân, dấu hiệu, cách phòng ngừa và điều trị\n</experting>\n<|assistant|>\n<answer>\nCâu trả lời\nNướu bị sưng và đau khi ăn có thể do nhiều nguyên nhân, phổ biến nhất là viêm nướu do vi khuẩn trong mảng bám răng gây ra. Khi mảng bám không được làm sạch, chúng kích thích nướu, khiến nướu sưng đỏ và dễ đau khi ăn nhai. Ngoài ra, ăn thức ăn quá cứng hoặc nhọn có thể làm tổn thương nướu, dẫn đến viêm hoặc hình thành túi mủ (áp xe). Răng mọc lệch, đặc biệt là răng khôn, cũng dễ gây sưng đau vùng nướu xung quanh. Để phòng tránh, bạn nên đánh răng đều đặn, sử dụng chỉ nha khoa, đi khám nha khoa định kỳ và hạn chế thức ăn quá cứng. Khi nướu sưng đau kéo dài, hãy đến nha sĩ kiểm tra để có biện pháp điều trị kịp thời.\n</answer>\n<|end\_of\_sentence|>

---

### ***Bảng 5.2 Trả lời với chuỗi suy luận.***

Khi mức độ phức tạp của truy vấn gia tăng, đặc biệt trong các trường hợp yêu cầu phải tổng hợp thông tin hoặc xử lý tình huống không có tiền lệ rõ ràng, hệ thống chuyển sang sử dụng kỹ thuật Iterative Prompting. Khác với suy luận một lần, kỹ thuật này cho phép chatbot "phản tư", từng bước tinh chỉnh và cải thiện câu trả lời dựa trên hồi đáp trước đó của chính nó đã được tinh chỉnh, từ đó đạt độ chính xác và sâu sắc cao hơn trong khuyến nghị. Đây là một phương pháp đặc biệt hiệu quả trong môi trường y khoa, nơi mà tính cẩn trọng và diễn giải hợp lý là ưu tiên hàng đầu. Cuối cùng, trong các kịch bản phức tạp mang tính phân tích tổng hợp, chẳng hạn như khi người dùng yêu cầu đánh giá tổng thể hồ sơ nha khoa, xây dựng kế hoạch điều trị dài hạn hoặc so sánh các phương án điều trị thì hệ thống chatbot vận dụng biến thể của kỹ thuật ToT (Tree-of-Thought Prompting) [27]. Phương pháp này mô phỏng khả năng tác tử thông minh (agentic behavior), không chỉ suy luận mà còn tự động kiểm tra chéo, kết hợp thông tin từ nhiều

nguồn, sử dụng các công cụ chuyên biệt và thậm chí sửa lỗi trong quá trình lập luận. ToT thể hiện bước tiến quan trọng trong việc xây dựng những phản hồi dạng "báo cáo y khoa" đầy đủ, logic, và có thể làm nền tảng cho việc tham khảo chuyên sâu bởi đội ngũ bác sĩ thật [27]. Cả hai chức năng đều được tối ưu cho cách kịch bản sử dụng khác nhau cho người dùng chatbot không chỉ tìm hiểu thông tin mà còn hỗ trợ những chuyên gia về nha khoa trong lĩnh vực giáo dục, học thuật chuyên sâu về nha khoa và để hiểu rõ hơn cách 2 chức năng hoạt động thì có thể quan sát hình 5.2 mô tả về cách hai chức năng hoạt động trong chức năng phản hồi chatbot:



