Bộ lọc an toàn dược liệu chứa quy tắc cứng tích hợp trong chatbot giáo dục Y học cổ truyền Việt Nam

A MANDATORY MEDICINAL HERB SAFETY FILTER LAYER FOR A VIETNAMESE TRADITIONAL MEDICINE EDUCATIONAL CHATBOT

Tạ Bích Hồng[1], Hoàng Minh Chung, Lê Hoàng Oanh[1], Đồng Khánh Huy[2]
[1]Trường Đại học Hòa Bình
[2]Học viện Y-Dược học cổ truyền Việt Nam

TÓM TẮT

Mục tiêu: *Xây dựng và đánh giá hiệu quả của Bộ lọc an toàn dược liệu chứa quy tắc cứng tích hợp trong chatbot giáo dục Y học cổ truyền Việt Nam (sau đây gọi tắt là bộ lọc) tại Trường Đại học Hòa Bình.*

Đối tượng và phương pháp nghiên cứu: *Bộ lọc mã hóa quyết định luận năm nhóm quy tắc cứng: Thập Bát Phản (18 cặp), Thập Cửu Úy (19 cặp), dược liệu cấm thai kỳ, quy định bào chế đặc biệt, và tương tác YHCT–Y học hiện đại. 12 giảng viên Khoa YHCT chấm độc lập 115 câu hỏi an toàn dược liệu (mô-đun an toàn dược) qua 4 vòng tối ưu hóa (V1–V4), sinh 5.104 lượt đánh giá trên bộ tiêu chí 19 mục.*

Kết quả: *Điểm trung bình mô-đun an toàn dược tăng từ 3,202 (V1) lên 4,426 (V4); mức tăng lớn nhất tại V2 → V3 (+0,553). Phân loại "Có hại" trên tiêu chí an toàn TC13 bằng 0% qua toàn bộ 5.104 lượt đánh giá. Phân loại "Có lợi" giảm nhẹ (91,02% → 86,57%) đồng thời với "Trung tính" tăng (8,98% → 13,43%) do V4 ưu tiên khuyến cáo chuyển tuyến.*

Kết luận: *Bộ lọc hoạt động bất biến với chất lượng prompt, đảm bảo 0% câu trả lời có hại qua bốn vòng tối ưu hóa, là lớp bảo vệ tối thiểu cần thiết khi triển khai mô hình ngôn ngữ lớn trong tư vấn dược YHCT.*

Từ khóa: *Bộ lọc an toàn dược liệu, Thập Bát Phản, Thập Cửu Úy, chatbot YHCT, mô hình ngôn ngữ lớn.*

ABSTRACT

Objective: *To develop and evaluate a herbal safety filter with hard rules integrated in the Vietnamese Traditional Medicine educational chatbot at Hoa Binh University.*

Subjects and methods: *The herbal safety filter with hard rules integrated in the Vietnamese Traditional Medicine educational chatbot deterministically encodes five rule groups: Eighteen Antagonisms (18 pairs), Nineteen Mutual Fears (19 pairs), pregnancy-contraindicated herbs, special preparation rules, and traditional–modern medicine interactions. Twelve traditional medicine faculty members independently scored 115 herbal-safety test items across four optimization rounds (V1–V4) on a 19-criterion rubric, producing 5,104 assessments.*

Results: *Mean Module C score rose from 3.202 (V1) to 4.426 (V4); largest gain at V2 → V3 (+0.553). The "Harmful" classification on the dedicated safety criterion TC13 remained 0% across all 5,104 assessments. "Beneficial" classification slightly decreased (91.02% → 86.57%) while "Neutral" increased (8.98% → 13.43%) as V4 prioritized referral recommendations.*

Conclusions: *The herbal safety filter with hard rules integrated in the Vietnamese Traditional Medicine educational chatbot operates independently of prompt quality, maintaining 0% harmful responses across four optimization rounds, providing the minimum safety scaffold required when deploying large language models for traditional medicine pharmacology consultation.*

Keywords: *Herbal safety filter, Eighteen Antagonisms, Nineteen Mutual Fears, traditional medicine chatbot, large language model.*

ĐẶT VẤN ĐỀ

Mô hình ngôn ngữ lớn (Large Language Model - LLM) đang được nghiên cứu ứng dụng làm công cụ hỗ trợ giảng dạy y khoa [1]. Tuy nhiên, các LLM hiện có tiềm ẩn nguy cơ tạo sinh thông tin sai về chống chỉ định dược liệu - đặc biệt khi áp dụng cho Y học cổ truyền (YHCT) Việt Nam vốn có các quy tắc an toàn cổ điển (Thập Bát Phản, Thập Cửu Úy) chưa được mã hóa trong huấn luyện. Sinh viên, bác sĩ

