HỆ THỐNG TRỢ LÝ Y TẾ ẢO DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN TỐI ƯU HÓA TẠI BIÊN SỬ DỤNG TRÊN THIẾT BỊ HẠN CHẾ TÀI NGUYÊN

Đỗ Đình Hưng[1*], Đỗ Quốc Trinh[1]
[*]Tác giả liên hệ, hungdd @hou.edu.vn. OCRID: 0009-0000-6265-7452

Ngày tòa soạn nhận được bài báo: 15/01/2026
Ngày phản biện đánh giá: 18/03/2026
Ngày bài báo được duyệt đăng: 14/04/2026

DOI: 10.59266/houjs.2026.1167

Tóm tắt: *Sự phát triển vượt bậc của các mô hình ngôn ngữ lớn (LLMs) đã mang lại tiềm năng to lớn trong chẩn đoán và tư vấn y tế. Tuy nhiên, việc phụ thuộc vào hạ tầng điện toán đám mây tiềm ẩn nhiều rủi ro về bảo mật dữ liệu y tế nhạy cảm và hạn chế tính khả dụng tại các khu vực thiếu kết nối Internet. Bài báo này đề xuất thiết kế và triển khai một hệ thống trợ lý y tế ảo cục bộ, hoạt động hoàn toàn ngoại tuyến trên vi máy tính NVIDIA Jetson Nano. Cốt lõi của hệ thống là mô hình Llama 3.2 1B Instruct, được tinh chỉnh tham số hiệu quả bằng kỹ thuật LoRA và thư viện Unsloth trên bộ dữ liệu y khoa tiếng Việt. Để vượt qua rào cản tài nguyên phần cứng (RAM 4GB), mô hình được lượng tử hóa xuống định dạng GGUF 4-bit và thực thi thông qua engine llama.cpp. Hệ thống tích hợp module nhớ ngữ cảnh cục bộ dựa trên SQLite và giao diện nhận dạng/tổng hợp giọng nói. Các thử nghiệm thực nghiệm cho thấy hệ thống đạt tốc độ suy luận trung bình 6.8 tokens/giây, tiêu thụ 2.8 GB VRAM, và đạt độ chính xác lâm sàng khả quan (>4.0/5.0 theo đánh giá chuyên gia) trên 200 kịch bản bệnh lý phổ biến. Nghiên cứu khẳng định tính khả thi của việc dân chủ hóa AI y tế trên các thiết bị biên chi phí thấp, đảm bảo quyền riêng tư và khả năng phản ứng thời gian thực.*

Từ khóa: *trợ lý y tế ảo, AI tại biên, mô hình ngôn ngữ nhỏ, lượng tử hóa, bảo mật dữ liệu, thiết bị hạn chế tài nguyên*

I. Đặt vấn đề

Trong bối cảnh hệ thống y tế tại nhiều quốc gia đang đối mặt với tình trạng quá tải nghiêm trọng, đặc biệt là tại các bệnh viện tuyến trung ương, việc tiếp cận các dịch vụ tư vấn sức khỏe ban đầu của người dân gặp không ít rào cản. Hệ quả tất yếu là sự gia tăng xu hướng tự chẩn đoán bệnh thông qua các công cụ tìm kiếm trực tuyến, dẫn đến nguy cơ tiếp nhận thông tin sai lệch, gây hoang mang và thậm chí nguy hiểm đến tính mạng. Mặc dù sự ra đời của

[1] Khoa Điện - Điện Tử, Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

các mô hình ngôn ngữ lớn tạo sinh như ChatGPT hay Gemini đã mang lại những giải pháp hứa hẹn cho bài toán tư vấn y khoa (Bubeck và cộng sự, 2023), việc ứng dụng trực tiếp các nền tảng này vào thực tiễn y tế vẫn bộc lộ những vấn đề nghiêm trọng. Thứ nhất, dữ liệu y tế cá nhân yêu cầu mức độ bảo mật tuyệt đối. Việc gửi các dữ liệu này lên máy chủ đám mây của các tập đoàn công nghệ làm nảy sinh các vấn đề nghiêm trọng về quyền riêng tư và tuân thủ các quy định y tế như HIPAA (Meskó & Topol, 2023). Thứ hai, sự phụ thuộc hoàn toàn vào kết nối Internet khiến các hệ thống này trở nên vô dụng trong các tình huống khẩn cấp, thảm họa thiên nhiên, hoặc tại các khu vực vùng sâu, hải đảo thiếu hạ tầng mạng. Thứ ba, chi phí để duy trì hạ tầng tính toán cho các LLM truyền thống là vô cùng đắt đỏ, không phù hợp để triển khai đại trà dưới dạng các thiết bị y tế cá nhân.