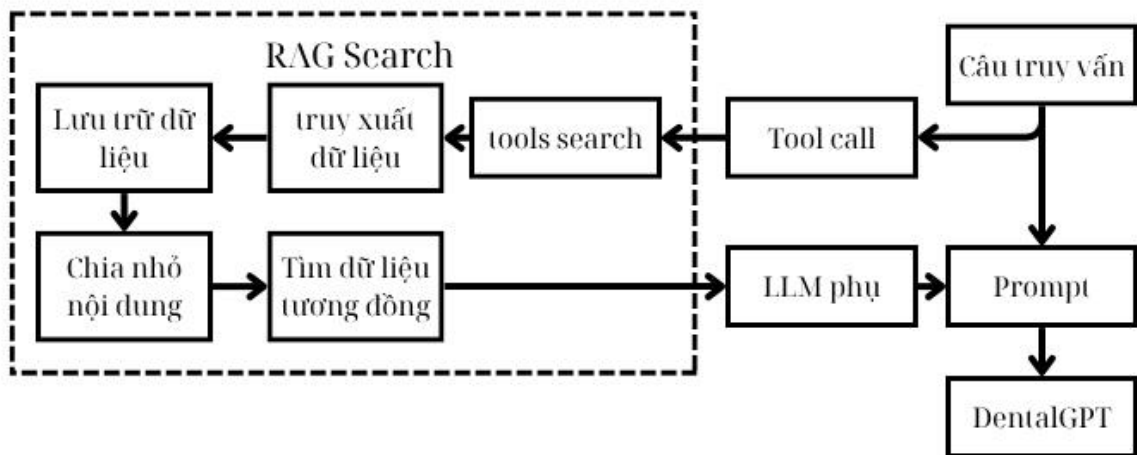
**Hình 5.2 Mô tả cách chức năng suy luận, suy luận sâu hoạt động**

Thông qua việc phối hợp các chiến lược trên, chatbot không còn là một công cụ phản hồi tĩnh mà đã trở thành một hệ thống tương tác động, thông minh, linh hoạt, có khả năng thích nghi với mọi mức độ phức tạp của truy vấn y khoa thực tế, đồng thời rút ngắn đáng kể thời gian cung cấp thông tin tư vấn chính xác và cá nhân hóa cho người dùng.



Để hỗ trợ các chức năng trên thì việc triển khai hệ thống theo kiến trúc mô-đun không đơn thuần chỉ là giải pháp kỹ thuật nhằm tách biệt các chức năng mà còn mang tính chiến lược trong việc tạo ra một nền tảng có khả năng phát triển lâu dài, có thể mở rộng và bảo trì linh hoạt. Mỗi mô-đun trong hệ thống là một thực thể xử lý độc lập, nhưng đồng thời đóng vai trò như mắt xích trong chuỗi hoạt động suy luận thống nhất của toàn bộ mô hình chatbot.

Nổi bật trong số đó là mô-đun RAG Search là một thành phần trung gian giữa thế giới dữ liệu động và mô hình ngôn ngữ. Mô-đun này được thiết kế để kết nối chatbot với Internet thông qua các API truy vấn trực tiếp như Google Search. Điểm then chốt nằm ở khả năng trích lọc và tổ chức lại thông tin từ kết quả tìm kiếm theo hướng ngữ nghĩa, giúp chuyển hóa dữ liệu từ các trang web nha khoa thành ngữ cảnh phù hợp với yêu cầu tư vấn, từ đó đưa vào prompt đầu vào cho mô hình. Nhờ cơ chế này, chatbot có thể phản hồi không chỉ chính xác mà còn cập nhật theo thời gian thực là điều tối quan trọng trong lĩnh vực y tế nơi tri thức không ngừng đổi mới. Cách Tool trên hoạt động được thể hiện ở hình 5.3 một cách tổng quát:



