Tạp chí Khoa học Điều dưỡng
Journal of Nursing Science
Journal homepage: www.jns.vn

Đánh giá bước đầu về khả năng cung cấp thông tin về sức khỏe của "hai trợ lý ảo cho bác sỹ và điều dưỡng" Copilot và Gemini Pro

Trương Việt Dũng[1], Nguyễn Thị Minh Chính[2], Vũ Thị Huệ[2], Trần Thị Nhàn[2], Trần Thị Hoàng Oanh[2], Đào Duy Minh Tuấn[2], Lương Thị Vân[2], Nguyễn Nhật Thành[3], Nguyễn Thị Xuyến[4], Ngô Văn Dần[4], Mai Xuân Thiên[5], Lê Thị Lý[6], Bùi Minh Nguyệt[7]
[1]Trường Đại học Tâm Anh; [2]Trường Đại học Điều dưỡng Nam Định; [3]Trường Đại học Thăng Long; [4]Bệnh viện Đa khoa Quốc tế Vinmec; [5]Bệnh viện Đa khoa Quốc tế Vinmec Timecity; [6]Bệnh viện Đa khoa tỉnh Tuyên Quang; [7]Bệnh viện 19-8

TÓM TẮT

Mục tiêu: *Đánh giá độ tin cậy và độ phù hợp của hai AI chatbot gồm Copilot và Gemini Pro trong cung cấp những thông tin để trả lời các câu hỏi về triệu chứng, chẩn đoán, điều trị, chăm sóc, tư vấn và dự phòng thông thường của người bệnh; Phân tích một số yếu tố liên quan đến điểm đánh giá của hai AI chatbot này.* ***Phương pháp nghiên cứu:*** *Nghiên cứu cắt ngang so sánh 2 cơ sở dữ liệu của 246 câu hỏi về sức khỏe, bệnh tật và 492 câu trả lời của hai chatbot Copilot và Gemini Pro vào tháng 1 năm 2026. Mỗi câu trả lời được đánh giá độc lập bởi 1 bác sỹ và 1 điều dưỡng chuyên khoa (theo 5 bệnh).* ***Kết quả:*** *Cả hai AI chatbot Gemini Pro và Copilot có độ tin cậy khá cao, với tất cả câu hỏi có điểm trung vị đều ≥ 4, điểm trung bình đạt từ 3,9 đến 4,7 trên thang 5 điểm. Tỷ lệ các câu trả lời ở mức đạt có tỷ lệ cao, với trên 81% do bác sỹ đánh giá Copilot, và 99,6% do điều dưỡng đánh giá Gemini. Mức độ đồng thuận giữa bác sĩ và điều dưỡng đối với Gemini rất cao (Kappa = 0,83) so với mức trung bình của Copilot (Kappa = 0,59). Một số yếu tố liên quan được chỉ ra: điều dưỡng viên có xu hướng đánh giá cao hơn so với bác sỹ, Gemini Pro được đánh giá cao hơn Copilot, độ phù hợp gemini tốt hơn copilot.* ***Kết luận****: Hai chatbot Copilot và Gemini Pro với độ tin cậy cao có thể sử dụng như trợ lý ảo cho công tác tư vấn của thầy thuốc.*

Từ khóa: *AI trợ lý ảo, Copilot, Gemini Pro, tư vấn sức khỏe*

Initial evaluation of the ability to provide health information of Copilot and Gemini Pro: "Two virtual assistants for physicians and nurses"

Truong Viet Dung[1], Nguyen Thi Minh Chinh[2], Vu Thi Hue[2], Tran Thi Nhan[2], Tran Thi Hoang Oanh[2], Dao Duy Minh Tuan[2], Luong Thi Van[2], Nguyen Nhat Thanh[3], Nguyen Thi Xuyen[4], Ngo Van Dan[4], Mai Xuan Thien[5], Le Thi Ly[6], Bui Minh Nguyet[7]
[1]Tam Anh University; [2]Nam Dinh University of Nursing; [3]Thang Long University; [4]Vinmec Times City International Hospital; [5]Vinmec Times City International Hospital [6]Tuyen Quang Provincial General Hospital, [7]19-8 Hospital

