# Tổng quan mô hình ngôn ngữ lớn và kỹ thuật tăng cường truy xuất trong chẩn đoán bệnh da liễu

Bùi Thị Thanh Tú[1]*, Phạm Thị Anh Thư[2]
[1, 2]Trường Đại học Ngoại ngữ - Tin học Thành phố Hồ Chí Minh, Việt Nam
*tubtt@huflit.edu.vn

Tóm tắt

Nghiên cứu này tiến hành tổng quan hệ thống các công trình từ tháng 1/2023 đến tháng 5/2025 về ứng dụng mô hình ngôn ngữ lớn và kỹ thuật tăng cường truy xuất trong chẩn đoán bệnh da liễu. Từ hơn 40 tài liệu được tìm kiếm trên các cơ sở dữ liệu học thuật, 10 nghiên cứu tiêu biểu đã được phân tích chuyên sâu theo các tiêu chí về mô hình, dữ liệu, hiệu suất và khả năng ứng dụng lâm sàng. Kết quả cho thấy các mô hình ngôn ngữ lớn, đặc biệt khi kết hợp với kỹ thuật tăng cường truy xuất và dữ liệu đa phương thức, có tiềm năng nâng cao hiệu quả chẩn đoán và hỗ trợ ra quyết định lâm sàng, dù vẫn tồn tại các thách thức như hạn chế xử lý hình ảnh y khoa, nguy cơ "hallucination" và vấn đề bảo mật dữ liệu. Nghiên cứu đề xuất hướng phát triển bao gồm tối ưu hóa mô hình ngôn ngữ lớn chuyên biệt cho da liễu, tăng cường khả năng thị giác y khoa, chuẩn hóa bộ dữ liệu, và triển khai thử nghiệm lâm sàng có kiểm soát, nhằm tiến tới xây dựng hệ thống hỗ trợ chẩn đoán thông minh và tin cậy.

Nhận 14/10/2025
Được duyệt 14/11/2025
Công bố 30/11/2025

Từ khóa
Mô hình ngôn ngữ lớn; Kỹ thuật tăng cường truy xuất; Chẩn đoán bệnh da liễu; Suy luận đa phương thức; Hỗ trợ ra quyết định lâm sàng.



## 1. Đặt vấn đề

Trong những năm gần đây, sự phát triển nhanh chóng của các mô hình ngôn ngữ lớn (Large Language Models - LLMs) đã mở ra nhiều cơ hội ứng dụng trong lĩnh vực y học, đặc biệt là ở các nhiệm vụ chẩn đoán và hỗ trợ điều trị. Song song với đó, kỹ thuật tăng cường truy xuất (Retrieval-Augmented Generation - RAG) cho phép tích hợp tri thức bên ngoài vào quá trình tạo phản hồi, qua đó nâng cao độ chính xác và giảm thiểu hiện tượng "hallucination" - một hạn chế phổ biến ở LLM truyền thống.

Chẩn đoán bệnh da liễu là một lĩnh vực có tiềm năng đặc biệt trong việc ứng dụng LLM và RAG, nhờ đặc trưng dữ liệu đa dạng (văn bản, hình ảnh lâm sàng, mô bệnh học) và nhu cầu ngày càng cao đối với các hệ thống hỗ trợ ra quyết định lâm sàng (Clinical Decision Support Systems - CDSS). Giai đoạn 2023-2025 ghi nhận nhiều bước tiến đáng chú ý, bao gồm: SkinGPT-4 [1] - hệ thống thị giác – ngôn ngữ tương tác đầu tiên chuyên biệt cho da liễu, đánh dấu mốc quan trọng trong tích hợp LLM với xử lý hình ảnh y tế; SkinGEN [2] - mở rộng khả năng giải thích lâm sàng thông qua sinh ảnh minh họa cá nhân hóa; các bộ dữ liệu quy mô lớn như MM-Skin [3] và Derm1M [4] - được xây dựng theo cấu trúc ontology, cung cấp nguồn dữ liệu chuẩn hóa cho huấn luyện mô hình; và MRD-RAG [5] - ứng dụng cách tiếp cận đa vòng, kết hợp RAG với mô hình đa phương thức nhằm nâng cao khả năng suy luận lâm sàng.

Tuy nhiên, việc triển khai LLM và RAG trong chẩn đoán da liễu vẫn còn đối mặt với nhiều thách thức, bao gồm hạn chế trong xử lý hình ảnh y khoa phức tạp, nguy cơ sai lệch thông tin, thiếu minh bạch trong giải thích, cùng các rào cản đạo đức và pháp lý.

Ngoài ra, các tổng quan hiện có về ứng dụng trí tuệ nhân tạo trong da liễu trong giai đoạn này còn hạn chế về phạm vi và độ sâu. Lewandowski et al. [6] là công trình duy nhất tập trung vào LLM trong da liễu, nhưng chủ yếu đánh giá các mô hình thuần văn bản, chưa xem xét mô hình đa phương thức hoặc kỹ thuật RAG. Abdalla et al. [7] tổng hợp vai trò của multimodal imaging trong chẩn đoán da liễu nhưng không đề cập đến LLM hoặc khả năng tích hợp tri thức bên ngoài. Các tổng quan rộng hơn như Shool et al. [8], Alkalbani et al. [9] và Amugongo et al. [10] khảo sát LLM hoặc RAG trong y học nói chung, nhưng thiếu phân tích chuyên sâu về da liễu và các mô hình thị giác – ngôn ngữ hiện đại.