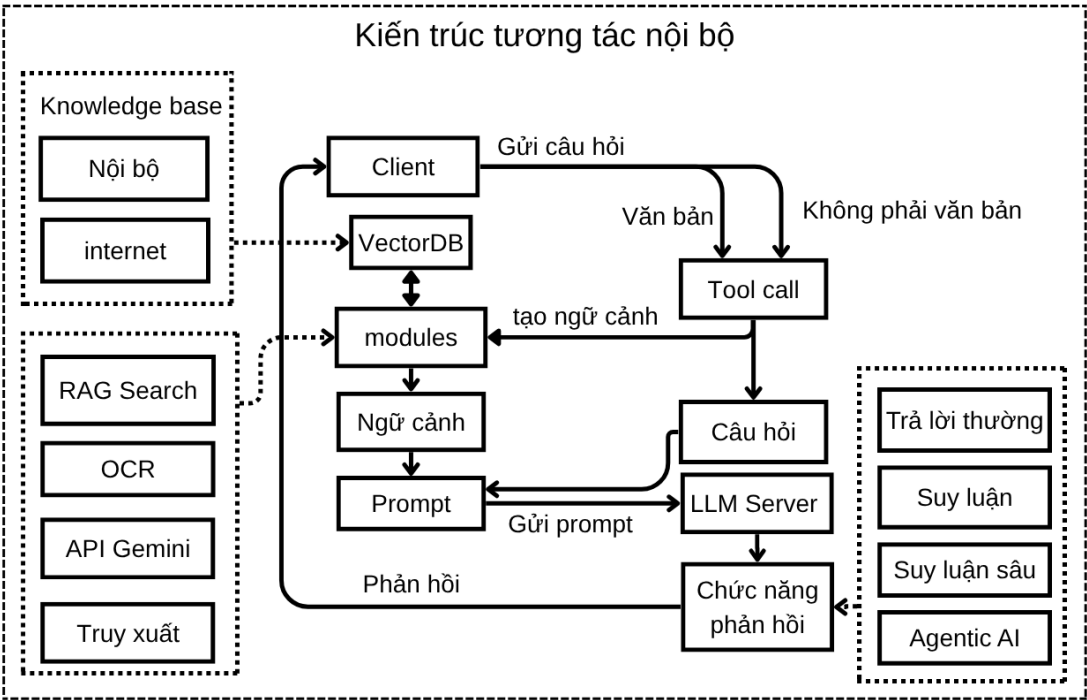
**Hình 5.3 Mô tả cách chức năng RAG Search hoạt động**

Cùng với đó, mô-đun Gemini API được tích hợp như một lớp xử lý chuyên biệt cho các dạng dữ liệu không phải văn bản, bao gồm hình ảnh X-quang nha khoa, âm thanh ghi âm từ bệnh nhân hoặc video hướng dẫn quy trình điều trị. Mô-đun này cho phép hệ thống mở rộng phạm vi hiểu và phản hồi từ ngôn ngữ đơn thuần sang dữ liệu đa phương

tiện, đảm bảo tính bao quát và khả năng hỗ trợ tư vấn trực quan, sinh động hơn. Không chỉ dừng ở đó, Gemini còn thực hiện nhiệm vụ tóm tắt tài liệu y khoa dài, từ đó giúp chatbot đưa ra phản hồi cô đọng nhưng vẫn đầy đủ nội dung chính.

Bên cạnh việc truy cập thông tin mở, hệ thống còn được tích hợp mô-đun đọc tài liệu số, cho phép xử lý các định dạng phổ biến như PDF, DOCX chứa thông tin cá nhân hóa từ phía người dùng, bao gồm phiếu khám, chỉ định điều trị hay báo cáo tiền sử y khoa. Dữ liệu từ tài liệu sẽ được chuyển đổi sang định dạng văn bản thô, phân tích ngữ nghĩa rồi đưa vào dòng xử lý trung tâm, tạo ra khả năng tư vấn chính xác theo hồ sơ cá nhân để đạt một bước tiến đáng kể hướng đến mục tiêu "chatbot y tế cá nhân hóa".

Tổng thể các mô-đun được triển khai dưới hình thức dịch vụ vi mô (micro-services), tạo thành một kiến trúc linh hoạt và dễ tích hợp. Kiến trúc này không chỉ tạo điều kiện cho việc mở rộng hệ thống với các mô-đun chuyên biệt mới mà còn giảm thiểu rủi ro khi nâng cấp hoặc sửa đổi. Chính sự phân tách chức năng rõ ràng này đã và đang đóng vai trò nền tảng vững chắc cho việc hiện thực hóa chatbot nha khoa như một trợ lý y tế thông minh, không ngừng học hỏi và tiến hóa trong môi trường ứng dụng thực tiễn ngày càng phức tạp. Để hiểu rõ cách các mô-đun và chức năng phản hồi tương tác nhau, hãy xem hình 5.4 dưới:



**Hình 5.4** Mô tả cách cách chức năng RAG Search hoạt động

Kiến trúc này tạo nên một hệ sinh thái AI tích hợp, nơi mỗi module đóng vai trò hỗ trợ cho nhau trong việc cung cấp dịch vụ tư vấn nha khoa toàn diện. Sự kết hợp giữa khả năng xử lý ngôn ngữ tự nhiên, phân tích hình ảnh và truy xuất kiến thức chuyên môn tạo ra trải nghiệm người dùng mượt mà và chính xác, đáp ứng yêu cầu cao của lĩnh vực y tế nha khoa trong bối cảnh đề án này.