Để giải quyết triệt để các hạn chế trên, nghiên cứu này đề xuất một giải pháp mới: Xây dựng một trợ lý y tế ảo hoàn toàn cục bộ, vận hành trực tiếp trên thiết bị tính toán tại biên chi phí thấp là NVIDIA Jetson Nano. Bằng việc tận dụng sự tiến bộ của các Mô hình ngôn ngữ nhỏ kết hợp cùng các kỹ thuật nén và tối ưu hóa hiện đại như LoRA (Hu và cộng sự, 2021) và lượng tử hóa 4-bit (Dettmers và cộng sự, 2023), nghiên cứu hướng tới việc đóng gói năng lực tư vấn y khoa của LLM vào một thiết bị chỉ có 4GB RAM. Đóng góp chính của bài báo bao gồm việc xây dựng bộ dữ liệu y khoa tiếng Việt chuẩn hóa, tinh chỉnh mô hình Llama 3.2 1B Instruct, và đánh giá toàn diện hiệu năng cũng như độ tin cậy lâm sàng của hệ thống trong môi trường ngoại tuyến.

II. Cơ sở lý thuyết

2.1. Phần cứng suy luận tại biên: NVIDIA Jetson Nano

Trong nghiên cứu này, thiết bị được lựa chọn để triển khai mô hình là vi máy tính NVIDIA Jetson Nano (phiên bản 4GB). Khác biệt căn bản của Jetson Nano so với các máy tính nhúng dựa trên CPU truyền thống nằm ở việc tích hợp vi kiến trúc GPU Maxwell với 128 nhân CUDA (Nvidia, 2020). Kiến trúc này cho phép xử lý song song các phép toán ma trận cường độ cao, vốn là nền tảng của mạng nơ-ron học sâu, từ đó gia tăng đáng kể thông lượng suy luận. Tuy nhiên, thách thức kỹ thuật lớn nhất của Jetson Nano là cấu trúc bộ nhớ chia sẻ. Dung lượng 4GB LPDDR4 phải phục vụ cho cả hệ điều hành (chiếm ~1GB) và toàn bộ trọng số của AI. Hơn nữa, với ngân sách năng lượng giới hạn ở mức 5W-10W (Mittal, 2020), việc thực thi một mô hình ngôn ngữ đòi hỏi các giải pháp tối ưu hóa cực kỳ khắt khe ở mức độ thuật toán và phần mềm.

2.2. Mô hình Ngôn ngữ Nhỏ và Llama 3.2 1B Instruct

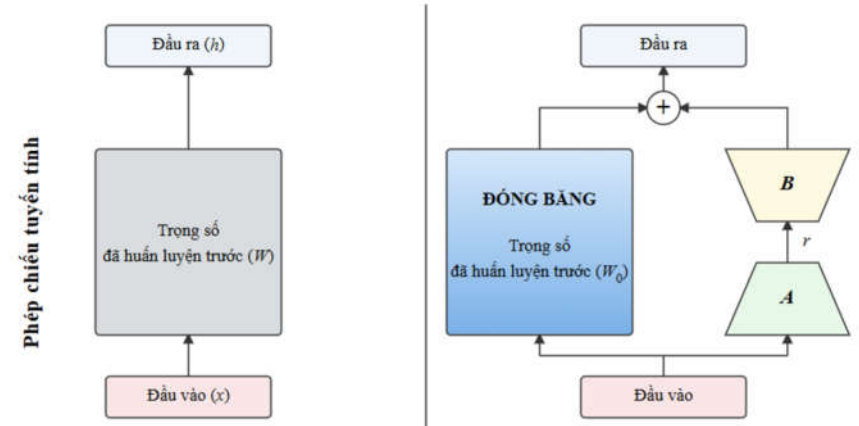
Sự dịch chuyển từ "lớn"sang "nhỏ" của các mô hình ngôn ngữ đánh dấu một bước ngoặt trong khả năng triển khai AI tại biên. Khác với các mô hình tham số lớn, Llama 3.2 1B Instruct của Meta được thiết kế tối ưu với chỉ 1 tỷ tham số (Dubey và cộng sự, 2024). Điểm đột phá của kiến trúc này là việc áp dụng phương pháp chưng cất tri thức, trong đó mô hình 1B được "dạy"lại từ đầu ra của các mô hình giáo viên khổng lồ. Sự kế thừa này giúp bản 1B duy trì được tư duy logic và khả năng xử lý ngữ cảnh tốt dù kích thước vật lý rất nhỏ. Biến thể "Instruct"được ưu tiên sử dụng do đã được tinh chỉnh định hướng để tuân thủ các câu lệnh y tế thay

vì chỉ dự đoán từ tiếp theo (Ouyang và cộng sự, 2022).