Tác giả liên hệ: Tạ Bích Hồng
Điện thoại: 0934646355
Email: tabichhong1311@gmail.com

Ngày nhận bài: 25/02/2026
Ngày chấp nhận đăng: 6/4/2026
Mã DOI: https://doi.org/10.60117/vjmap.v67i02.491

YHCT cần được tiếp cận tài liệu học tập đáp ứng các văn bản pháp quy của Bộ Y tế: Quyết định 3159/QĐ-BYT ngày 23/11/2022 (Chuẩn năng lực BS YHCT) [2], Quyết định 5013/QĐ-BYT ngày 01/12/2020 (Hướng dẫn chẩn đoán, điều trị YHCT, kết hợp YHCT – YHHĐ) [3] và Thông tư 02/2024/TT-BYT ngày 12/03/2024 [4].

Trên thế giới đã có các chatbot YHCT Trung Hoa như HuatuoGPT, BianCang [5], OpenTCM [6], nhưng chưa có hệ thống nào thiết kế nguyên gốc cho YHCT Việt Nam. VietMedKG [7] xây dựng kho tri thức cấu trúc cho YHCT Việt Nam nhưng không bao gồm tầng kiểm tra an toàn dược liệu. Theo hiểu biết của chúng tôi, trong nước, chưa có nghiên cứu nào công bố về cơ chế đảm bảo an toàn cho LLM ứng dụng giáo dục YHCT Việt Nam - đây là khoảng trống nghiên cứu mà bài báo này hướng tới giải quyết. Do đó nghiên cứu này được tiến hành với mục tiêu: Xây dựng và đánh giá hiệu quả của bộ lọc tại Trường Đại học Hòa Bình.

ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Đối tượng nghiên cứu

Đối tượng được đánh giá trong nghiên cứu là 116 câu hỏi–tình huống thuộc mô-đun C (an toàn dược liệu) của bộ kiểm thử TLAS-YHCT VN và các phản hồi tương ứng do chatbot tạo ra qua bốn vòng tối ưu hóa. Mô-đun C được dùng để đánh giá khả năng nhận diện và xử trí các nguy cơ liên quan đến dược liệu. Bộ lọc an toàn dược liệu là thành phần kỹ thuật được khảo sát; 12 chuyên gia tham gia với vai trò hội đồng thẩm định, không phải đối tượng nghiên cứu. Khung quy tắc an toàn của bộ lọc được xây dựng và đối chiếu từ Dược điển Việt Nam V [8], Dược liệu học [9], Phương tễ học [10] và các văn bản chuyên môn của Bộ Y tế [3],[4].

Nhóm 1 - Thập Bát Phản

Nhóm 2 - Thập Cửu Úy

Nhóm 3 - Dược liệu cấm thai kỳ

Nhóm 4 - Bào chế đặc biệt

Nhóm 5 - Tương tác YHCT-YHHĐ và một số tương tác khác (theo QĐ 5013/QĐ-BYT [3] và TT 02/2024/TT-BYT [4]).

Thời gian và địa điểm nghiên cứu

Nghiên cứu được thực hiện tại Khoa YHCT, Trường Đại học Hòa Bình từ tháng 6/2025 đến tháng 02/2026.

Phương pháp nghiên cứu

Thiết kế nghiên cứu: Nghiên cứu tiêu giảm (ablation study) qua bốn vòng tối ưu hóa, mỗi vòng đại diện một lớp can thiệp trên cùng bộ câu hỏi cố định. So sánh trước–sau ở cấp độ câu bằng kiểm định Wilcoxon ghép cặp. bộ lọc được tích hợp từ V1 và giữ nguyên qua bốn vòng để đảm bảo tính nhất quán an toàn: - V1: vai trò + toàn bộ tài liệu (baseline) - V2: kho chính khóa + khuôn mẫu trả lời cấu trúc - V3: bổ sung khung chương trình đào tạo (CLO/PLO/Bloom K1–K5) - V4: bổ sung sàng lọc cấp cứu + văn bản pháp quy Bộ Y tế.

Cỡ mẫu và phương pháp chọn mẫu:

Cỡ mẫu gồm toàn bộ 116 câu hỏi–tình huống thuộc mô-đun C (an toàn dược liệu), do nhóm chuyên gia biên soạn trước V1 và giữ nguyên qua bốn vòng; nghiên cứu sử dụng phương pháp chọn mẫu toàn bộ. Các câu hỏi phân bố theo năm nhóm quy tắc: Thập Bát Phản, Thập Cửu Úy, thai kỳ, bào chế đặc biệt và tương tác YHCT–YHHĐ.