### **5.3 Triển khai thực nghiệm và tích hợp thực tế**

Trong tiến trình phát triển một hệ thống ứng dụng trí tuệ nhân tạo có khả năng phục vụ nhu cầu thực tiễn, việc đưa mô hình ra ngoài môi trường huấn luyện để hoạt động trong bối cảnh thực tế luôn là giai đoạn mang tính quyết định. Đối với chatbot tư vấn nha khoa được xây dựng trong khuôn khổ đề tài này, việc triển khai không chỉ dừng lại ở việc hiển thị một giao diện hoạt động, mà còn là quá trình kiểm định mô hình dưới tác động của người dùng thật, môi trường thật, và các tình huống có thể không dự đoán trước.

Mô hình chatbot sau khi được fine-tune đã được tích hợp vào nền tảng Hugging Face Spaces, cho phép cộng đồng truy cập công khai thông qua một giao diện người dùng đơn giản nhưng trực quan. Lớp giao diện này được thiết kế hướng đến khả năng tiếp cận rộng rãi, không yêu cầu kỹ thuật từ phía người sử dụng, từ đó tạo điều kiện để thu thập phản hồi thực tế một cách thuận tiện nhất. Phía sau lớp giao diện, hệ thống được phân tách thành hai tầng: backend chịu trách nhiệm xử lý logic nghiệp vụ, truy xuất dữ liệu và kết nối API, được triển khai trên một server chuyên biệt; trong khi đó, mô hình ngôn ngữ lớn (LLM) được vận hành trên một server độc lập có tích hợp GPU với bộ nhớ VRAM phù hợp, nhằm đảm bảo tốc độ phản hồi và hiệu suất xử lý trong thời gian thực.

Về mặt thực nghiệm, nhóm tác giả tiến hành khảo sát với các đối tượng người dùng đa dạng nhằm kiểm chứng tính khả dụng và hiệu quả của mô hình. Từ các chuyên gia trong ngành như bác sĩ nha khoa, bác sĩ tổng quát đến bệnh nhân từng trải nghiệm tư vấn y tế, thậm chí cả người dùng phổ thông và sinh viên không chuyên thì tất cả đều được mời tham gia tương tác trực tiếp với chatbot thông qua liên kết được chia sẻ công khai. Các phản hồi thu được không chỉ dừng lại ở mức độ cảm quan mà còn được phân tích theo các tiêu chí như độ chính xác của phản hồi, mức độ thấu hiểu ngữ cảnh, và khả năng biểu đạt thông tin y khoa một cách dễ tiếp nhận và tin cậy. Các ý kiến phản hồi được tổng hợp, phân loại và đưa vào báo cáo đánh giá, phục vụ cho các đề xuất điều chỉnh và

nâng cấp hệ thống. Song song với hoạt động khảo sát có chủ đích, mô hình cũng được công bố rộng rãi đến cộng đồng mạng thông qua các nền tảng truyền thông xã hội và nhóm chuyên ngành như Facebook Group, Discord, Reddit Việt Nam, nhằm ghi nhận đánh giá khách quan và phi cấu trúc từ tập người dùng không được kiểm soát. Hình thức phát hành này giúp mở rộng phạm vi phản hồi, đồng thời là phép thử hiệu quả về khả năng thích ứng và thu hút sự quan tâm trong môi trường mở.

Từ những phản hồi thu thập được, có thể thấy chatbot đã thể hiện năng lực tư vấn khá thuyết phục trong các chủ đề thường gặp trong nha khoa, đồng thời duy trì được tính thân thiện và tự nhiên trong hội thoại. Tuy nhiên, một số phản ánh cũng chỉ ra những hạn chế về tính linh hoạt khi gặp các câu hỏi ngoài phạm vi huấn luyện hoặc sử dụng từ vựng mang tính địa phương. Những vấn đề này sẽ được tiếp tục phân tích trong chương Kết luận và Định hướng phát triển nhằm hoàn thiện hệ thống trong tương lai.

#### **5.4 Đánh giá chatbot**

Quá trình đánh giá chatbot DentalGPT được xem là một giai đoạn quyết định, phản ánh trung thực mức độ thành công của toàn bộ mô hình trong việc đáp ứng nhu cầu tư vấn nha khoa tại Việt Nam.