Nhìn chung, chưa có công trình nào đồng thời kết nối đầy đủ ba hướng nghiên cứu quan trọng: (1) hiệu năng LLM trong da liễu, (2) bộ dữ liệu thị giác – ngôn ngữ chuyên biệt, và (3) tiềm năng của RAG trong suy luận lâm sàng. Khoảng trống này nhấn mạnh nhu cầu thực hiện một tổng quan hệ thống, cập nhật và toàn diện về LLM, RAG và các mô hình đa phương thức trong chẩn đoán da liễu để đánh giá hiện trạng, xác định tiềm năng


https://doi.org/10.55401/5e54wg60

và định hướng phát triển các hệ thống AI y tế thông minh, an toàn và hiệu quả.

## 2. Phương pháp nghiên cứu

Nghiên cứu này được thực hiện dưới dạng bài tổng quan hệ thống (systematic review), nhằm khảo sát và phân tích các công trình nghiên cứu tiêu biểu trong giai đoạn từ tháng 1 năm 2023 đến tháng 5 năm 2025 liên quan đến việc tích hợp LLMs và kỹ thuật RAG trong chẩn đoán da liễu.

### 2.1. Chiến lược tìm kiếm và chọn lọc tài liệu

Dữ liệu được thu thập từ các cơ sở dữ liệu học thuật uy tín bao gồm PubMed, IEEE Xplore, ACM Digital Library, SpringerLink và arXiv. Các từ khóa chính được sử dụng trong quá trình tìm kiếm bao gồm: “Large Language Models”, “LLM”, “Retrieval-Augmented Generation”, “RAG”, “Dermatology”, “Dermatological Dataset”, “Skin Disease Diagnosis”, "Dermatological Diagnosis" và “Clinical Decision Support”. Tìm kiếm được thực hiện với bộ lọc về thời gian (trong khoảng từ tháng 1/2023 đến tháng 5/2025), ngôn ngữ (tiếng Anh) và loại tài liệu (bài báo bình duyệt, preprint có ảnh hưởng). Sau quá trình sàng lọc tiêu đề, tóm tắt và toàn văn, tổng cộng 10 bài báo đáp ứng tiêu chí đã được lựa chọn để đưa vào phân tích chi tiết.

Tiêu chí chọn lọc bao gồm: (1) các nghiên cứu có tích hợp LLM trong chẩn đoán hoặc hỗ trợ lâm sàng da liễu; (2) có thể có ứng dụng hoặc đề xuất kỹ thuật RAG (truy xuất tri thức ngoài mô hình) hoặc phương pháp truy xuất dữ liệu tương đương; (3) mô tả chi tiết về hệ thống, kiến trúc hoặc thử nghiệm lâm sàng; (4) công bố trong giai đoạn tháng 1/2023 đến tháng 5/2025. Những bài viết không liên quan đến lĩnh vực bệnh da liễu, không sử dụng LLM hoặc không sử dụng truy xuất dữ liệu đều bị loại trừ.

### 2.2. Chiến lược phân tích và đánh giá

Việc phân tích được thực hiện dựa trên các tiêu chí định tính và định lượng sau:

Các bài báo được phân tích theo nhiều khía cạnh, bao gồm: mục tiêu và phạm vi nghiên cứu, mô hình LLM sử dụng, dữ liệu huấn luyện và thử nghiệm, chỉ số đánh giá, khả năng ứng dụng thực tiễn trong môi trường lâm sàng.

Ngoài ra, nghiên cứu cũng tiến hành so sánh các phương pháp theo ba khía cạnh chính: (1) hiệu quả mô hình và độ chính xác trong chẩn đoán; (2) chiến lược tích hợp tri thức và khả năng mở rộng; (3) yêu cầu tài nguyên tính toán. Dữ liệu từ các bài báo được chuẩn hóa và tổ chức trong bảng tính Excel.

Các bước chọn lọc được thực hiện bởi hai nhà nghiên cứu độc lập và đối chiếu kết quả để đảm bảo tính khách quan. Một số công trình preprint được đưa vào nếu có chất lượng học thuật cao và có ảnh hưởng đáng kể trong cộng đồng nghiên cứu. Tuy nhiên, nghiên cứu vẫn tồn tại giới hạn nhất định, bao gồm khả năng bỏ sót các tài liệu ngoài hệ thống cơ sở dữ liệu chính hoặc các nghiên cứu công bố bằng ngôn ngữ khác ngoài tiếng Anh.

## 3. Kết quả và thảo luận

Sau quá trình chọn lọc và phân tích các công trình nghiên cứu cho thấy việc tích hợp LLM với kỹ thuật RAG trong lĩnh vực chẩn đoán da liễu vẫn còn là một lĩnh vực mới mẻ, số lượng bài nghiên cứu được công bố chính thức vẫn còn ít. Trong số đó, 10 công trình đã lọc được được đánh giá là tiêu biểu dựa trên mức độ ảnh hưởng, tính đổi mới kỹ thuật và độ chi tiết về thực nghiệm.