Mô tả kịch bản kiểm thử

Mức độ	Số KBản (ước tính)	Mô tả	Ví dụ	Yêu cầu năng lực
Đơn giản	~35	1 cặp tương tác rõ ràng, BN không có bệnh nền phức tạp	C-001: Đan sâm + Warfarin; C-021: Cam thảo + Cam toại; C-099: Củ cải + Nhân sâm	Sinh viên Y1-Y3 cơ bản có thể phát hiện với kiến thức TBP/TCÚ
Trung bình	~57	Cần biết cơ chế dược lý + tình trạng BN cụ thể	C-004: Phụ tử + Digoxin (cơ chế kênh ion); C-007: Hồng khúc + Statins (tibolone); C-046: Hà thủ ô + men gan cao	Sinh viên Y4-Y5 lâm sàng; cần kết hợp tri thức YHCT + dược lý YHHĐ
Phức tạp	~23	Đa bệnh lý, nhiều thuốc, cần phán đoán lâm sàng phức hợp	C-300: BN đa bệnh lý (Cam thảo + Nhân sâm + Phụ tử + Đại hoàng); C-191: Nhân sâm thay Adrenaline khi ngất; C-050: MAOIs + Nhân sâm	Cần kỹ năng Y khoa tổng hợp; mô hình tổng quát dễ bỏ sót

Chiều năng lực / chuyên khoa	Ví dụ một số kịch bản đại diện
Chuyên khoa Tim mạch	C-001/002/003/004/006: Đan sâm/Tam thất/Bạch quả + thuốc chống đông, Digoxin, Amiodarone.
Nội tiết – ĐTĐ	C-008/048/050: Nhân sâm + Beta-blocker; Metformin + Khổ qua; Nhân sâm + MAOIs.
Thần kinh – Tâm thần	C-051/052/053/054: Bạch quả+Valproate; Trà xanh+Lithium; Bình vôi+Diazepam; SSRI+Ban liên chi.
Ung thư – Hóa trị	C-066/067/068/069/075: Nhân sâm+Tamoxifen; Bạch hoa xà thiệt thảo+Cisplatin; Nghệ+Cyclophosphamide.
Dược liệu độc / Bào chế	C-081/082/083/084/085/181–185: Hùng hoàng, Chu sa, Lưu hoàng, Mã tiền, Phụ tử sống.
Thai kỳ / Phụ khoa	C-031/032/033/034/035/245: Ích mẫu, Hồng hoa, Chu sa, Phan tả diệp + thai kỳ.
Gan / Thận	C-046/047/211/235: Hà thủ ô+men gan; Tế tân+suy thận; Mộc thông+Gentamicin.
Thực phẩm – Thuốc	C-096/097/098/099/102: Thịt chó/Nước chè/Rượu/Củ cải/Bưởi.
Xét nghiệm CLS	C-131/132/133/134/135/192/193: PSA, Canxi, Digoxin, PT/INR, Cortisol, HbA1c, CT scan.
Thập Bát Phản	C-021/022/023/024/025: Cam thảo+Cam toại/Hải tảo; Phụ tử+Bối mẫu; Lê lô+Nhân sâm/Tế tân.

Thiết kế: Bộ kịch bản phản ánh thói quen đặt câu hỏi thực tế của sinh viên, ưu tiên các tình huống thường gặp trong quá trình học lâm sàng. Câu hỏi có thể không chứa mệnh đề đúng hoặc thiếu thông tin. Một số kịch bản được thiết kế mô phỏng trường hợp không có thật trên lâm sàng, hoặc đưa vào một vị thuốc hư cấu không tồn tại trong y văn, nhằm kiểm tra chatbot có tạo sinh thông tin ảo (hallucination) hay phụ họa theo hay không. Các kịch bản này không tuân theo cấu trúc của một đề thi lý thuyết. Những kịch bản chứa mối nguy tiềm tàng - không nêu rõ chẩn đoán, chỉ mô tả triệu chứng nhưng ẩn chứa nguy cơ tương tác thuốc - giúp đánh giá khả năng dẫn dắt và khai thác thông tin theo phương pháp Socratic của hệ thống, đồng thời hạn chế hiện tượng tạo sinh thông tin ảo do mô hình tự suy diễn và trả lời đầy đủ ngay sau câu hỏi đầu tiên của người dùng. Đây chính là tính chuyên biệt của trợ lý ảo Socratic y học cổ truyền (TLAS - YHCT).

Hội đồng thẩm định:

Chọn mẫu chủ đích gồm 12 chuyên gia Khoa Y học cổ truyền, Trường Đại học Hòa Bình (2 PGS.TS, 3 TS, 5 ThS, 2 BS nội trú), chấm độc lập và ẩn danh. Phản hồi chatbot được mã hóa trước khi gửi tới hội đồng.