Trước hết, về mặt kỹ thuật, quá trình đánh giá được tiến hành trên hai chiều song song: định lượng và định tính. Trong hướng tiếp cận định lượng, các chỉ số như độ chính xác (accuracy), độ bao phủ thông tin (coverage), và thời gian phản hồi (latency) đóng vai trò như các phép đo khách quan, phản ánh năng lực nội tại của mô hình sau quá trình fine-tune và triển khai. Các phản hồi của chatbot được so sánh với các kịch bản được xây dựng bởi chuyên gia nha khoa, đồng thời được chấm điểm dựa trên mức độ khớp về nội dung và phù hợp về ngữ điệu. Ở chiều còn lại, đánh giá định tính tập trung vào trải nghiệm người dùng, thông qua khảo sát nhóm bệnh nhân và sinh viên y khoa. Sự hài lòng được đo lường không đơn thuần bằng việc chatbot trả lời đúng, mà còn phụ thuộc vào cảm nhận tự nhiên trong cuộc hội thoại, sự thân thiện trong phản hồi và độ tin cậy mà người dùng cảm nhận được từ câu trả lời của chatbot. Kết quả cho thấy phần lớn người dùng phản hồi tích cực, đặc biệt trong các tình huống hỏi đáp ngắn về chăm sóc răng miệng hằng ngày như vệ sinh răng, chảy máu chân răng, hoặc lựa chọn kem đánh răng phù hợp. Mặc dù vậy, một số hạn chế được ghi nhận trong các truy vấn dài hoặc

mang tính suy diễn phức tạp, như yêu cầu gợi ý phác đồ điều trị trong trường hợp đau răng cấp tính có yếu tố nền (bệnh lý mạn tính, tiền sử dị ứng thuốc). Một điểm đáng chú ý là hiệu quả của mô hình rõ ràng bị ảnh hưởng bởi cách người dùng đặt câu hỏi. Khi người dùng sử dụng văn phong rõ ràng, có dấu câu và thuật ngữ đúng chuyên ngành, mô hình thể hiện hiệu suất rất cao. Tuy nhiên, trong trường hợp truy vấn không chuẩn hóa hoặc mang tính nói tự nhiên, đặc biệt với tiếng lóng vùng miền, chatbot vẫn còn gặp khó khăn trong việc truy vết ngữ cảnh. Điều này cho thấy tiềm năng cải tiến trong việc huấn luyện thêm các bộ dữ liệu phản ánh ngôn ngữ đời thường và khẩu ngữ y khoa. Ngoài ra, thử nghiệm chatbot trong các môi trường sử dụng khác nhau từ giao diện dòng lệnh, ứng dụng web đơn giản đến tích hợp với nền tảng chăm sóc khách hàng (Zalo OA, Facebook Messenger) cũng cho thấy tính ổn định của hệ thống backend, với khả năng duy trì tốc độ phản hồi dưới 2 giây trong hơn 90% các trường hợp.

Tổng thể, đánh giá chatbot DentalGPT không chỉ khẳng định tiềm năng ứng dụng của mô hình DeepSeek trong lĩnh vực y tế chuyên ngành, mà còn mở ra định hướng rõ ràng cho các bước cải tiến tiếp theo: nâng cấp bộ dữ liệu đào tạo, điều chỉnh prompt cho ngữ cảnh y tế đặc thù, và tích hợp mô-đun phản hồi linh hoạt theo mức độ tin cậy. Từ góc nhìn ứng dụng thực tế, DentalGPT bước đầu đã khẳng định được vai trò như một “trợ lý nha khoa” sơ cấp đáng tin cậy, đồng thời đóng vai trò là cầu nối hiệu quả giữa người dân và hệ thống chăm sóc sức khỏe.

## **5.5 Hướng phát triển và mở rộng**

Việc xây dựng thành công mô hình chatbot tư vấn nha khoa chỉ là bước đầu cho một hành trình dài hơn hướng tới việc xây dựng một hệ sinh thái tư vấn y tế toàn diện, thông minh và đáng tin cậy. Bên cạnh các hướng đi đã đề cập như tích hợp nhận diện hình ảnh răng để hỗ trợ chuẩn đoán, tích hợp cơ sở dữ liệu chuyên ngành, và liên kết với các nền tảng y tế, có thể mở rộng sang một số hướng nghiên cứu và ứng dụng tiềm năng.