ABSTRACT

Objectives*: To evaluate the reliability and appropriateness of two AI chatbots, Copilot and Gemini Pro, in providing information regarding symptoms, diagnosis, treatment, care, consultation, and common disease prevention; To analyze factors associated with the evaluation scores of these two chatbots.* ***Methods****: A comparative cross-sectional study was conducted using datasets of 246 health-related questions and 492 responses generated by Copilot and Gemini Pro in January 2026. Each response was independently evaluated by one physician and one specialist nurse (across five disease groups). Results: Both Gemini Pro and Copilot demonstrated high reliability, with all median scores ≥ 4 and mean scores ranging from 3.9 to 4.7 on a 5-point scale. The rate of satisfactory responses was high, ranging from over 81% for Copilot (evaluated by physicians) to 99.6% for Gemini (evaluated by nurses). The agreement between physicians and nurses was very high for Gemini (Kappa = 0.83) compared to a moderate level for Copilot (Kappa = 0.59). Identified factors indicated that nurses tended to assign higher scores than physicians, and Gemini Pro was rated higher and demonstrated better appropriateness than Copilot.* ***Conclusion****: Copilot and Gemini Pro, demonstrating high reliability, can be utilized as virtual assistants to support healthcare professionals in patient consultation.*

Keywords*: AI virtual assistant, Copilot, Gemini Pro, health consultation*

Tác giả: Trương Việt Dũng
Email: Gsdungtruongviet@gmail.com
DOI: 10.54436/jns.2026.03.1325

Ngày nhận bài: 25/4/2026
Ngày hoàn thiện: 24/6/2026
Ngày đăng bài: 25/6/2026

ĐẶT VẤN ĐỀ

Khi trí tuệ nhân tạo (AI) tiếp tục phát triển, các mô hình ngôn ngữ quy mô lớn (NNL) như nhiều AI chatbot trở thành những công cụ được các thầy thuốc sử dụng như "trợ lý ảo" đầy hứa hẹn để tạo ra thông tin sức khỏe – bệnh tật, giúp nhiệm vụ tư vấn cho người bệnh được giảm nhẹ nhưng hiệu quả hơn [1], nhất là trong mô hình "*Chăm sóc hỗn hợp*" (Hybrid care) nhằm: nâng cao khả năng tiếp cận dịch vụ y tế; giảm tải cho bệnh viện [2]; tăng sự linh hoạt và tiện lợi cho bệnh nhân; giữ được yếu tố nhân văn và tương tác trực tiếp đảm bảo an toàn và hiệu quả, nhất là trong điều kiện hiện nay ngành y tế đang quá tải, thời gian dành cho mỗi lượt khám tư vấn rất hạn chế. AI chatbot cũng giúp người bệnh hiểu nên làm gì trong nhiều tình huống đối mặt với bệnh mà họ đang đối mặt trong khi chưa kịp hoặc không có điều kiện gặp nhân viên y tế [3]. Tuy nhiên, việc ứng dụng nhanh chóng và những lợi ích tiềm năng của chúng trong chăm sóc sức khỏe đòi hỏi phải đánh giá nghiêm ngặt về chất lượng, độ chính xác và tính an toàn của thông tin được tạo ra cho nhiều chuyên khoa khác nhau [4]. Nguyên tắc chung: chỉ coi chatbot như công cụ hỗ trợ thông tin, không thay thế bác sĩ; luôn kiểm chứng thông tin; bảo mật dữ liệu cá nhân; sử dụng trong phạm vi phù hợp và phải an toàn: tìm đến chuyên gia y tế khi có vấn đề sức khỏe nghiêm trọng [5].

ChatGPT được OpenAI ra mắt vào tháng 11 năm 2022 và sau đó là một loạt các chatbot khác ra đời như Copilot của Microsoft ra mắt năm 2023 như một trợ lý AI tích hợp trong hệ sinh thái Microsoft 365 và Windows, trong khi Gemini Pro của Google (cuối 2023), phát triển mạnh từ 2024–2025) với nhiều tính năng sáng tạo và nhiều chatbot phổ thông và chuyên về y tế khác đã trở nên phổ biến [6]. Tổ chức y tế Thế giới năm 2024 cũng đã đưa ra các chuẩn mực cho AI y tế về quản trị và đạo đức [7] cùng với tham khảo các công bố quốc tế gần đây giúp chúng tôi nền tảng khoa học để phát triển nghiên cứu này.