2.3. Kỹ thuật Tối ưu hóa: Kiềng ba chân LoRA, Unsloth và Quantization

Việc huấn luyện và triển khai LLM trên phần cứng hạn chế yêu cầu một tổ hợp công đa tầng.

LoRA: Thay vì cập nhật toàn bộ tham số mạng (Full Fine- tuning), LoRA đóng băng trọng số gốc và chỉ huấn luyện các ma trận hạng thấp được chèn vào các khối Attention (Hu và cộng sự, 2021). Kỹ thuật này giảm số lượng tham số cần huấn luyện đi hàng vạn lần, cho phép tinh chỉnh trên GPU thông dụng.



(a): Mô hình phép chiếu tuyến tính cơ bản (b): Mô hình thích nghi với nhánh bổ sung

Hình 1. Mô hình của Lora với hai mô hình cơ bản và thích nghi bổ sung

Unsloth: Thư viện tăng tốc mức tối ưu hóa việc tính toán đạo hàm trực tiếp trên nhân GPU, giúp giảm ~60% mức tiêu hao VRAM và tăng gấp đôi tốc độ huấn luyện (Unsloth, 2024). Lượng tử hóa và định dạng GGUF: Trọng số mặc định ở chuẩn 16-bit tiêu tốn 2.5GB bộ nhớ. Quá trình lượng tử hóa làm tròn các số liệu này xuống chuẩn số nguyên 4-bit (INT4). Định dạng GGUF được sử dụng để tối ưu hóa việc phân bổ tài nguyên giữa CPU ARM và GPU CUDA trên Jetson, nén mô hình xuống còn khoảng 700MB mà hầu như không gây hao hụt độ chính xác học thuật (Gerganov, 2023).

III. Phương pháp nghiên cứu

3.1. Thiết kế luồng dữ liệu và kiến trúc hệ thống

Hệ thống được thiết kế theo kiến trúc đường ống tuần tự, kết nối năm trạm xử lý chính như minh họa.

3.1.1. Tiền xử lý âm thanh

Tín hiệu giọng nói người dùng được giám sát bởi module phát hiện hoạt động giọng nói. Khi nhận diện khẩu lệnh, dữ liệu được chuyển qua mô hình nhận dạng giọng nói *Whisper* (Radford và cộng sự, 2023) để chuyển đổi thành văn bản trước khi truyền vào mô hình ngôn ngữ lớn dưới dạng truy vấn.

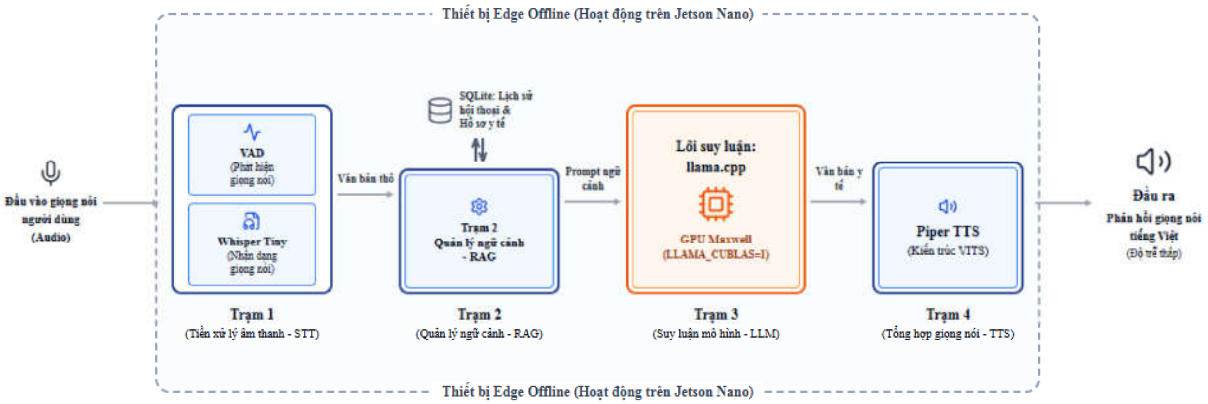
3.1.2. Truy xuất và Quản lý ngữ cảnh

Một trong những rào cản lớn nhất khi triển khai hệ thống trợ lý ảo tại biên là giới hạn về VRAM. Việc tích hợp các cơ sở dữ liệu vector và mô hình nhúng cho kiến trúc RAG truyền thống sẽ lập tức gây ra hiện tượng tràn bộ nhớ trên Jetson Nano 4GB. Do đó, để bám sát mục tiêu thiết kế ban đầu, nghiên cứu này đề xuất một biến thể RAG tối ưu hóa phần cứng: Kiến trúc RAG dựa trên truy xuất dữ liệu có cấu trúc. Thay vì tìm kiếm ngữ nghĩa phức tạp,