Các công trình này sẽ trình bày thành ba nhóm:

i. Đánh giá năng lực của các mô hình ngôn ngữ lớn LLM trong lĩnh vực chẩn đoán bệnh da liễu thông qua việc kiểm nghiệm các mô hình LLM bằng các bài test chuyên môn y khoa.

ii. Tăng cường hiểu biết lâm sàng của LLM bằng tri thức y học bên ngoài (knowledge bases, guidelines, EMR) thông qua việc xây dựng các bộ dữ liệu chuẩn dùng trong lĩnh vực huấn luyện LLM chẩn đoán bệnh da liễu.

iii. Đánh giá triển vọng của việc tích hợp LLM với RAG trong lĩnh vực y sinh nói chung, và xây dựng ứng dụng hỗ trợ bác sĩ, hỏi-đáp và phân tích triệu chứng da liễu nói riêng.

### 3.1. Nhóm 1: Đánh giá tiềm năng của các mô hình ngôn ngữ lớn trong chẩn đoán bệnh da liễu

Hai nghiên cứu [11] và [12] cùng đánh giá hiệu suất của các mô hình ngôn ngữ lớn (LLM) trong trả lời câu hỏi chuyên ngành da liễu, nhằm xác định tiềm năng ứng dụng trong đào tạo và hỗ trợ lâm sàng. Nghiên cứu [11] khảo sát hiệu năng tổng hợp của năm mô hình (ChatGPT-4o, Claude-3.5 Sonnet, Copilot, Gemini, Perplexity) trên 100 câu hỏi từ kỳ thi SCE tại Anh, chủ yếu là câu hỏi văn bản, chỉ có 4 câu chứa hình ảnh. Trong khi đó, nghiên cứu [12] tập trung phân tích riêng ChatGPT-4o trên 300 câu hỏi luyện thi kiểu hội đồng, với tỷ lệ lớn hơn các câu hỏi có hình ảnh (133/300), bao gồm ảnh lâm sàng và mô bệnh học.

Cả hai công trình đều ghi nhận ChatGPT-4o vượt ngưỡng điểm đạt (≥ 70 %): 90% ở [11] và 77,3 % ở [12]. Tuy nhiên, khi phân tích sâu hơn, hiệu suất mô hình giảm rõ rệt với các câu hỏi hình ảnh (67,7 %), đặc biệt với ảnh mô bệnh học (58,8 %), cho thấy hạn chế trong xử lý dữ liệu thị giác. Cả hai nghiên cứu cũng chỉ ra nguy cơ phát sinh thông tin sai lệch (hallucination), rủi ro đạo đức và bảo mật khi sử dụng hình ảnh bệnh nhân, cũng như sự thiếu tối ưu hóa mô hình cho chuyên ngành da liễu.

Tổng thể, hai nghiên cứu bổ sung cho nhau về phạm vi và chiều sâu đánh giá: một cung cấp góc nhìn so sánh đa mô hình, một đi sâu vào phân tích hiệu năng theo từng loại dữ liệu. Cả hai cùng cho thấy LLMs, đặc biệt là ChatGPT, đang dần đạt đến ngưỡng có thể ứng dụng thực tiễn trong đào tạo và hỗ trợ lâm sàng da liễu, nhưng vẫn cần cải thiện về thị giác và tối ưu hóa chuyên biệt.

Bảng 1. So sánh tổng hợp hai bài báo [11] và [12]

| Tiêu chí | Dermatological Knowledge of LLMs [11] | Computer Vision Meets LLMs [12] |
|---|---|---|
| **Mô hình đánh giá** | ChatGPT-4o, Claude-3.5 Sonnet, Copilot, Gemini, Perplexity | Chỉ ChatGPT 4.0 |
| **Số lượng câu hỏi** | 100 câu (4 câu có hình ảnh) từ đề thi SCE | 300 câu (169 văn bản, 133 hình ảnh) từ DermQbank |
| **Loại câu hỏi** | Chủ yếu văn bản, rất ít hình ảnh | Văn bản + hình ảnh (lâm sàng & mô bệnh học) |
| **Hiệu suất từng mô hình** | ChatGPT-4o: **90 %**<br>Claude-3.5 Sonnet: **87 %**<br>Copilot: **88 %**<br>Perplexity: **87 %**<br>Gemini: **75 %** | ChatGPT 4.0:<br>Tổng thể: **77,3 %**<br>Văn bản: **85,2 %**<br>Hình ảnh: **67,7 %**<br>• Ảnh lâm sàng: **69,0 %**<br>• Ảnh mô bệnh học: **58,8 %** |
| **So sánh với người thật** | So với chuẩn đầu ra bác sĩ nội trú SCE (≥ 70 %) | So với bác sĩ nội trú năm cuối (PGY-4), phân vị 46 |
| **Khả năng giải thích và minh bạch** | Có giải thích; chưa đánh giá chất lượng lời giải thích | Tỷ lệ đồng thuận với đáp án bác sĩ: **75,3 %** |
| **Hạn chế chính** | Thiếu câu hỏi hình ảnh, thiếu điều chỉnh theo ngữ cảnh chuyên ngành, chưa đánh giá phản hồi cải tiến | Hiệu suất thấp với hình ảnh, hallucination, dữ liệu cũ, lo ngại về bảo mật ảnh |
| **Phân tích thống kê** | Không có phân tích định lượng | Có phân tích hồi quy logistic ($P < .001$) |
| **Ý nghĩa chính** | LLMs có khả năng hỗ trợ đào tạo y khoa; ChatGPT-4o và Copilot nổi bật | ChatGPT 4.0 có thể hỗ trợ học tập, cải thiện nhiều qua thời gian, nhưng vẫn hạn chế về hình ảnh |

### 3.2. Nhóm 2. Xây dựng bộ dữ liệu chẩn đoán da liễu

Để giải quyết bài toán thiếu hụt dữ liệu chuẩn hóa phục vụ huấn luyện AI trong chẩn đoán da liễu, hai công trình [3] và [4] đã xây dựng các bộ dữ liệu ngôn ngữ – thị giác chuyên biệt MM-Skin [3] và Derm1M [4], kết hợp hình ảnh da liễu với mô tả y khoa theo chuẩn chuyên ngành.