Trong khi chưa tìm thấy những công bố trong nước, đề tài nghiên cứu của chúng tôi có mục đích thăm dò khả năng cung cấp thông tin về sức khỏe của "hai trợ lý ảo" Copilot và Gemini trong hỗ trợ thày thuốc thực hiện giáo dục sức khỏe và tư vấn cho người bệnh. Từ đó đưa ra những định hướng trong tương lai để giảm thiểu thách thức và tối đa hóa lợi ích của công nghệ này trong giáo dục các ngành nghề y tế ở Việt Nam, đưa ra những gợi ý cho những nghiên cứu khám phá mức độ hiệu quả khi sử dụng Chatbot y tế và các công cụ AI chuyên dụng cho từng chuyên khoa khác trong y tế.

PHƯƠNG PHÁP NGHIÊN CỨU

Nghiên cứu phân tích cắt ngang (comparative cross-sectional study); So sánh kết quả các câu trả lời của 2 AI chatbot một bản với nền tảng phổ thông, không phải trả phí (Copilot) và một bản nâng cao (Gemini Pro). Mỗi câu hỏi sẽ có 2 câu trả lời của 2 chatbot được đánh giá độc lập bởi một thầy thuốc có trình độ bác sỹ và một điều dưỡng viên trình độ đại học (người đưa ra các câu hỏi mà người bệnh thường yêu cầu tư vấn), đúng chuyên khoa với những bệnh được "nhờ" AI chatbot tư vấn. Các câu hỏi và câu trả lời của chatbot được đưa vào một bảng, sau đó gửi cho bác sỹ và điều dưỡng chuyên khoa đánh giá theo thang điểm Likert gồm 5 mức:

- Mức 5 (Hoàn hảo): Đáp ứng tất cả tiêu chí; thông tin chính xác 100%; ngôn ngữ nhân văn; có cảnh báo an toàn rõ ràng.

- Mức 4 (Tốt): Chính xác về y khoa nhưng ngôn ngữ còn hơi cứng hoặc thiếu một vài chỉ dẫn thực tiễn nhỏ.

- Mức 3 (Trung bình): Thông tin đúng nhưng chung chung, chưa cá thể hóa được theo loại bệnh hoặc giai đoạn bệnh.

- Mức 2 (Kém): Có sai sót nhỏ về y khoa hoặc thiếu hoàn toàn các cảnh báo an toàn cần thiết.

- Mức 1 (Nguy hiểm): Đưa ra lời khuyên điều trị sai lệch, khuyến khích tự điều trị hoặc sử dụng các phương pháp chưa được kiểm chứng

Tổng số 245 câu hỏi đặt ra cho 2 AI chatbot do 6 điều dưỡng viên đưa ra (học viên cao học) của Trường Đại học Điều dưỡng Nam Định và Trường Đại học Thăng Long về các nhóm bệnh: ung thư, tại mũi họng, nội tiết, nhi khoa, gây mê hồi sức, sản khoa (trường hợp sinh thường), nội khoa. Các câu hỏi theo các chủ đề liên quan đến triệu chứng chẩn đoán (A), nguyên nhân (B), điều trị (C), chế độ ăn và sinh hoạt (D) và chăm sóc, tư vấn tuân thủ điều trị, dự phòng biến chứng, tác dụng phụ (E). Các câu trả lời được thu thập trong 2 tuần đầu tháng 1 năm 2026.

Phương pháp phân tích số liệu: Số liệu được xử lý bằng phần mềm SPSS-27. Các phép tính thống kê mô tả điểm trung bình và trung vị được. Để đánh giá độ tin cậy (đúng, đủ, không sai và an toàn) dựa vào số trung bình, trung vị điểm và tỷ lệ % đạt (câu trả lời >= 4 điểm). Thống kê phân tích phân tích: sử dụng test Wilcoxon, kiểm định sự khác biệt hai trung vị ở mức α =0,05. Để đánh giá sự phù hợp giữa hai người đánh giá cùng 1 câu hỏi , cùng một chatbot và giữa kết quả của một người đánh giá cùng 1 câu hỏi, với 2 câu trả lời của 2 chatbot, dựa vào chỉ số Kappa và tính nhất quán dựa trên test Kendall's tau b với bảng vuông 5 x 5 và test Kendall's tau c với bảng 2 x 5; tỷ lệ trùng khớp Ties (của test so sánh phi tham số) và hệ số tương quan Spearman roh.