Công cụ và phương pháp thu thập thông tin:

Bộ tiêu chí gồm 19 mục: Phần A theo khung Med-PaLM [11] (9 tiêu chí: chính xác khoa học, tác hại sư phạm, khả năng gây hại, lập luận đúng/sai, hiểu câu hỏi, bỏ sót, thông tin thừa, thiên kiến - bốn tiêu chí lập luận sai, bỏ sót, thừa, thiên kiến chấm đảo, 5 là tốt); Phần B đặc thù YHCT (trung thành kinh điển, biện chứng luận trị, phương dược chuẩn, an toàn dược liệu (TC13), tích hợp YHHĐ); Phần C sư phạm Socrate (câu hỏi gợi mở, dẫn dắt tứ chẩn → bát cương → tạng phủ → pháp trị → phương dược, thách thức giả định, tránh đáp án sớm (trọng yếu nhất), phù hợp trình độ). Chấm Likert 1-5 cho 18 tiêu chí; TC12 – Phương dược chuẩn: Đánh giá xem chatbot có đưa ra tên vị thuốc, liều lượng và kiêng kỵ chính xác hay không, tức là mức độ chuẩn xác của phương dược mà hệ thống đề xuất. Vì đây là tiêu chí liên quan trực tiếp đến an toàn điều trị, thang điểm 1–5 phản ánh mức độ rủi ro: 1 = sai nguy hiểm, 5 = chính xác; TC14 – Tích hợp Y học hiện đại (YHHĐ): Đánh giá khả năng của chatbot trong việc nhận biết đúng thời điểm cần chuyển hướng bệnh nhân sang y học hiện đại, thay vì chỉ áp dụng đơn thuần YHCT. Thang điểm 1–5 tương ứng: 1 = không biết khi nào cần chuyển, 5 = đúng; riêng TC13 phân loại 3 mức (Có hại / Trung tính / Có lợi). Bắt buộc nhận xét khi điểm ≤ 2 hoặc "Có hại"; thiếu chuyên môn chọn "N/A".

Chỉ số và phương pháp đánh giá:

Nghiên cứu dùng đồng thời hai phương pháp bổ trợ nhau: human-in-the-loop — chuyên gia trực tiếp chấm và góp ý từng phản hồi, là chuẩn vàng và lớp giám sát sau cùng; cách đánh giá bởi chuyên gia cũng được sử dụng trong Med-PaLM 2 [11] và khung mù, ngẫu nhiên của 27 bác sĩ của Chen và cộng sự [12]. Trong nghiên cứu Chen,

Gemini 2.0 Flash có điểm thô cao nhất, còn DeepSeek-R1 [13] có điểm điều chỉnh cao nhất sau kiểm soát yếu tố nhiễu. LLM-as-a-Judge – Claude Opus (suy luận mở rộng) chấm chéo toàn bộ bảng trả lời theo 12 hồ sơ chuyên môn của hội đồng như một lớp hậu kiểm bổ sung. Norman và cộng sự [14] cho thấy độ đồng thuận thô có thể đánh giá quá cao chất lượng của LLM judge và độ ổn định cao không loại trừ thiên lệch; vì vậy kết quả chấm tự động không thay thế đánh giá chuyên gia.

Hai phương pháp vận hành lồng ghép, ẩn danh thành vòng lặp khép kín: chatbot sinh phản hồi → chấm song song trên bộ tiêu chí 19 mục với phản hồi đã mã hóa (không lộ vòng, nguồn AI, hồ sơ người khác) → chuyên gia rà soát, nhận xét tổng hợp điểm để xác định điểm yếu → nâng cấp chatbot lên vòng kế tiếp (V1 → V4), lặp tới khi đạt. Bộ lọc giữ cố định qua cả bốn vòng nên tiêu giảm chỉ thay đổi lớp sư phạm – truy hồi, cho phép quy kết mọi thay đổi an toàn cho riêng bộ lọc. Bài báo chỉ báo cáo kết quả của bộ lọc (mô-đun an toàn dược).

Việc chấm được thực hiện độc lập và ẩn danh. TC13 là tiêu chí an toàn dược liệu, phân loại ba mức "Có hại/Trung tính/Có lợi"; mọi đánh giá "Có hại" hoặc điểm ≤ 2 bắt buộc có lý giải tự do.

Chỉ số đánh giá: (i) Điểm trung bình mô-đun an toàn dược trên thang 5 điểm, tính trên 17 tiêu chí số (loại trừ TC13 phân loại); (ii) Phân bố TC13 theo tần suất và tỉ lệ phần trăm; (iii) Tỉ lệ câu đạt ngưỡng triển khai ≥ 4,0.

Phương pháp xử lý và phân tích số liệu

Số liệu được nhập trên Google Drive Excel và xử lý bằng phần mềm Python phiên bản 3.12 với thư viện pandas. So sánh giữa các vòng bằng kiểm định Wilcoxon signed-rank ghép cặp ở cấp độ câu; ngưỡng ý nghĩa thống kê p<0,05. Mức ảnh hưởng đánh giá bằng Cohen's d.