MM-Skin [3] là bộ dữ liệu đa phương thức đầu tiên trong da liễu, gồm hơn 10.000 cặp ảnh – văn bản được trích từ 15 giáo trình chuyên ngành, bao gồm ảnh lâm sàng, ảnh soi da và mô bệnh học. Bộ dữ liệu này được sử dụng để huấn luyện SkinVL, một mô hình Vision - Language được tinh chỉnh từ kiến trúc LlaVA-Med [8]. SkinVL thể hiện hiệu suất nổi bật ở các tác vụ như phân loại bệnh, hỏi đáp hình ảnh (VQA), và phân loại zero-shot, nhờ tận dụng tri thức từ văn bản chuyên môn và hướng dẫn QA sinh bởi LLM. Tuy nhiên, MM-Skin [3] còn hạn chế về quy mô, thiếu dữ liệu thực địa và phân bố ảnh chưa đồng đều, ảnh hưởng đến khả năng tổng quát hóa.

Derm1M [4], ra đời sau, cung cấp bộ dữ liệu lớn nhất hiện nay trong lĩnh vực với hơn 1 triệu cặp ảnh – văn bản, thu thập từ nhiều nguồn như video giáo dục, tài liệu khoa học, diễn đàn và dữ liệu công khai. Điểm nổi bật của Derm1M [4] là tích hợp ontology lâm sàng, liên kết hình ảnh với 390 bệnh và 130 khái niệm y khoa (vị trí tổn thương, màu da, dạng tổn thương, ...). Dựa trên nền tảng này, nhóm tác giả phát triển DermLIP – một kiến trúc tương tự CLIP, có khả năng phân loại, truy xuất và nhận diện khái niệm hiệu quả trong thiết lập zero-shot và few-shot. Tuy nhiên, Derm1M [4] cũng gặp thách thức như nhiễu từ nguồn dữ liệu không kiểm chứng (ví dụ YouTube), và thiếu kiểm định thủ công khi gán nhãn hoặc tạo mô tả bằng LLM.

Tóm lại, MM-Skin [3] nổi bật về chất lượng chuyên môn và ứng dụng đa nhiệm, còn Derm1M [4] có ưu thế vượt trội về quy mô, tính đại diện và khả năng hỗ trợ mô hình tổng quát hóa. Cả hai cùng cho thấy xu hướng tích hợp thị giác và ngôn ngữ với tri thức y học để hỗ trợ chẩn đoán lâm sàng hiệu quả hơn.

Bảng 2. Bảng thống kê các đặc điểm của 02 bộ dữ liệu chẩn đoán da liễu MM-Skin và Derm1M

| Bộ dữ liệu | MM-Skin [3] | Derm1M [4] |
|---|---|---|
| **Quy mô dataset** | 11.039 ảnh, 9.369 mô tả, 27.412 cặp hỏi đáp (QA) | 1.029.761 cặp ảnh – văn bản |
| **Nguồn dữ liệu** | 15 sách giáo khoa chuyên ngành da liễu | YouTube, PubMed, diễn đàn y học, sách giáo khoa, bộ dữ liệu công khai |
| **Loại ảnh** | Ảnh lâm sàng (clinical),<br>Ảnh soi da (dermoscopic),<br>Ảnh mô bệnh học (pathological) | Ảnh lâm sàng,<br>Ảnh soi da |
| **Mô tả văn bản** | Mô tả chuyên sâu, chính thống từ sách; dài và giàu thông tin y khoa | Sinh từ video, bài viết, diễn đàn; tăng cường bằng ontology và mô hình ngôn ngữ lớn (LLM) |
| **Ontology / Nhãn khái niệm** | Không dùng ontology chính thức; có phân loại mô bệnh học và trích xuất thuộc tính bệnh nhân | Có 390 bệnh, 130 khái niệm lâm sàng; phân cấp thành 4 tầng theo cấu trúc ontology |
| **Thông tin** | Tuổi, giới tính (trích từ mô tả sách giáo khoa) | Giới tính, màu da, vị trí cơ thể,... |
| **Nhiệm vụ hỗ trợ** | Hỏi đáp thị giác (VQA), phân loại không giám sát | Phân loại không cần huấn luyện trước (zero-shot), học với ít dữ liệu (few-shot), truy hồi đa mô thức, phát hiện khái niệm |
| **Ưu điểm** | Mô hình thể hiện hiệu năng vượt trội trong nhiều tác vụ, có khả năng tổng quát cao, tạo phản hồi giàu ngữ cảnh và giảm thiểu "hallucination" | Derm1M là bộ dữ liệu đa phương thức có quy mô và tính đa dạng cao, tích hợp kiến thức y khoa chuẩn hóa, hỗ trợ xây dựng các mô hình học sâu chuyên biệt trong da liễu mà không yêu cầu nhãn thủ công. |
| **Hạn chế chính** | Dữ liệu mất cân bằng (ảnh da chiếm <10%)<br>Thiếu nhãn chẩn đoán trong SkinVL-MM | Một số nguồn dữ liệu có thể chứa nhiễu hoặc thông tin chưa xác thực.<br>Phụ thuộc vào mô hình ngôn ngữ để tạo nhãn có thể gây sai lệch nếu không được kiểm định.<br>Còn bất cập trong gán nhãn ontology<br>Chưa có đánh giá thực nghiệm trong môi trường lâm sàng. |