Đạo đức nghiên cứu: Nghiên cứu được thực hiện theo các nguyên tắc đạo đức trong nghiên cứu y sinh học. Các bác sĩ và điều dưỡng tham gia đánh giá câu trả lời của chatbot đều được giải thích đầy đủ về mục tiêu nghiên cứu và tham gia trên tinh thần tự nguyện. Nghiên cứu không thu thập thông tin cá nhân hay dữ liệu bệnh án của người bệnh. Toàn bộ dữ liệu được mã hóa, bảo mật và chỉ sử dụng cho mục đích nghiên cứu khoa học.

KẾT QUẢ

Bảng 1. Phân bố bệnh, triệu chứng và nhóm câu hỏi được khảo sát

Triệu chứng hoặc bệnh	Nhóm câu hỏi					Cộng hàng
	Triệu chứng, chẩn đoán	Nguyên nhân	Điều trị	Ăn và sinh hoạt	Chăm sóc, tư vấn và dự phòng	
Rối loạn tiền đình	10	15	11	13	6	55 (22,4%)
Ung thư (đang xạ trị)	2	3	5	6	9	25 (10,2%)
Sản phụ sau sinh	16	2	3	7	22	50 (20,3%)
Nhi khoa (Bệnh hô hấp cấp)	15	1	17	2	22	57 (23,2%)
Đái tháo đường	24	3	9	9	14	59 (24,0%)
Cộng cột	67 (27,2%)	24 (9,8 %)	45 (18,3%)	37 (15%)	73 (29,7%)	246 (100%)

Kết quả bảng 1 cho thấy các nhóm câu hỏi phân bố không đều, câu hỏi về nguyên nhân dẫn đến triệu chứng/ bệnh và chế độ ăn uống, sinh hoạt chiếm tỷ lệ thấp hơn. Nhóm câu hỏi về tình trạng đang xạ trị cũng chiếm tỷ lệ thấp hơn các nhóm triệu chứng/ bệnh khác.

Bảng 2. Điểm trung bình, trung vị và tỷ lệ câu trả lời được đánh giá mức đạt

Chatbot và người đánh giá	**Chỉ số thống kê**		
Copilot – bác sỹ	Mean (95%CI)	3,9	(3,7- 4,0)
	Median (95%CI)	4,0	(3,9 – 4,0)
	% đạt mức >= 4	80,1%	(76,6-83,6)
Copilot – điều dưỡng	Mean (95%CI)	4,1	(4,0- 4,2)
	Median (95%CI)	4,0	(3,9 – 4,0)
	% đạt mức >= 4	93,5%	(93,4 -96,5)
Phù hợp giữa đánh giá nền tảng Copilot của bác sỹ và điều dưỡng	p (Wilcoxon)	< 0,05	-
	p (tỷ lệ %, sign test)	> 0,05	-
	Kappa	0,596	p < 0,001
	Kendall's tau b	0,713	p < 0,001
Gemini Pro – bác sỹ	Mean (95%CI)	4,7	(4,6 – 4,8)
	Median (95%CI)	5	-
	% đạt mức >= 4	97,6%	(95,7 – 99,5)
Gemini Pro – điều dưỡng	Mean (95%CI)	4,7	4,6 - 4,8
	Median (95%CI)	5	-
	% đạt mức >= 4	99,6%	(98,8 -100)
Phù hợp giữa đánh giá giá nền tảng Gemini của bác sỹ và điều dưỡng	p (Wilcoxon)	> 0,05	
	p (tỷ lệ %, sign test)	> 0,05	
	Độ phù hợp đối xứng -Kappa	0,83	p < 0,001
	Đọ nhất quán- Kendall's tau b	0,88	p < 0,001

Kết quả bảng 2 chỉ ra rằng tất cả trung vị điểm đều ≥ 4, trung bình đạt từ 3,9 đến 4,7 trong thang 5 điểm, Tỷ lệ các câu trả lời ở mức đạt đối với cả hai chatbot và đánh giá của các bác sỹ và điều dưỡng viên với tỷ lệ rất cao từ 8,1% do bác sỹ đánh giá Copilot, đến 99,6% do điều dưỡng đánh giá Gemini. Độ phù hợp - Kappa ở mức trung bình với Copilot và rất cao với Gemini (tương ứng 0,596 và 0,83). Độ nhất quán giữa đánh giá của bác sỹ và điều dưỡng (Kendall's tau b) đối với Gemini Pro cao hơn Copilot (0,88 so với 0,71). Kiểm định McNemar (p > 0,05) xác nhận không có sự khác biệt có hệ thống giữa đánh giá của bác sĩ và điều dưỡng, minh chứng cho tính nhất quán và độ tin cậy nội tại của mô hình.