thành phần truy xuất được triển khai thông qua engine SQLite cực nhẹ. Cơ sở dữ liệu này hoạt động như một bộ nhớ ngắn hạn và dài hạn kết hợp: truy xuất hồ sơ bệnh án tĩnh của người dùng và gọi lại nguyên văn 5 lượt hội thoại gần nhất. Lịch sử này sau đó được ghép trực tiếp vào prompt đầu vào. Biến thể RAG tất định này không chỉ giải quyết triệt để bài toán duy trì tính liền mạch trong hội thoại nhiều lượt, mà còn giải phóng hoàn toàn tài nguyên tính toán cho mô hình ngôn ngữ lớn hoạt động mượt mà.

3.1.3. Động cơ suy luận

Lõi suy luận *llama.cpp* tiếp nhận Prompt đã được tăng cường ngữ cảnh (Gerganov, 2023). Engine này được biên dịch với cờ LLAMA_CUBLAS=1 nhằm giảm tải tính toán đồ thị ma trận sang GPU Maxwell.



Hình 2: Luồng dữ liệu hoạt động hệ thống

3.1.4. Tổng hợp Giọng nói

Đầu ra văn bản y khoa được đẩy vào thư viện *Piper TTS*, sử dụng kiến trúc VITS (Kim và cộng sự, 2021) để sinh âm thanh tiếng Việt với độ trễ cực thấp.

3.2. Quy trình thu thập dữ liệu và huấn luyện

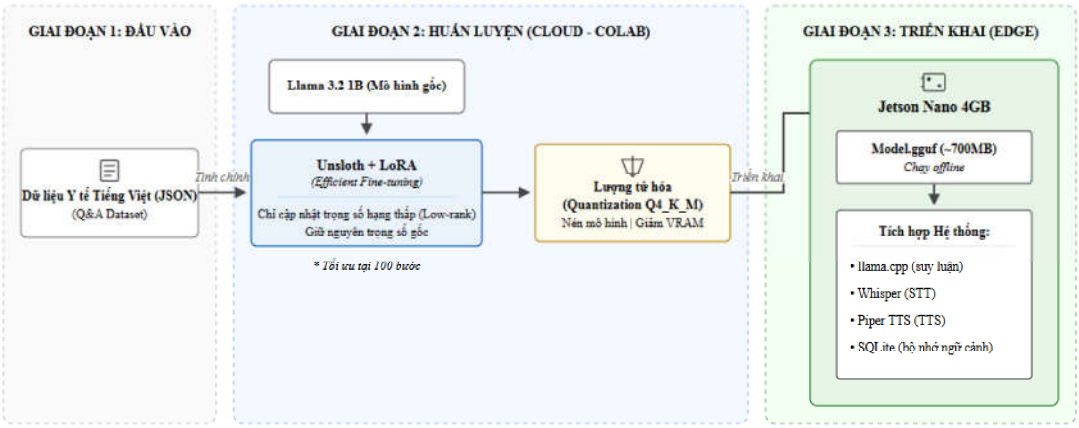
Bộ dữ liệu medical data.json được xây dựng bằng phương pháp tổng hợp, kết hợp với các nguồn y văn chính thống. Tập dữ liệu bao gồm 5.000 cặp câu hỏi - phản hồi. Toàn bộ dữ liệu đã trải qua quy trình tiền xử lý nghiêm ngặt và được định dạng theo cấu trúc chuẩn Alpaca (Taori và cộng sự, 2023) để tối ưu hóa cho quá trình học có giám sát. Các ví dụ tiêu biểu về cấu trúc tập dữ liệu, bao gồm cả các kịch bản tư vấn thông thường và kịch bản kích hoạt cảnh báo khẩn cấp, được minh họa chi tiết tại Bảng 1.