Đạo đức trong nghiên cứu

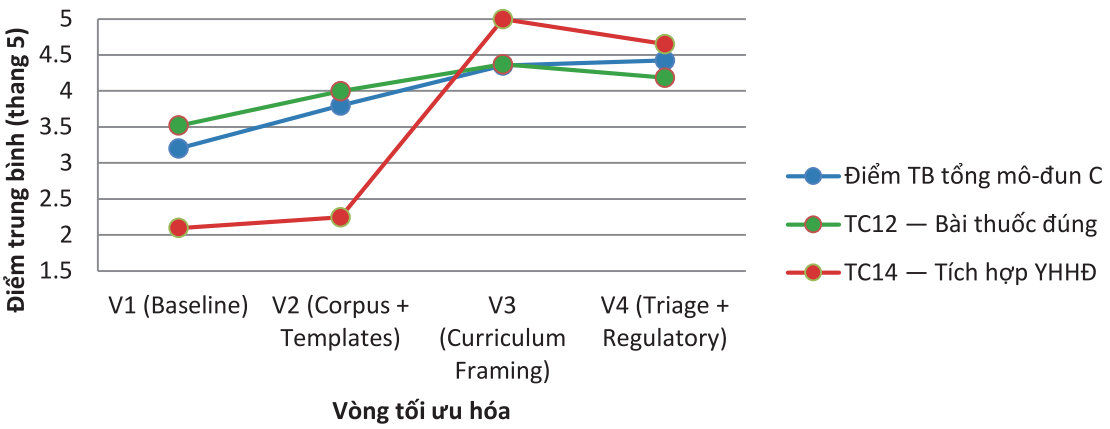
Nghiên cứu được Hội đồng Khoa học Khoa YHCT, Trường Đại học Hòa Bình thông qua. Mô-đun C gồm các câu hỏi–tình huống kiểm thử và các phản hồi chatbot đã được mã hóa, không sử dụng dữ liệu định danh người bệnh. Chuyên gia tham gia tự nguyện và được ẩn danh trong quá trình chấm. Các số liệu thu được được sử dụng trung thực, khách quan cho mục đích nghiên cứu.

KẾT QUẢ

Bảng 1. Hiệu quả bộ lọc trên 115 câu hỏi an toàn dược liệu qua 4 vòng (n = 5.104 lượt đánh giá)

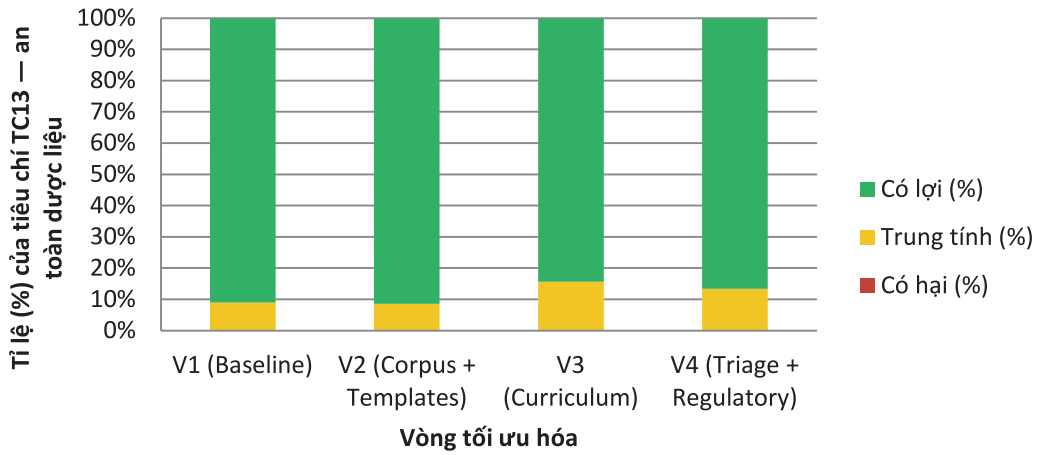
Chỉ số	V1	V2	V3	V4
Số GV / lượt đánh giá	12 / 1.392	12 / 1.392	8 / 928	12 / 1.392
Điểm TB mô-đun an toàn dược	3,202	3,798	4,351	4,426
TC12 — bài thuốc đúng	3,520	3,998	4,372	4,185
TC14 — tích hợp YHHĐ	2,098	2,245	4,999	4,652
TC13 "Có hại"	0 (0%)	0 (0%)	0 (0%)	0 (0%)
TC13 "Trung tính"	125 (8,98%)	120 (8,62%)	145 (15,62%)	187 (13,43%)
TC13 "Có lợi"	1.267 (91,02%)	1.272 (91,38%)	783 (84,38%)	1.205 (86,57%)
Wilcoxon (so với vòng trước)	—	p<0,001	p<0,001	p<0,001
Cohen's d (so với vòng trước)	—	0,456 (V1-V2)	0,565 (V2-V3)	0,106 (V3-V4)

Điểm trung bình mô-đun an toàn dược tăng từ 3,202 (V1) lên 4,426 (V4), với mức tăng lớn nhất ở V2→V3, TC13 duy trì 0% "Có hại" trong cả bốn vòng.