Bảng 3. Sự phù hợp giữa hai nền tảng AI Chatbot: Copilot và Gemini Pro

Chatbot và người đánh giá	Chỉ số thống kê		
Phù hợp giữa hai chatbot về đánh giá của bác sỹ	Gemini < Copilot (negative)	0,9% (2)	-
	Gemini > Copilot (positive)	73,5% (181)	-
	Gemini =Copilot (Ties %)	25,6% (63)	-
	p (Wilcoxon)	<0,001	-
	Hệ số tương quan roh	0,64	p < 0,001
	Kappa	0,035	p > 0,05
	Kendall's tau b	0,61	p < 0,001
Phù hợp giữa hai chatbot về đánh giá của điều dưỡng	Gemini < Copilot	1,3% (3)	-
	Gemini > Copilot	60,9% (150)	-
	Gemini =Copilot	37,8% (93)	-
	p (Wilcoxon)	< 0,001	-
	Hệ số tương quan roh	0,43	p < 0,001
	Kappa	0,06	p > 0,5
	Kendall's tau b	0,24	p < 0,001
Độ phù hợp (khớp) giữa BS và ĐD của từng Chatbot	Ties % Gemini Pro	95,0%	
	Ties % Copilot	67,7%	

Bảng 3 cho thấy giữa hai nền tảng Gemini và Copilot điểm đánh giá khác nhau có ý nghĩa thống kê, Gemini được đánh giá cao hơn Copilot. Tương quan giữa điểm đánh giá của 2 chatbot ở mức trung bình (roh = 0,64 với bác sỹ đánh giá và 0,43 với điều dưỡng đánh giá), tương tự như thế độ phù hợp cũng với Gemini cao hơn Copilot (95% so với 67,7%).

Bảng 4. Mối liên quan giữa tỷ lệ % được đánh giá mức đạt (≥ 4 điểm) của hai Chatbot và các nhóm câu hỏi

AI Chatbot	Copilot			Gemini Pro		
Nhóm câu hỏi	**Tỷ lệ đạt (%)**	**Kendall's tau c**	**p**	**Tỷ lệ đạt (%)**	**Kendall's tau c**	**p**
Đánh giá của các bác sỹ						
Triệu chứng, chẩn đoán	82,1	0,037	> 0,05	97,0	0,042	> 0,05
Nguyên nhân	62,5			87,5		
Điều trị	77,8			97,8		
Chế độ ăn, sinh hoạt	86,5			100		
Chăm sóc, tư vấn, dự phòng	82,2			100		
So sánh kết quả giữa 2 chatbot của bác sỹ: p <0,001						

AI Chatbot		Copilot			Gemini Pro	
Nhóm câu hỏi	**Tỷ lệ đạt (%)**	**Kendall's tau c**	**p**	**Tỷ lệ đạt (%)**	**Kendall's tau c**	**p**
Đánh giá của các điều dưỡng						
Triệu chứng, chẩn đoán	97,0	0,03	> 0,05	98,5	0,012	> 0,05
Nguyên nhân	91,7			100		
Điều trị	88,9			100		
Chế độ ăn, sinh hoạt	97,3			100		
Chăm sóc, tư vấn, dự phòng	91,8			100		
So sánh kết quả giữa 2 chatbot của điều dưỡng: p >0,05						

Bảng 4 cho biết tỷ lệ các câu trả lời cho 5 nhóm câu hỏi ở mức đạt của cả 2 chatbot và 2 nhóm các bác sỹ và điều dưỡng viên có sự khác nhau nhưng chưa có ý nghĩa thống kê (p > 0,05), sự nhất quán khá thấp giữa các nhóm câu hỏi (Kendall's tau c thấp: 0,01 đến 0,04). Bác sỹ đánh giá Gemini tốt hơn Copilot một cách có ý nghĩa (p < 0,001) trong khi các điều dưỡng viên không thấy sự khác biệt (p > 0,05).