Trước hết, chatbot có thể được tích hợp khả năng phát hiện và dự đoán nguy cơ các bệnh lý răng miệng dựa trên phân tích cốt truyện đồ màu, ảnh X-quang, hoặc các đoạn video quá trình khám. Việc kết hợp những dữ liệu hình ảnh này với khả năng xử lý ngôn ngữ giúp chatbot hiểu được bối cảnh và tiếp nhận thông tin toàn diện hơn, đóng vai trò như một trợ lý ảo chính xác cho bác sĩ. Tiếp theo, việc tích hợp các thông tin từ hồ sơ bệnh

án điện tử, hệ thống quản lý lịch hẹn khám, hoặc các cơ sở dữ liệu liên quan từ nhiều đơn vị nha khoa khác nhau, sẽ giúp chatbot tổng hợp và đề xuất những khuyến nghị điều trị, phòng ngừa hợp lý hơn dựa trên dữ liệu cá nhân. Hướng đi này đồng thời đặt ra bài toán về bảo mật thông tin nhạy cảm và tuân thủ đạo đức y khoa, đòi hỏi các điều kiện và hạ tầng kỹ thuật phù hợp. Một hướng nghiên cứu khác là phát triển các mô hình học tập chuyển giao (transfer learning) giữa các lĩnh vực y tế liên quan như chẩn đoán ung thư miệng, phân tích dấu hiệu sinh học từ nước bọt hoặc máu, và để từ đó đa dạng hóa đối tượng tư vấn của chatbot. Việc kết hợp đa dạng nguồn dữ liệu sinh học sẽ giúp chatbot phát triển tầm nhìn toàn diện hơn về sức khỏe răng miệng nói riêng và sức khỏe toàn thân nói chung.

Song song đó, việc phát triển các mô hình đa phương thức (multi-modal models) là một hướng đi đầy hứa hẹn. Các mô hình này có khả năng xử lý đồng thời nhiều dạng dữ liệu đầu vào như văn bản, hình ảnh, âm thanh và video. Trong lĩnh vực nha khoa, chatbot có thể nhận đầu vào là một đoạn hội thoại kết hợp với ảnh răng hoặc đoạn video ghi lại tình trạng bệnh nhân, từ đó đưa ra đánh giá tổng hợp chính xác và sinh động hơn. Khả năng này đặc biệt phù hợp với môi trường khám chữa bệnh từ xa (telemedicine), nơi bác sĩ cần hỗ trợ ra quyết định nhanh chóng qua các kênh dữ liệu khác nhau. Cuối cùng, một định hướng mở rộng quan trọng là tích hợp các công nghệ AI tương tác âm thanh để phát triển chatbot nâng cao khả năng giao tiếp tự nhiên với người dùng. Việc ứng dụng xử lý tiếng nói, nhận dạng cảm xúc từ âm điệu và nhịp điệu câu nói sẽ giúp chatbot trở nên thân thiện, cá nhân hóa hơn và tăng tốc độ lan tỏa sử dụng.

Như vậy, việc tiếp tục nghiên cứu, tích hợp và mở rộng hệ thống chatbot không chỉ mang lại giá trị kỹ thuật mà còn góp phần xây dựng một nền tảng tư vấn y tế thông minh, lâu dài và bền vững cho tương lai.

## KẾT LUẬN

Đề tài “Nghiên cứu và xây dựng mô hình chatbot tư vấn nha khoa sử dụng mạng kết hợp chuyên gia” đã triển khai toàn diện một quy trình xây dựng hệ thống hội thoại trí tuệ nhân tạo có khả năng xử lý ngôn ngữ tự nhiên chuyên sâu, đáp ứng yêu cầu tư vấn trong lĩnh vực nha khoa. Trên cơ sở tích hợp các tiến bộ về mô hình ngôn ngữ lớn (LLM), đặc biệt là kiến trúc DeepSeek kết hợp Mixture of Experts (MoE), đề án đã góp phần đưa ra một hướng tiếp cận khả thi trong việc phát triển các hệ thống chatbot chuyên ngành, hướng đến khả năng ứng dụng thực tiễn trong môi trường y tế kỹ thuật số.