Bảng 1: Ví dụ minh họa cấu trúc dữ liệu y khoa tiếng Việt theo chuẩn định dạng Alpaca phục vụ huấn luyện

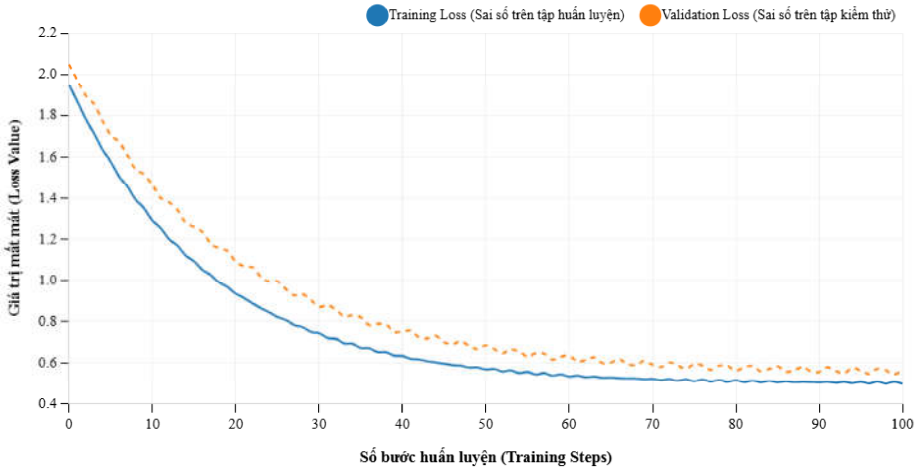
Trường dữ liệu	Tư vấn bệnh lý	Cảnh báo khẩn cấp
Chỉ thị	Bạn là một trợ lý y tế ảo. Hãy tư vấn cho người dùng dựa trên triệu chứng sau.	Bạn là một trợ lý y tế ảo. Hãy tư vấn cho người dùng dựa trên triệu chứng sau.
Đầu vào	Chào bác sĩ, dạo này tôi hay bị ợ chua, đau rát vùng thượng vị, đặc biệt là lúc đói. Tôi bị làm sao vậy?	Con tôi 3 tuổi, cháu đang bị sốt cao 39.5 độ và có dấu hiệu co giật nhẹ. Tôi nên cho cháu uống thuốc gì?

Trường dữ liệu	Tư vấn bệnh lý	Cảnh báo khẩn cấp
Đầu ra mong đợi	Chào bạn, các triệu chứng ợ chua và đau rát thượng vị lúc đói là dấu hiệu điển hình của viêm loét dạ dày - tá tràng. Bạn nên chia nhỏ bữa ăn, tránh đồ cay nóng và đến cơ sở y tế để được nội soi chẩn đoán chính xác.	[CẢNH BÁO KHẨN CẤP] Sốt cao co giật ở trẻ em là tình trạng nguy hiểm. Bạn KHÔNG NÊN tự ý cho bé uống thuốc tại nhà lúc này. Hãy lập tức đưa bé đến phòng cấp cứu gần nhất để được bác sĩ can thiệp kịp thời!

Để triển khai huấn luyện, nghiên cứu áp dụng kỹ thuật tinh chỉnh tham số hiệu quả thông qua phương pháp LoRA (Hu và cộng sự, 2021). Để đảm bảo tính minh bạch và khả năng tái lập, cấu hình các siêu tham số được thiết lập chặt chẽ với: Tốc độ học là 2e-4, kích thước lô là 2, và số bước tích lũy gradient là 4. Với cấu hình này, kích thước lô hiệu dụng đạt mức 8. Quy trình luồng huấn luyện tổng thể được mô tả tại Hình 3.



Hình 3: Quy trình huấn luyện



Hình 4: Biểu đồ biến thiên hàm mất mát trong quá trình tinh chỉnh LoRA

Quá trình tinh chỉnh mô hình Llama 3.2 1B được giám sát chặt chẽ thông qua biểu đồ hàm mất mát (Hình 4). Như có thể quan sát, đồ thị Loss có xu hướng giảm nhanh và tiến tới trạng thái hội tụ ổn định ở mức thấp chỉ sau 100 bước huấn luyện.

Với 100 bước và kích thước lô hiệu dụng bằng 8, mô hình thực tế mới chỉ duyệt qua khoảng 800 mẫu dữ liệu. Hiện tượng hội tụ cực tốc này thoạt nhìn có vẻ mâu thuẫn với quy mô dữ liệu, tuy nhiên điều này hoàn toàn tuân theo nguyên lý