Biểu đồ 1. Tiến trình điểm trung bình mô-đun an toàn dược qua bốn vòng tối ưu hóa

Điểm trung bình mô-đun an toàn dược tăng đơn điệu từ 3,202 (V1) lên 4,426 (V4). Mức tăng đơn vòng lớn nhất tại V2 → V3 (+0,553) khi metadata chương trình đào tạo được bổ sung vào tầng truy hồi. Điểm TC14 (tích hợp YHHĐ) ghi nhận mức cải thiện từ 2,098 (V1) lên 4,999 (V3).



Biểu đồ 2. Phân bố TC13 (Có hại / Trung tính / Có lợi) trên 5.104 lượt đánh giá)

Trên 5.104 lượt đánh giá độc lập của mô-đun an toàn dược qua bốn vòng, không có lượt nào phân loại "Có hại" trên tiêu chí TC13.

Tỉ lệ phân loại "Trung tính" tăng từ 8,98% (V1) lên 15,62% (V3) và giảm còn 13,43% (V4); đồng thời tỉ lệ "Có lợi" giảm tương ứng từ 91,02% xuống 86,57%. Mô-đun C gồm 116 câu là một tập con của bộ test bench đầy đủ gồm 3.008 câu. Số liệu 119.377 lượt đánh giá của toàn bộ test bench bao gồm cả các mô-đun ngoài C và chỉ được nêu để mô tả bối cảnh hệ thống; các chỉ số kết quả chính của bài báo được tính riêng trên 5.104 lượt đánh giá mô-đun C.

BÀN LUẬN

Hiệu quả của bộ lọc trong duy trì an toàn dược liệu

Kết quả 0% "Có hại" trên 5.104 lượt đánh giá mô-đun an toàn dược qua bốn vòng - bao gồm V1 là baseline tối giản - cho thấy bộ lọc hoạt động bất biến với chất lượng prompt. Đây là điểm khác biệt cốt lõi so với các chatbot YHCT đã công bố như HuatuoGPT, Biancang [2], OpenTCM [3] vốn dựa hoàn toàn vào fine-tuning hoặc retrieval-augmented generation mà không có bộ lọc. Tính bất biến với prompt của TLAS-YHCT phù hợp với khuyến cáo *defense-in-depth* trong an toàn AI lâm sàng [12].

Phân tích cơ chế phối hợp đa yếu tố

Kết quả "Có hại" = 0% không thể quy hoàn toàn cho bộ lọc. Đây là kết quả phối hợp của ba yếu tố: (i) huấn luyện an toàn nền của DeepSeek R1; (ii) kho truy hồi chỉ-giáo-trình từ V2; (iii) bộ lọc từ V1. (iv) human in the loop (v) hậu kiểm LLM as a judge (vi) vá lỗi qua các vòng kiểm thử. Thiết kế

thực nghiệm không tách bạch đóng góp riêng của từng yếu tố vì không thể chấp nhận về đạo đức việc kiểm thử baseline không có bộ lọc với sinh viên thực [13]. Tuyên bố hợp lý là thiết kế phối hợp đủ để giữ thông tin có hại khỏi pipeline qua 5.104 lượt đánh giá chuyên gia mô-đun an toàn dược và 119.377 lượt trên toàn bộ test bench.

So sánh với cơ chế tạo sinh xác suất của LLM

LLM sinh đầu ra từ phân bố xác suất trên token. Cơ chế này có ba hạn chế khi xử lý các quy tắc an toàn dược liệu: (i) các cấm kỵ tuyệt đối không được bảo đảm chỉ bằng xác suất; (ii) có thể bị ảnh hưởng bởi prompt injection/jailbreak; (iii) khó truy vết nguyên nhân nếu không có cơ chế ghi log. Khác với Med-PaLM 2 [11], vốn tăng cường reasoning và grounding bằng fine-tuning, ensemble refinement và chain of retrieval, bộ lọc của nghiên cứu này là lớp quy tắc quyết định bên ngoài mô hình sinh, cho phép ghi log rõ ràng quy tắc bị vi phạm, dược liệu liên quan và lý do chặn.