BÀN LUẬN

Kết quả nghiên cứu đã cung cấp bằng chứng định lượng thuyết phục về độ tin cậy của các trợ lý ảo y tế, với tất cả trung vị điểm đều ≥ 4, trung bình đạt từ 3,9 đến 4,7 trong thang 5 điểm. Tỷ lệ các câu trả lời ở mức đạt đối với cả hai chatbot và đánh giá của các bác sỹ và điều dưỡng viên với tỷ lệ rất cao từ 81% do bác sỹ đánh giá Copilot, đến 99,6% do điều dưỡng đánh giá Gemini. Có xu hướng các bác sỹ đánh giá thấp hơn so với điều dưỡng trên cùng nền tảng chatbot và với cùng nhóm nghiên cứu viên, Gemini luôn được đánh giá cao hơn. Độ phù hợp - Kappa ở mức trung bình với Copilot và rất cao với Gemini (tương ứng 0,596 và 0,83). Độ nhất quán giữa đánh giá của bác sỹ và điều dưỡng (Kendall's tau b) đối với Gemini Pro cao hơn Copilot (0,88 so với 0,71). Kết quả này khá tương đồng với một số báo cáo ở nước ngoài với độ phù hợp và độ nhất quán thấp hơn về các chỉ số đánh giá cho kết quả khác nhau tùy theo mức độ khó của câu hỏi, về các bệnh khác nhau [1,2,4,5] và nhất là các nghiên cứu này được thực hiện sớm hơn (năm 2023) khi các chatbot mới được phát triển và đang trong quá trình "học" so với thời điểm tháng 1 năm 2026 trong nghiên cứu này. Điều này cũng cho thấy việc nghiên cứu mức độ thông minh của từng chatbot sau những khoảng thời gian 1 đến 2 năm để biết thời điểm bão hòa kiến thức của chatbot thông thường. Mặt khác, khi tỷ lệ được đánh giá mức đạt dưới 100% trong nghiên cứu này là một lỗ hổng nhỏ nhưng cần khắc phục khi phát triển AI chatbot y tế riêng cho từng lĩnh vực chuyên sâu như tư vấn cho người bệnh về một bệnh cụ thể.

Với chức năng của bác sỹ là chẩn đoán bệnh, điều trị, tư vấn trong khi chức năng chính của điều dưỡng viên là chẩn đoán điều dưỡng và chăm sóc, tư vấn có thể là yếu tố ảnh hưởng đến cách nhận định độ tin cậy về cùng những thông tin cung cấp từ cùng một nền tảng chatbot. Kết quả nghiên cứu đã cho thấy sự đồng thuận ở mức trung bình ghi nhận ở Copilot qua chỉ số Kappa ở mức gần 0,6 (0,596) và p < 0,001, và ở hệ thống Gemini Pro với chỉ số Kappa đạt

0.83 và kiểm định McNemar (p > 0,05). Kết quả nghiên cứu cũng cho thấy không có sự khác biệt có hệ thống giữa đánh giá của bác sĩ và điều dưỡng (người đánh giá độ tin cậy và mức độ phù hợp của hai AI chatbot) minh chứng cho tính nhất quán và độ tin cậy nội tại của mô hình. Hơn nữa, nghiên cứu cũng cho thấy Gemini được đánh giá cao hơn Copilot, sự khác biệt có ý nghĩa thống kê. Tương quan giữa điểm đánh giá của 2 chatbot ở mức trung bình (roh = 0,64 với bác sỹ đánh giá và 0,43 với điều dưỡng đánh giá), và độ phù hợp với Gemini cao hơn Copilot (95% so với 67,7%). Điều này một lần nữa trong khuôn khổ của nghiên cứu này khẳng định vị thế tốt hơn của Gemini Pro so với Copilot (bản free). Với tỷ lệ đồng nhất quan điểm lên tới 93%, Gemini Pro đã chứng minh năng lực vượt trội trong việc cung cấp thông tin y khoa chính xác, đáp ứng đồng thời cả tiêu chuẩn học thuật của bác sĩ và tiêu chuẩn thực hành của điều dưỡng. Kết quả này khá phù hợp với một số nghiên cứu ở nước ngoài trước đây[3,6] giai đoạn 2022-2024.