### 3.3. Nhóm 3: Đánh giá triển vọng của việc tích hợp các mô hình ngôn ngữ lớn và kỹ thuật tăng cường truy xuất trong lĩnh vực y sinh và xây dựng hệ thống chẩn đoán da liễu

Giai đoạn 2023 - 2025 ghi nhận nhiều công trình tiêu biểu ứng dụng LLM và kỹ thuật đa phương thức trong chẩn đoán da liễu, phản ánh rõ xu hướng phát triển từ các hệ thống chẩn đoán đơn giản sang nền tảng AI y tế thông minh, có khả năng giải thích và tương tác linh hoạt hơn.

SkinGPT-4 [1] ) là hệ thống tiên phong kết hợp LLM với mô hình thị giác (Visual LLM), đánh dấu bước chuyển từ LLM thuần văn bản sang mô hình đa phương thức. Hệ thống được xây dựng trên MiniGPT-4, ViT và Vicuna, huấn luyện với dữ liệu da liễu lớn kèm chú thích chuyên sâu. Đánh giá lâm sàng cho thấy độ chính xác và tính hữu ích cao (≈78 - 83) %, đồng thời đảm bảo quyền riêng tư nhờ triển khai cục bộ. Đây là nền tảng kỹ thuật quan trọng cho các hệ thống kế tiếp như SkinGEN.

SkinGEN [2] mở rộng từ SkinGPT-4, chuyển trọng tâm từ chẩn đoán sang giải thích kết quả thông qua hình ảnh minh họa, đáp ứng yêu cầu tăng tính minh bạch và khả năng giải thích (explainability) trong AI y tế. Hệ thống tích hợp Stable Diffusion để sinh ảnh mô phỏng vùng tổn thương, giúp cải thiện đáng kể độ tin cậy, khả năng hiểu và mức độ sẵn sàng sử dụng lại của người dùng.

PanDerm [13] hướng đến xây dựng mô hình nền (foundation model) cho da liễu, áp dụng học không giám sát trên hơn 2 triệu ảnh từ nhiều trung tâm y tế. Sử dụng ViT-L và CLIP teacher, PanDerm đạt hiệu suất vượt trội so với bác sĩ và các mô hình SOTA, đồng thời chứng minh tiềm năng của self-supervised learning

trong y tế, đặc biệt ở các lĩnh vực hạn chế dữ liệu gán nhãn.

MAKE [14] tập trung khai thác tri thức lâm sàng đa khía cạnh và tối ưu cho thiết lập zero-shot, cho phép mô hình hoạt động hiệu quả ngay cả khi thiếu dữ liệu huấn luyện mới. Phương pháp chia nhỏ mô tả lâm sàng, liên kết từng phần với vùng ảnh tương ứng và gán trọng số theo giá trị chẩn đoán giúp MAKE đạt hiệu suất cao ở phân loại bệnh, chú thích khái niệm và truy xuất ảnh.

MRD-RAG [5] đánh dấu sự dịch chuyển từ chẩn đoán tĩnh sang mô phỏng quy trình hội thoại lâm sàng đa vòng, tích hợp RAG để tăng khả năng suy luận theo ngữ cảnh. Bằng cách kết hợp LLM với cây tri thức y khoa (hiện đại và Đông y), MRD-RAG cải thiện đáng kể độ chính xác chẩn đoán so với RAG đơn vòng, đặt nền tảng cho các hệ thống CDSS hội thoại thông minh.

Tóm lại, từ SkinGPT-4 [1] (chẩn đoán ảnh một vòng) → SkinGEN [2] (giải thích bằng ảnh cá nhân hóa) → PanDerm [13] (mô hình thị giác nền) → MAKE [14] (mô phỏng ra quyết định zero-shot) → MRD-RAG [5] (đối thoại suy luận nhiều vòng), ta thấy một tiến trình hội tụ giữa thị giác, ngôn ngữ và lập luận lâm sàng. Đây là xu hướng rõ ràng cho AI y tế tương lai - nơi các hệ thống không chỉ chẩn đoán, mà còn giải thích, tương tác và học hỏi liên tục từ tri thức chuyên ngành.