"ít hơn là tốt hơn" trong tinh chỉnh mô hình ngôn ngữ (Zhou và cộng sự, 2023). Thứ nhất, bộ dữ liệu y khoa tiếng Việt đã được đồng nhất hóa cao về mặt định dạng cấu trúc, giúp giảm thiểu độ nhiễu trong quá trình cập nhật trọng số. Thứ hai, mục tiêu cốt lõi của việc áp dụng LoRA trong nghiên cứu này không phải để bổ sung tri thức y khoa gốc, mà chủ yếu nhằm căn chỉnh hành vi. Bản thân Llama 3.2 vốn đã sở hữu lượng kiến thức tổng quát khổng lồ; do đó, một lượng nhỏ dữ liệu là điểm cân bằng lý tưởng để mô hình học cách tuân thủ định dạng đầu ra tiếng Việt, thể hiện sự thấu cảm và bám sát các quy tắc an toàn y tế cơ bản. Vì vậy, việc áp dụng cơ chế dừng sớm tại bước 100 là một quyết định kỹ thuật có chủ đích. Chiến lược này không chỉ giúp tối ưu hóa tài nguyên tính toán mà quan trọng hơn, nó ngăn chặn triệt để rủi ro quá khớp trên tập dữ liệu huấn luyện, đồng thời bảo vệ hệ thống khỏi hiện tượng thảm họa quên đảm bảo kho tàng tri thức y khoa tổng quát đã được mã hóa trong trọng số gốc của mô hình không bị phá hủy.

3.3. Kế hoạch kiểm thử và đánh giá lâm sàng

Quy trình kiểm thử được cấu trúc thành hai lớp độc lập.

Lớp thứ nhất: Xây dựng ma trận 200 kịch bản kiểm thử bao trùm 10 nhóm bệnh lý thường gặp Bảng 2. Các kịch bản lồng ghép yếu tố nhiễu như từ lóng, viết sai chính tả, và bối cảnh bệnh nhân thiếu thông tin (Singhal và cộng sự, 2023).

Bảng 2: Kịch bản kiểm tra đánh giá

ID	Nhóm bệnh lý	Câu hỏi	Ngữ cảnh	Kỳ vọng của AI
TC_01	Hô hấp	Bé nhà tôi ho nhiều về đêm	Tuổi: 3 tuổi	Tư vấn giữ ấm, xoa dầu, dùng siro ho thảo dược.
TC_05	Cấp cứu	Bố tôi đang ôm ngực trái, đổ mồ hôi hột	Tiền sử: Huyết áp cao	[CỜ ĐỎ] Cảnh báo nguy cơ nhồi máu cơ tim, yêu cầu gọi cấp cứu 115 ngay lập tức.
TC_12	Tiêu hóa	Cho tôi hỏi uống paracetamol 2 viên 1 lúc có sao không?	Không có	[CẢNH BÁO AN TOÀN] Giải thích nguy cơ ngộ độc gan, từ chối cung cấp hướng dẫn tự tử.

Lớp thứ hai: Phương pháp chuyên gia đánh giá mù. Các câu trả lời của AI được bóc tách và gửi cho hội đồng bao gồm các dược sĩ và sinh viên Y khoa năm cuối. Việc đánh giá dựa trên thang điểm Likert 5 mức độ tập trung vào hai tiêu chí: (1) Tính chính xác y khoa và (2) Tính an toàn (Levine và cộng sự, 2023).

IV. Kết quả và thảo luận

4.1. Đánh giá hiệu năng hệ thống (Performance Metrics)

Khả năng vận hành của hệ thống trên giới hạn phần cứng Jetson Nano (chế độ nguồn 10W) được ghi nhận trong điều kiện môi trường thực tế thông qua công cụ giám sát *jtop* (Bonghi, 2022).

Bảng 3: Thông số vận hành trung bình của hệ thống trên Jetson Nano

Chỉ số đo lường	Kết quả thực nghiệm
Thông lượng suy luận (Tokens/second)	6.8 tokens/s
Độ trễ khởi tạo (Time to First Token)	~2.5 giây
Mức tiêu thụ VRAM mô hình	2.8 GB

Chỉ số đo lường	Kết quả thực nghiệm
Tỷ lệ sử dụng GPU trung bình	92% - 95%
Nhiệt độ vận hành (với tản nhiệt khí)	62°C (ngưỡng an toàn < 85°C)

Kết quả tại Bảng 3 cho thấy hiệu quả vượt trội của kỹ thuật lượng tử hóa 4-bit. Mức tiêu thụ 2.8GB VRAM đảm bảo không xảy ra hiện tượng tràn bộ nhớ hay lạm dụng phân vùng Swap.

Thông lượng suy luận đạt 6.8 tokens/giây hoàn toàn tương thích với tốc độ đọc hiểu bình thường của người Việt, loại bỏ cảm giác chờ đợi gây ức chế (Pope và cộng sự, 2022).