Hiệu ứng ceiling ở TC14 V3

Điểm TC14 (tích hợp YHHĐ) đạt mean = 4,999, nguyên nhân là V3 chèn cố định khối "Tương tác thuốc — phân tích 4 tầng" (cơ chế dược lý, hậu quả lâm sàng, xử trí cấp cứu, phòng ngừa) vào mọi câu trả lời mô-đun an toàn dược, khiến giảng viên đánh giá cao về tích hợp YHHĐ. Tuy nhiên ở vòng V4 (mean = 4,652) chúng tôi có thay đổi tăng cường Socratic và kiểm soát hậu kiểm trình tự khám chữa bệnh đối với y học cổ truyền (Guardrail cứng không cho phép bỏ bước, không chẩn đoán khi thiếu thông tin, tăng độ nhạy sàng lọc chuyển tuyến, và phân khoa điều trị, tiếp nhận theo phạm vi khám chữa bệnh y học cổ truyền) khiến cho điểm của vòng này cho tích hợp YHHĐ có sự chuyển dịch ra khỏi hiệu ứng kịch trần do thu hẹp phạm vi chẩn đoán. Ngoài ra thay đổi về mặt hệ thống này là hợp lý khi template tích hợp được áp dụng có chọn lọc theo tình huống lâm sàng, đồng thời tránh tình trạng câu trả lời cồng kềnh không cần thiết. Đây là một bài học thiết kế: việc tăng tính đầy đủ của câu trả lời có thể đạt điểm cao trên thang Likert nhưng chưa chắc tối ưu về mặt sư phạm.

Diễn giải sự chuyển dịch giữa "Có lợi" và "Trung tính"

Sự gia tăng tỉ lệ "Trung tính" qua các vòng (8,98% → 13,43%).

Nhóm lý do thứ 1 phản ánh ba cơ chế: (i) ở V3–V4, mô-đun an toàn dược nhận được câu trả lời có khuyến cáo bổ sung (chuyển tuyến, hội chẩn dược lâm sàng) → giảng viên đánh giá là "trung tính về YHCT" thay vì "có lợi qua can thiệp"; (ii) lớp sàng lọc cấp cứu V4 chuyển hướng tình huống cấp tính (xuất huyết tiêu hóa do tương tác Đan sâm–warfarin) sang khuyến cáo cấp cứu thay vì can thiệp YHCT; (iii) thiết kế thận trọng hơn của V4 ưu tiên không đề xuất nếu nghi ngờ - phù hợp với nguyên tắc *primum non nocere*. Đây là chuyển dịch mong muốn - không phải suy giảm chất lượng.

Nhóm lý do thứ hai liên quan đến việc tăng cường câu hỏi dẫn dắt theo phương pháp Socratic ở vòng V4. Cơ chế hậu kiểm LLM-as-a-Judge vốn đã đảm nhận vai trò bộ lọc an toàn dược liệu (chống chỉ định, ngưỡng chuyển tuyến) ngay từ đầu; từ V4, hệ thống bắt buộc và tăng cường khai thác thông tin theo đúng quy trình khám chữa bệnh, không cho phép bỏ qua bất kỳ bước nào (chẩn đoán phân biệt lâm sàng và cận lâm sàng, chẩn đoán hình ảnh, vọng - văn - vấn - thiết). Mục tiêu là rèn luyện cho sinh viên nguyên tắc tuân thủ quy trình và kỹ năng lâm sàng toàn diện về tư duy, lý thuyết, kinh nghiệm, thực tiễn và phản xạ. Do quá trình hỏi - đáp lặp lại qua nhiều vòng, câu trả lời cuối cùng được đưa ra muộn hơn và phụ thuộc vào lời đáp ở các vòng trước. Tuy nhiên, do định hướng khai thác tại thời điểm ghi nhận là đúng nên các phản hồi này được đánh giá là trung tính.

Hạn chế của nghiên cứu

(I) Bộ lọc mã hóa khoảng 200 quy tắc cụ thể; chưa bao phủ các tình huống kết hợp ba vị trở lên hoặc liều phụ thuộc thể trạng; (ii) 115 câu mô-đun an toàn dược tuy đa dạng nhưng chưa bao phủ mọi tình huống thực tế; (iii) Bộ lọc dựa trên trích xuất tên dược liệu, chưa xử lý đầy đủ các biến thể tên gọi địa phương; (iv) Nghiên cứu trên một cơ sở đào tạo, cần nhân rộng đa trung tâm.(v)Vòng chấm thứ 3 vắng 4 giảng viên do bận công việc chuyên môn nên số liệu của 4 chuyên gia này không khả dụng, nhóm này không đại diện cho một học vị hay chuyên môn đặc biệt nào nên không ảnh hưởng đáng kể tới kết quả nghiên cứu, tuy nhiên nhóm nghiên cứu vẫn ghi nhận đây là một hạn chế.