Bảng 3. Bảng tổng hợp so sánh 5 mô hình chẩn đoán da liễu ứng dụng các mô hình ngôn ngữ lớn hoặc mô hình ngôn ngữ thị giác

| Tiêu chí | SkinGPT-4 [1] | SkinGEN [2] | PanDerm [13] | MAKE [14] | MRD-RAG [5] |
|---|---|---|---|---|---|
| **Mục tiêu** | Chẩn đoán bệnh da liễu từ ảnh + mô tả bệnh | Sinh ảnh minh họa bệnh lý da để hỗ trợ tư vấn và giải thích kết quả | Học biểu diễn tổn thương da không giám sát | Chẩn đoán bệnh ít gặp bằng kết hợp ảnh và tri thức y học | Mô phỏng quy trình hội thoại chẩn đoán nhiều vòng như bác sĩ thực thụ |
| **Kiến trúc chính** | Visual LLM (miniGPT-4 + Vision Encoder) Vision Encoder sử dụng ViT + BLIP-2 (Q-Former) + Vicuna | Visual LLM (SkinGPT-4) + Diffusion Model + GroundingDINO + skinSAM | ViT-L + CLIP + Self-supervised latent alignment | ViT + mô hình VLM + trọng số lâm sàng đa khía cạnh | Multi-Round RAG (retriever → analyzer → doctor) với LLM và DI-Tree |
| **Đặc điểm chính** | Kết hợp ảnh + mô tả bệnh để chẩn đoán | Tăng khả năng giải thích bằng cách sinh ảnh minh họa bệnh | Học không giám sát từ ảnh tổn thương da | Gắn kết tri thức đa chiều (hình ảnh, ngôn ngữ, ontology) | Tương tác hội thoại đa vòng, tích hợp tri thức Đông Tây y |
| **Đầu vào** | Ảnh bệnh + mô tả văn bản | Ảnh bệnh + mô tả văn bản | Ảnh tổn thương không nhãn | Ảnh + mô tả bệnh + tri thức y khoa + ontology | Mô tả bệnh, tiền sử bệnh, dữ liệu triệu chứng, hội thoại nhiều vòng |
| **Dữ liệu huấn luyện** | Ảnh có nhãn + mô tả từ bộ dữ liệu da liễu công khai 2929 ảnh (DermNet + nội bộ + SKINCON) | Ảnh có nhãn + mô tả + ảnh tổng hợp sinh bởi mô hình 27000 ảnh (Fitzpatrick17k, SCIN) + mô phỏng ảnh | Hình ảnh không gán nhãn quy mô lớn 2 triệu ảnh từ 11 bệnh viện (4 loại ảnh: lâm sàng, soi da, kính hiển vi, toàn thân) | Hình ảnh có nhãn + mô tả + ontology bệnh SkinCAP dataset (có mô tả đa chiều, cấu trúc tốt) | Dữ liệu mô phỏng triệu chứng + DI-Tree + tiền sử bệnh nhân |
| **Chiến lược huấn luyện** | Fine-tuning LLM/VLM | Fine-tuning VLM + huấn luyện Stable Diffusion | Contrastive Learning (không giám sát) | Pretraining Vision-Language Alignment + | Prompting LLM với pipeline truy xuất đa vòng, không cần fine-tune |

| Tiêu chí | SkinGPT-4 [1] | SkinGEN [2] | PanDerm [13] | MAKE [14] | MRD-RAG [5] |
|---|---|---|---|---|---|
| | | | | Ontology Injection | |
| **Kỹ thuật chính** | GPT-4 + Vision Encoder | Stable Diffusion + Visual LLM | ViT + BYOL / SimCLR | Multi-aspect Knowledge Embedding + Visual-Language Mapping | Multi-Round Retrieval-Augmented Generation (MRD-RAG) |
| **Khả năng tương tác** | Có – đối thoại đơn vòng | Có – sinh ảnh + giải thích trực quan | Không – chỉ encode ảnh | Có – tương tác dạng truy vấn + phản hồi | Có – đối thoại nhiều vòng giống bác sĩ |
| **Khả năng giải thích** | Trung bình – văn bản tổng hợp | Cao – sinh ảnh minh họa bệnh | Thấp – không có phần giải thích | Cao – truy vấn dựa vào tri thức và so sánh đa chiều | Trung bình – giải thích qua phản hồi và truy xuất bệnh |
| **Hiệu suất** | Khá cao – tùy thuộc chất lượng ảnh và mô tả<br>Chính xác: 78,13 %,<br>hữu ích: 83,13 % | Tăng độ tin cậy thông qua hình ảnh minh họa<br>Độ dễ hiểu: 4,38/5, độ tin cậy: 4,09/5 | Tốt cho tiền xử lý cho các task khác<br>AUROC dự đoán di căn: 0,964;<br>Phát hiện melanoma: tăng 10,2 % so bác sĩ | Hiệu quả tốt trong zero-shot và few-shot<br>Zero-shot classification: 49,29 % (tăng 5.09 %),<br>AUROC: 0,7369 | Hiệu suất vượt RAG đơn vòng: tăng 9,4 % (so với GPT) và 21,75 % (so với bác sĩ đánh giá) |
| **Ưu điểm** | Hệ thống đa modal dễ tích hợp, hỗ trợ đa dạng đầu vào | Giải thích sinh động bằng hình ảnh, nâng cao niềm tin người dùng | Không cần nhãn - mở rộng nhanh | Giải thích tốt, mô hình hiểu sâu bằng tri thức, mạnh trong zero - shot reasoning | Giống quy trình chẩn đoán thực tế, hiệu quả cao, dùng được cho cả Y học hiện đại và Y học cổ truyền |
| **Hạn chế** | Giải thích chưa sâu, phụ thuộc chất lượng mô tả và ảnh | Tốn tài nguyên tính toán, phụ thuộc chất lượng sinh ảnh | Thiếu tương tác, không có chẩn đoán | Cần tri thức cấu trúc tốt, huấn luyện phức tạp | Chưa đánh giá trên dữ liệu bệnh nhân thật, phụ thuộc truy xuất đúng bệnh |
| **Ứng dụng chính** | Hỗ trợ chẩn đoán lâm sàng, tư vấn bệnh nhân | Tư vấn và giáo dục sức khỏe qua ảnh minh họa bệnh | Học biểu diễn tổn thương da, hỗ trợ tiền xử lý cho các task khác | Chẩn đoán bệnh lạ, bệnh hiếm trong bối cảnh thiếu dữ liệu<br>Mở rộng hệ thống AI y tế thông minh | Tích hợp chatbot bác sĩ AI trong bệnh viện hoặc nền tảng khám bệnh trực tuyến |