Trong quá trình thực hiện, đề án đã xác lập rõ ràng lộ trình phát triển từ khâu thu thập dữ liệu đến huấn luyện và đánh giá mô hình. Bộ dữ liệu được xây dựng trên cơ sở tổng hợp từ các nguồn chuyên môn đáng tin cậy, bao gồm văn bản học thuật, hội thoại thực tế giữa bác sĩ và bệnh nhân, và các tài liệu hướng dẫn lâm sàng. Quá trình tiền xử lý được thực hiện chặt chẽ nhằm chuẩn hóa ngữ nghĩa, gán nhãn vai trò hội thoại, và cấu trúc hóa thành tập dữ liệu phù hợp cho bài toán học máy. Mô hình DeepSeek-R1 được chọn lựa và tinh chỉnh thông qua hai giai đoạn: huấn luyện có giám sát (SFT) để đảm bảo tuân thủ cấu trúc phản hồi chuyên môn, và Reinforcement Learning từ phản hồi con người (RLHF) để tăng tính tự nhiên và sát với mong đợi của người dùng.

Đóng góp lớn nhất của đề án là việc ứng dụng mạng kết hợp chuyên gia trong hệ thống chatbot y tế chuyên ngành có thể coi là một hướng đi còn ít được khai thác trong các mô hình hiện tại ở Việt Nam. Kiến trúc này cho phép phân tách nhiệm vụ phản hồi theo chuyên môn, đồng thời tối ưu tài nguyên tính toán mà vẫn giữ được chất lượng suy luận. Hệ thống chatbot không chỉ phản hồi chính xác về mặt nội dung mà còn thể hiện khả năng thích ứng ngữ cảnh, duy trì mạch hội thoại và đưa ra các phản hồi phù hợp với mức độ hiểu biết của từng nhóm người dùng khác nhau. Những đặc điểm này tạo nền tảng để chatbot có thể vận hành như một trợ lý tư vấn nha khoa ảo đáng tin cậy, hỗ trợ tốt cho các tác vụ sàng lọc, giải thích bệnh lý hoặc cung cấp thông tin chăm sóc răng miệng.

Dẫu đạt được những kết quả tích cực, hệ thống vẫn tồn tại một số hạn chế nhất định. Dữ liệu đầu vào, dù đã được xử lý kỹ, vẫn còn thiếu sự đa dạng trong các trường hợp đặc thù, ví dụ như các bệnh lý hiếm gặp, hội thoại chuyên sâu giữa bác sĩ chuyên khoa với nhau, hoặc tình huống khẩn cấp y khoa. Điều này dẫn đến việc mô hình đôi khi đưa ra

phản hồi ở mức khái quát, chưa đáp ứng đầy đủ yêu cầu chẩn đoán chính xác trong các trường hợp biên. Thêm vào đó, khả năng xử lý hội thoại nhiều lượt, duy trì ngữ cảnh dài và thích ứng với phong cách diễn đạt đa dạng của người dùng vẫn còn dư địa để cải thiện. Về mặt triển khai, mô hình hiện vẫn vận hành ở môi trường thử nghiệm, chưa được tích hợp vào các nền tảng quản lý y tế hoặc đánh giá theo tiêu chuẩn lâm sàng, điều này làm hạn chế khả năng kiểm chứng tính an toàn và hiệu quả ở quy mô lớn.

Hướng phát triển tiếp theo có thể được định hình theo ba trục chính. Thứ nhất là mở rộng quy mô và chiều sâu của bộ dữ liệu huấn luyện, đặc biệt bổ sung thêm các nguồn đa phương thức như hình ảnh X-quang, dữ liệu âm thanh, văn bản hội chẩn hoặc phản hồi từ hệ thống chăm sóc sức khỏe điện tử. Việc tích hợp dữ liệu từ các nguồn khác nhau sẽ giúp mô hình học được nhiều biểu hiện triệu chứng và phong cách hội thoại đa dạng hơn, từ đó nâng cao năng lực phản hồi chuyên biệt. Thứ hai, phát triển năng lực duy trì ngữ cảnh dài và học liên tục (continual learning) nhằm giúp mô hình tích lũy kiến thức từ các tương tác trước, qua đó cải thiện khả năng cá nhân hóa và nâng cao độ tin cậy trong các phiên tư vấn lặp lại. Thứ ba, triển khai thử nghiệm trên nền tảng y tế thực tế, chẳng hạn tích hợp vào hệ thống quản lý hồ sơ bệnh án, ứng dụng tư vấn từ xa hoặc các trợ lý số tại cơ sở nha khoa để nhằm đánh giá tính hiệu quả trong điều kiện vận hành thực tế, đồng thời kiểm tra độ chấp nhận từ phía người dùng chuyên môn và cộng đồng.