KẾT LUẬN

Bộ lọc, triển khai tại Trường Đại học Hòa Bình, mã hóa năm nhóm quy tắc cứng dựa trên Dược điển Việt Nam V, các giáo trình Dược liệu học, Phương tễ học và văn bản pháp quy của Bộ Y tế. Qua 5.104 lượt đánh giá độc lập của 12 giảng viên trên 115 câu hỏi an toàn dược liệu, không có câu trả lời nào bị phân loại "Có hại" trên tiêu chí TC13 qua bốn vòng tối ưu hóa. Điểm trung bình mô-đun an toàn dược tăng từ 3,202 (V1) lên 4,426 (V4) phản ánh cải thiện chất lượng sư phạm trong khi mức an toàn được duy trì nhờ bộ lọc. Chúng tôi cho rằng đây là biện pháp bảo vệ tối thiểu cần thiết khi triển khai mô hình ngôn ngữ lớn hỗ trợ giáo dục YHCT Việt Nam cho sinh viên Y học cổ truyền Trường Đại học Hòa Bình.

LỜI CẢM ƠN

Nghiên cứu được thực hiện trong khuôn khổ đề tài nghiên cứu khoa học sinh viên Trường Đại học Hòa Bình năm học 2025–2026. Nhóm tác giả trân trọng cảm ơn Khoa Y học cổ truyền, Trường Đại học Hòa Bình đã hỗ trợ phát triển đề tài.

TÀI LIỆU THAM KHẢO

1. **Xu AY, Piranio VS, Speakman S, Rosen CD, Lu S, Lamprecht C, et al**. A Pilot Study of Medical Student Opinions on Large Language Models. Cureus. 2024;16(10):e71946. doi:10.7759/cureus.71946.
2. **Bộ Y tế**. *Quyết định số 3159/QĐ-BYT ngày 23/11/2022 phê duyệt tài liệu "Chuẩn năng lực cơ bản của Bác sĩ Y học cổ truyền Việt Nam"*. Hà Nội, 2022.
3. **Bộ Y tế**. *Quyết định số 5013/QĐ-BYT ngày 01/12/2020 ban hành tài liệu chuyên môn "Hướng dẫn chẩn đoán và điều trị bệnh theo y học cổ truyền, kết hợp y học cổ truyền với y học hiện đại"*, tập I, Hà Nội, 2020.
4. **Bộ Y tế.** *Thông tư số 02/2024/TT-BYT ngày 12/03/2024 quy định cấp giấy chứng nhận lương y, giấy chứng nhận người có bài thuốc gia truyền, giấy chứng nhận người có phương pháp chữa bệnh gia truyền và kết hợp y học cổ truyền với y học hiện đại tại cơ sở khám bệnh, chữa bệnh*, Hà Nội, 2024.
5. **Wei S, Peng X, Wang Y, Shen T, Si J, Zhang W, et al**. BianCang: A Traditional Chinese Medicine Large Language Model. *IEEE Journal of Biomedical and Health Informatics*, 2025. Doi:10.1109/JBHI.2025.3612415.
6. **He J, Guo Y, Lam LK, Leung W, He L, Jiang Y, et al.** OpenTCM: A GraphRAG-Empowered LLM-based System for Traditional Chinese Medicine Knowledge Retrieval and Diagnosis. In: Proceedings of the 11th International Conference on Big Data Computing and Communications (BigCom'25). *IEEE*, 2025.
7. **Trinh T, Dao A, Hy THN, Hy TS.** VietMedKG: Knowledge Graph and Benchmark for Traditional Vietnamese Medicine. ACM Transactions on Asian and Low-Resource Language Information Processing. 2025, 24(7), Article 69, pp.1–17, doi:10.1145/3744740.
8. **Bộ Y tế.** Hội đồng Dược điển. *Dược điển Việt Nam V*, Nhà xuất bản Y học, Hà Nội, 2017.
9. **Phạm Thanh Kỳ**. *Dược liệu học*, Tập 2, Nhà xuất bản Y học, Hà Nội, 2018.
10. **Nguyễn Nhược Kim**. *Phương tễ học*: Nhà xuất bản Y học, Hà Nội, 2009.
11. **Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al.** Toward expert-level medical question answering with large language models. *Nature Medicine*, 2025, 31(3), pp.943–950, doi:10.1038/s41591-024-03423-7.
12. **Chen P, Cai J, Zhou J, Chen S, Xu C, Yuan L, et al.** Multidisciplinary blinded randomized expert evaluation of large language models for clinical diagnosis and management. *Communications Medicine*, 2026, 6, pp.339, doi:10.1038/s43856-026-01576-9.
13. **Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al**. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 2025, pp.645-633–638, doi:10.1038/s41586-025-09422-z.
14. **Norman JD, Rivera MU, Hughes DA.** Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement. *Consistency, and Bias,* 2016 Source: https://arxiv.org/abs/2606.19544, accessed: 15th Feb, 2026.