## 4. Kết luận và đề xuất hướng phát triển

### 4.1. Kết luận

Việc tích hợp LLM và RAG trong chẩn đoán da liễu đang dần khẳng định tiềm năng ứng dụng thực tiễn trong đào tạo y khoa, hỗ trợ lâm sàng và xây dựng hệ thống chẩn đoán thông minh. Qua phân tích 10 công trình tiêu biểu, có thể thấy rằng:

1. ***Các mô hình LLM hiện đại như ChatGPT-4o, Claude-3.5 hay Copilot đã đạt hiệu suất khả quan trong các bài kiểm tra chuyên ngành da liễu***, vượt qua ngưỡng điểm chuẩn đầu ra của bác sĩ nội trú, đặc biệt trong xử lý văn bản. Tuy nhiên, khả năng xử lý hình ảnh y khoa vẫn còn hạn chế, nhất là với ảnh mô bệnh học – điều này cản trở việc ứng dụng toàn diện trong chẩn đoán thực tế.

2. ***Các nỗ lực xây dựng bộ dữ liệu ngôn ngữ – thị giác chuyên biệt*** như MM-Skin [3] và Derm1M [4] đóng vai trò cốt lõi trong việc huấn luyện và đánh giá các mô hình AI da liễu. Trong khi MM-Skin [3] chú trọng chất lượng chuyên môn và thiết kế mô hình đa nhiệm từ giáo trình, Derm1M [4] tạo lợi thế về quy mô, tính đa dạng và gắn kết với ontology y khoa.

3. Các mô hình kết hợp LLM và RAG (như MRD-RAG [5], MAKE [14]) đã bắt đầu được triển khai thử nghiệm trong ***hệ thống hỏi – đáp, hỗ trợ ra quyết định lâm sàng (CDSS) và nhận diện triệu chứng***, mở ra triển vọng tích hợp sâu vào hệ thống chăm sóc sức khỏe thông minh.

Tuy nhiên, vẫn tồn tại các thách thức cần giải quyết như:

- Hiệu suất thấp khi xử lý hình ảnh phức tạp hoặc không đồng nhất.
- Nguy cơ "hallucination" và sự thiếu minh bạch trong giải thích của LLM.
- Rào cản đạo đức, bảo mật và tính khả dụng của dữ liệu bệnh nhân trong môi trường triển khai thực tế.

### 4.2. Đề xuất hướng phát triển

Dựa trên các phân tích trên, một số hướng nghiên cứu và phát triển tương lai được đề xuất như sau:

1. **Tối ưu hóa LLM chuyên biệt cho ngành da liễu**: Thiết kế các mô hình ngôn ngữ - hình ảnh được huấn luyện từ đầu hoặc tinh chỉnh sâu (fine-tune) dựa trên bộ dữ liệu da liễu đa dạng, thay vì sử dụng các LLM tổng quát.
2. **Tăng cường khả năng thị giác y khoa**: Kết hợp các mô hình thị giác tiên tiến như ViT, SAM, hoặc foundation model y sinh (BioMedCLIP, MedSAM) với LLM để cải thiện hiệu suất trên ảnh lâm sàng và mô bệnh học.
3. ***Phát triển cơ chế kiểm định độ tin cậy và giải thích (trustworthiness & explainability):*** Tích hợp các module đánh giá lời giải thích, xác minh nguồn truy xuất và phản hồi bác sĩ nhằm tăng tính minh bạch và giảm rủi ro "hallucination".
4. ***Chuẩn hóa và mở rộng bộ dữ liệu***: Tăng cường nỗ lực xây dựng kho dữ liệu công khai chất lượng cao với ảnh – văn bản có gán nhãn xác thực và liên kết với hệ thống tri thức y khoa. Cần có sự hợp tác đa ngành giữa bác sĩ, chuyên gia AI và nhà phát triển hệ thống y tế.
5. ***Triển khai thử nghiệm lâm sàng có kiểm soát (pilot study):*** Trước khi ứng dụng thực tiễn, các mô hình cần được đánh giá toàn diện thông qua các nghiên cứu lâm sàng quy mô nhỏ, có đối chứng, nhằm đảm bảo tính an toàn và hiệu quả.
6. ***Xây dựng nền tảng CDSS kết hợp LLM – RAG có khả năng tương tác:*** Thiết kế hệ thống hỏi – đáp y khoa chuyên sâu, cá nhân hóa theo ngữ cảnh bệnh nhân và lịch sử y tế, tạo tiền đề cho trợ lý y tế thông minh trong tương lai.