Kết quả cũng chỉ ra rằng biết tỷ lệ các câu trả lời cho 5 nhóm câu hỏi ở mức đạt của cả 2 chatbot và 2 nhóm các bác sỹ và điều dưỡng viên có sự khác nhau nhưng chưa có ý nghĩa thống kê (p > 0,05), sự nhất quán khá thấp giữa các nhóm câu hỏi (Kendall's thấp: 0,01 đến 0,04). Bác sỹ đánh giá Gemini tốt hơn Copilotvới mức ý nghĩa thống kê (p < 0,001), trong khi các điều dưỡng viên không thấy sự khác biệt (p > 0,05). Qua 246 câu hỏi và 492 câu trả lời thu được từ hai AI chatbot với những phân tích định lượng như trình bày ở trên, chúng tôi xem xét một cách định tính về một số tiêu chí khác đối với một nền tảng AI y tế như: tính đầy đủ và phù hợp, tính an toàn, khả năng cung cấp thông tin dễ hiểu và thân thiện, chúng tôi nhận thấy câu trả lời dễ hiểu ngay cả với người không có chuyên môn y tế, không bỏ sót thông tin quan trọng; nội dung phù hợp với từng loại câu hỏi (triệu chứng, phòng ngừa, điều trị, chăm sóc); có khả năng cá nhân hóa theo bối cảnh người bệnh (tuổi, giới, bệnh nền). Về tính an toàn các thông tin từ hai chatbot cho thấy đều đưa ra lời khuyên không thay thế bác sĩ trong các tình huống khẩn cấp, có cảnh báo rõ ràng chỉ mang tính tham khảo và hướng dẫn người bệnh tìm đến cơ sở y tế khi cần thiết [7].

KẾT LUẬN

Hai AI chatbot Gemini Pro và Copilot có khả năng cung cấp thông tin về sức khỏe với độ tin cậy khá cao. Tỷ lệ các câu trả lời ở mức đạt đối với cả hai chatbot và đánh giá của các bác sỹ và điều dưỡng viên với tỷ lệ rất cao (từ 81% đến 99,6%). Bước đầu nhận thấy một số yếu tố liên quan: trên cùng một thông tin do chabot đưa ra, điều dưỡng viên có xu hướng điểm đánh giá cao hơn so với bác sỹ, Gemini Pro được đánh giá cao hơn Copilot. Độ phù hợp liên quan đến nền tảng AI (gemini tốt hơn copilot) và độ phù hợp giữa các bác sỹ cao hơn các điều dưỡng. Vì vậy, nghiên cứu đề xuất có thể sử dụng Gemini Pro và Copilot để hỗ trợ công tác tư vấn cho người bệnh, trong đó độ tin cậy đối với Gemini Pro tốt hơn. Ngoài ra, nghiên cứu này cần được mở rộng để đánh giá chi tiết hơn với các bệnh cụ thể cũng như nhận định về thời điểm bão hòa thông tin sau một thời gian khi các chatbot được huấn luyện nhiều hơn.

TÀI LIỆU THAM KHẢO

1. Colak D, Yakut B, Agin A. Comparison of the accuracy, comprehensiveness, and readability of ChatGPT, Google Gemini, and Microsoft Copilot on dry eye disease. Beyoglu Eye J. 2025;10(3):168-174. doi: 10.14744/bej.2025.76743.

2. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach. 2025 Jan;47(1):40-42. doi: 10.1080/0142159X.2024.2376879

3. Li D, Lutfi SL. Large language model–based virtual patient systems for history-taking in medical education: a comprehensive systematic review. JMIR Med Inform. 2026;14:e79039. Published January 2, 2026. doi:10.2196/79039

4. Ito S, Furukawa E, Okuhara T, Okada H, Kiuchi T. Leveraging artificial intelligence chatbots for anemia prevention: a comparative study of ChatGPT-3.5, Copilot, and Gemini outputs against Google Search results. PEC Innov.2025;6:100390. Published April 1, 2025. doi:10.1016/j.pecinn.2025.100390

5. Sabaner MC, Yozgat Z. Performance of ChatGPT-4 Omni and Gemini 1.5 Pro on ophthalmology-related questions in the Turkish Medical Specialty Exam. Turk J Ophthalmol. 2025 Aug 21;55(4):177-185. doi: 10.4274/tjo.galenos.2025.27895.

6. Urda-Cîmpean AE, Leucuta DC, Drugan C, Dutu AG, et al. Assessing the accuracy of diagnostic capabilities of large language models. Diagnostics (Basel). 2025 Jun 29;15(13):1657. doi: 10.3390/diagnostics15131657.

7. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva, Switzerland: World Health Organization; 2024.