Tổng thể, đề án đã đạt được các mục tiêu nghiên cứu cốt lõi và chứng minh được tính khả thi trong việc phát triển một hệ thống chatbot chuyên ngành ứng dụng trí tuệ nhân tạo hiện đại. Đề tài không chỉ góp phần mở rộng ứng dụng LLM vào lĩnh vực y tế chuyên sâu mà còn tạo tiền đề cho các hệ thống hỗ trợ ra quyết định, góp phần giải quyết bài toán quá tải nhân lực y tế, đặc biệt trong bối cảnh chuyển đổi số đang ngày càng mạnh mẽ.



## TÀI LIỆU THAM KHẢO

- [1] Juniper Research, "Chatbots in Healthcare: Market Trends and Success Rates," Juniper Research, 2021.
- [2] VinBrain, "DrAid™: Trợ lý y tế ảo ứng dụng trí tuệ nhân tạo hỗ trợ chẩn đoán hình ảnh," VinBrain.
- [3] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," arXiv preprint arXiv:1701.06538, 2017.
- [4] A. Kay, "Predicting the Future," Xerox PARC internal memo, 1971. Quoted in Computerworld, 1982.
- [5] D. AI, "DeepSeek-R1: Expert Language Model with Reasoning and Tool-Use," arXiv preprint arXiv:2405.18880, May 2024.
- [6] J. Ouyang et al., "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, p. 27730–27744.
- [7] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Proceedings of the ACL-04 Workshop on Text Summarization Branches Out, p. 74–81, 2004.
- [8] S. R. T. W. a. W.-J. Z. K. Papineni, "BLEU: a Method for Automatic Evaluation of Machine Translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, p. 311–318.
- [9] M. AI, "The Llama 3 Model Family: A new generation of open-source large language models," Meta AI Blog, Apr. 2024.
- [10] A. Vaswani et, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), p. 5998–6008, 2017.
- [11] D. AI, "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism," arXiv preprint arXiv:2312.14385, Dec. 2023.
- [12] D. AI, "DeepSeek-V2: Scaling Mixture of Experts with Multi-head Latent Attention," arXiv preprint arXiv:2405.04434, May 2024.

- [13] A. Liu, B. Feng, B. Xue, B. Wang, B. Huang, S. Zhou, et al., "“DeepSeek-V3 Technical Report,” arXiv:2412.19437v2 [cs.CL]," 18-Feb-2025.
- [14] L. Ouyang et al, "“Training language models to follow instructions with human feedback,” arXiv preprint arXiv:2203.02155," Mar. 2022.
- [15] J. D. J. K. M. M. M. R. R. D. L. M. F. a. M. L. J. S. H. Ji, "ORPO: Monolithic Preference Optimization without Reference Model," arXiv preprint arXiv:2403.07691, Mar. 2024.
- [16] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "“Deep reinforcement learning from human preferences,” arXiv:1706.03741," Jun. 2017.
- [17] S. Hawking, "“The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge”".
- [18] Google, "Gemini 1.5 Flash Model Card," Google AI, 2024.
- [19] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv preprint arXiv:2201.11903, 2022.
- [20] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, p. 1877–1901.
- [21] W. E. Deming (attributed), In God we trust, all others must bring data, Goodreads Quotes.
- [22] T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023.
- [23] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
- [24] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, p. 38–45.
- [25] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine

Translation and/or Summarization, 2005, p. 65–72.

- [26] V. K. F. W. K. Q. W. a. Y. A. T. Zhang, "BERTScore: Evaluating Text Generation with BERT," arXiv preprint arXiv:1904.09675, 2019.
- [27] S. Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," arXiv preprint arXiv:2305.10601, May 2023.