**Lời cảm ơn:** Nhóm tác giả xin chân thành cảm ơn PGS.TS. Bạch Long Giang (Đại học Nguyễn Tất Thành) và các nhà phản biện ẩn danh của Tạp chí đã có những nhận xét chi tiết và xây dựng, góp phần quan trọng nâng cao chất lượng bài báo.

**Tài liệu tham khảo**

**[1]** J. Zhou *et al*., "Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4," *Nature Communications*, 2023. [Online]. Available: https://www.nature.com/articles/s41467-024-50043-3

**[2]** B. Lin, Y. Xu, X. Bao, Z. Zhao, Z. Wang, and J. Yin, "SkinGEN: An explainable dermatology diagnosis to generation framework with interactive vision language models," in *Proc. 30th Int. Conf. Intelligent User Interfaces (IUI '25)*, ACM, 2025. doi: 10.1145/3708359.3712098.

**[3]** W. Zeng, Y. Sun, C. Ma, W. Tan, and B. Yan, "MM Skin: Enhancing dermatology vision language model with an image text dataset derived from textbooks," in *Proc. 33rd ACM Int. Conf. Multimedia (MM '25)*, ACM, 2025. doi: 10.1145/3746027.3755187.

**[4]** S. Yan *et al*., "Derm1M: A million-scale vision–language dataset aligned with clinical ontology knowledge for dermatology," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV 2025)*, pp. 12681–12690, Oct. 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2025/html/Yan_Derm1M_A_Million-scale_Vision-Language_Dataset_Aligned_with_Clinical_Ontology_Knowledge_ICCV_2025_paper.html

**[5]** Y. Chen, P. Sun, X. Li, and X. Chu, "The multi-round diagnostic RAG framework for emulating clinical reasoning," *arXiv preprint* arXiv:2504.07724, 2025. [Online]. Available: https://arxiv.org/abs/2504.07724
**[6]** M. Lewandowski, J. Kropidłowska, A. Kvinen, and W. Barańska-Rybak, "A systemic review of large language models and their implications in dermatology," *Australas. J. Dermatol.*, vol. 66, no. 4, pp. e202–e208, Jun. 2025. doi: 10.1111/ajd.14484.
**[7]** B. M. Z. Abdalla *et al*., "Noninvasive multimodal imaging and its role in diagnosing skin lesions in dermatology: A systematic review and meta-analysis," *Am. J. Clin. Dermatol.*, vol. 26, no. 5, pp. 711–722, 2025.
**[8]** S. Shool, S. Adimi, R. S. Amleshi, E. Bitaraf, R. Golpira, and M. Tara, "A systematic review of large language model (LLM) evaluations in clinical medicine," *BMC Med. Inf. Decis. Mak.*, vol. 25, p. 117, 2025.
**[9]** L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval-augmented generation for large language models in healthcare: A systematic review," *PLOS Digit. Health*, vol. 4, no. 6, e0000877, 2025.
**[10]** A. M. Alkalbani *et al*., "A systematic review of large language models in medical specialties: Applications, challenges and future directions," *Information*, vol. 16, no. 6, 489, 2025.
**[11]** K. S. Fan and K. H. Fan, "Dermatological knowledge and image analysis performance of large language models based on specialty certificate examination in dermatology," *AI*, vol. 4, no. 4, p. 13, 2024.
**[12]** L. R. Smith, R. E. Hanna, L. A. Hatch, and K. Hanna, "Computer vision meets large language models: Performance of ChatGPT 4.0 on dermatology boards-style practice questions," *SKIN: J. Cutaneous Med.*, 2024.
**[13]** S. Yan *et al*., "A multimodal vision foundation model for clinical dermatology," *Nature Medicine*, 2025.
**[14]** S. Yan, X. Li, M. Hu, Y. Jiang, Z. Yu, and Z. Ge, "MAKE: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment," *arXiv preprint* arXiv:2505.09372, 2025. [Online]. Available: https://arxiv.org/abs/2505.09372

# Overview of Large Language Models and Retrieval-Augmented Generation Techniques in Dermatological Diagnosis

Bùi Thị Thanh Tú[1]*, Phạm Thị Anh Thư[2.]
[1, 2]Ho Chi Minh City University of Foreign Languages - Information Technology, Hochiminh City, Vietnam
*tubtt@huflit.edu.vn

**Abstract** This study conducts a systematic review of research published between January 2023 and May 2025 on the application of Large Language Models and Retrieval-Augmented Generation techniques in dermatological diagnosis. From more than 40 articles retrieved from academic databases, 10 representative studies were selected for in-depth analysis based on model architecture, data sources, performance, and clinical applicability. The findings indicate that LLMs - particularly when integrated with RAG and multimodal data - hold significant potential to improve diagnostic accuracy and support clinical decision-making, although challenges remain, including limitations in processing complex medical images, the risk of hallucination, and data privacy concerns. The study proposes future development directions, including the optimization of domain-specific LLMs for dermatology, enhancement of medical imaging capabilities, standardization of datasets, and controlled clinical trials, with the goal of advancing toward intelligent and trustworthy diagnostic support systems.
**Keywords** Large Language Models, Retrieval-Augmented Generation; Dermatological Diagnosis, Multimodal Reasoning; Clinical Decision Support.