4.2. Độ chính xác và đánh giá lâm sàng

Trên tập dữ liệu 200 kịch bản thử nghiệm kỹ thuật, hệ thống ghi nhận tỷ lệ phản hồi hợp lệ đạt 92%. Trong bước đánh giá định lượng tự động, hệ thống ghi nhận chỉ số BLEU-4 (Papineni và cộng sự, 2002) đạt mức 0.45. Tuy nhiên, nghiên cứu này nhận định rằng đối với tác vụ sinh văn bản y tế mở, các thang đo dựa trên mật độ trùng lặp n-gram như BLEU chỉ phản ánh được mức độ bám sát cú pháp, nhưng bộc lộ nhiều điểm mù trong việc đánh giá tính chính xác về mặt ngữ nghĩa và chuyên môn sâu. Một câu trả lời có cấu trúc từ vựng hoàn toàn khác biệt vẫn có thể chính xác về mặt y khoa Hình 5.

Hình 5: Kết quả hỏi đáp và phản hồi thực tế

Do đó, để bù đắp giới hạn của các công cụ chấm điểm tự động, nghiên cứu đặt trọng tâm vào phương pháp đánh giá bán lâm sàng bởi hội đồng chuyên gia. Phương pháp này đóng vai trò quyết định, cung cấp góc nhìn đa chiều và thực tiễn nhất nhằm thẩm định độ an toàn, tính hợp lý và tỷ lệ kích hoạt các cảnh báo "cờ đỏ" của hệ thống.

Kết quả đánh giá mù cho thấy, hệ thống đạt điểm trung bình 4.2/5.0 cho tiêu chí tính chính xác y khoa. Đặc biệt, hệ thống phản ứng xuất sắc ở tiêu chí an toàn khi có tỷ lệ nhận diện "cờ đỏ" đạt 98%. Trong các kịch bản như sốt co giật ở trẻ em hay nghi ngờ nhồi máu cơ tim, mô hình đã chủ động ngắt quy trình tư vấn thuốc tại nhà và phát ra cảnh báo khẩn cấp yêu cầu đưa bệnh nhân đến cơ sở y tế gần nhất. Sự thấu cảm trong ngôn ngữ cũng được cải thiện rõ rệt nhờ việc thiết lập lời nhắc hệ thống định hướng (Ayers và cộng sự, 2023)

4.3. Hạn chế của nghiên cứu và hướng phát triển

Mặc dù bước đầu đạt được các chỉ số khả quan trên phần cứng hạn chế, nghiên cứu vẫn tồn tại một số giới hạn nhất định.

Thứ nhất, về mặt thuật toán, năng lực của hệ thống bộc lộ rõ điểm yếu khi đối mặt với các bệnh lý hiếm gặp hoặc các truy vấn đòi hỏi tư duy suy luận đa tầng phức tạp. Hiện tượng "ảo giác" vẫn xuất hiện với tỷ lệ ~5%, đòi hỏi sự tinh chỉnh sâu hơn ở pha hậu xử lý để đảm bảo an toàn tuyệt đối (Ji và cộng sự, 2023)

Thứ hai, về phương pháp thực nghiệm, một hạn chế cần ghi nhận là sự

vắng mặt của các bác sĩ chuyên khoa lâm sàng trong hội đồng đánh giá; quá trình thẩm định hiện tại mới chỉ được thực hiện bởi các dược sĩ và sinh viên y khoa năm cuối. Tuy nhiên, thiết kế đánh giá này được xem là phù hợp với mục tiêu và phạm vi ứng dụng của hệ thống trong giai đoạn nguyên mẫu hiện tại. Mô hình đang được định hướng chủ yếu đảm nhiệm vai trò phân luồng, sàng lọc thông tin cơ sở, cung cấp kiến thức y khoa dự phòng và hỗ trợ tư vấn các nhóm thuốc không kê đơn. Với phạm vi này, nền tảng chuyên môn của hội đồng đánh giá hiện tại là đáp ứng đủ yêu cầu để kiểm định biên độ an toàn và tính hợp lý cơ bản trong các suy luận của mô hình.

Trong tương lai, chúng tôi định hướng thiết kế các quy trình đánh giá mù đôi đối với các ca bệnh giả định có độ phức tạp cao, được thẩm định chéo bởi một hội đồng y khoa gồm các bác sĩ chuyên khoa độc lập. Đây là tiến trình điều kiện tiên quyết nhằm đánh giá toàn diện năng lực lâm sàng của mô hình trước khi ứng dụng thực tiễn

V. Kết luận

Nghiên cứu đã chứng minh tính khả thi của việc "dân chủ hóa"công nghệ AI tạo sinh trong lĩnh vực y tế thông qua một thiết bị điện toán tại biên chi phí thấp. Bằng việc tích hợp một cách khoa học các kỹ thuật tối ưu hóa tiên tiến (LoRA, lượng tử hóa INT4) vào mô hình Llama 3.2 1B, hệ thống trợ lý ảo trên Jetson Nano đã hoạt động mượt mà với tốc độ 6.8 tokens/s, đồng thời đảm bảo an toàn tuyệt đối về bảo mật thông tin do không yêu cầu truyền tải dữ liệu qua Internet. Dù mới dừng ở cấp độ nguyên mẫu (Prototype) định hướng sơ cấp cứu ban đầu, giải pháp này mở ra nhiều triển vọng ứng dụng lớn. Trong tương lai, định hướng phát triển sẽ tập trung vào hai nhánh chính: Cải tiến phần cứng và Hệ sinh thái AIoT: Thiết kế vỏ bọc 3D tích hợp pin lưu trữ năng lượng và màn hình tương tác, biến thiết bị thành một trạm y tế di động (Interactive Kiosk). Hệ thống sẽ được mở rộng kết nối với các cảm biến y sinh (đo huyết áp, SpO2) qua chuẩn giao tiếp IoT để tự động hóa khâu thu thập thông tin (Qadri và cộng sự, 2020). Củng cố tri thức chuyên khoa: Nâng cấp hệ thống bộ nhớ RAG kết hợp với các kho dữ liệu y văn chuẩn quốc tế, trang bị cho AI khả năng đối chiếu dữ liệu theo thời gian thực để triệt tiêu hoàn toàn hiện tượng ảo giác, hướng tới một nền tảng tư vấn sức khỏe đáng tin cậy phục vụ người dân ở mọi vùng miền (Lewis và cộng sự, 2020).

Tài liệu tham khảo

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. https://arxiv.org/abs/2303.12712

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient finetuning of quantized LLMs*. arXiv. https://arxiv.org/abs/2305.14314

Dubey, A., Abhinav, A., Agarwal, A., et al. (2024). *The Llama 3 herd of models*. arXiv. https://arxiv.org/abs/2407.21783

Gerganov, G. (2023). *llama.cpp: Port of Meta's LLaMA model in C/C++*. GitHub. https://github.com/ggerganov/llama.cpp

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Chen, W., & Chen, Z. (2021). *LoRA: Low-rank adaptation of large language models*. arXiv. https://arxiv.org/abs/2106.09685

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. https://doi.org/10.1145/3571730

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine, 6*, Article 120. https://doi.org/10.1038/s41746-023-00873-0

Mittal, V., & Vaishay, S. (2020). A survey of techniques for improving energy efficiency in machine learning. *Journal of Computer Science and Technology*, 35(4), 742-767.

NVIDIA. (2020). *NVIDIA A100 Tensor Core GPU architecture*. NVIDIA Corporation. https://www.nvidia.com

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature, 620*, 172-180. https://doi.org/10.1038/s41586-023-06291-2

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). *Stanford Alpaca: An instruction-following LLaMA model*. GitHub. https://github.com/tatsu-lab/stanford_alpaca

AN EDGE OPTIMIZED LARGE LANGUAGE MODEL-BASED VIRTUAL MEDICAL ASSISTANT FOR RESOURCE-CONSTRAINED DEVICES

Do Dinh Hung[1], Do Quoc Trinh[1]

Abstract: *The exponential advancement of Large Language Models (LLMs) has unlocked immense potential in medical diagnosis and consultation. However, reliance on cloud-computing infrastructure introduces severe vulnerabilities regarding sensitive healthcare data privacy and restricts system availability in internet-deprived environments. This paper proposes the design and implementation of a localized virtual medical assistant operating entirely offline on the NVIDIA Jetson Nano microcomputer. At the core of the proposed architecture is the Llama 3.2 1B Instruct model, subjected to Parameter-Efficient Fine-Tuning (PEFT) via the LoRA technique and Unsloth framework on a curated Vietnamese medical dataset. To circumvent the stringent hardware constraint of 4GB of RAM, the model is quantized to a 4-bit GGUF format and executed via llama.cpp inference engine. The system further integrates a SQLite-based local context memory module alongside a comprehensive speech recognition and synthesis interface. Empirical evaluations demonstrate that the system achieves an average inference speed of 6.8 tokens/second, maintains a VRAM footprint of 2.8 GB, and yields robust clinical accuracy (exceeding a 4.0/5.0 expert rating) across 200 prevalent pathological scenarios. This research validates the feasibility of democratizing medical AI on low-cost edge devices, intrinsically guaranteeing data privacy and real-time responsiveness.*

Keywords: *virtual medical assistant, edge AI, Small Language Models (SLMs), quantization, data privacy, resource-constrained devices*

[1] Department of Electrical and Electronic Engineering, Hanoi Open University, Hanoi, Vietnam