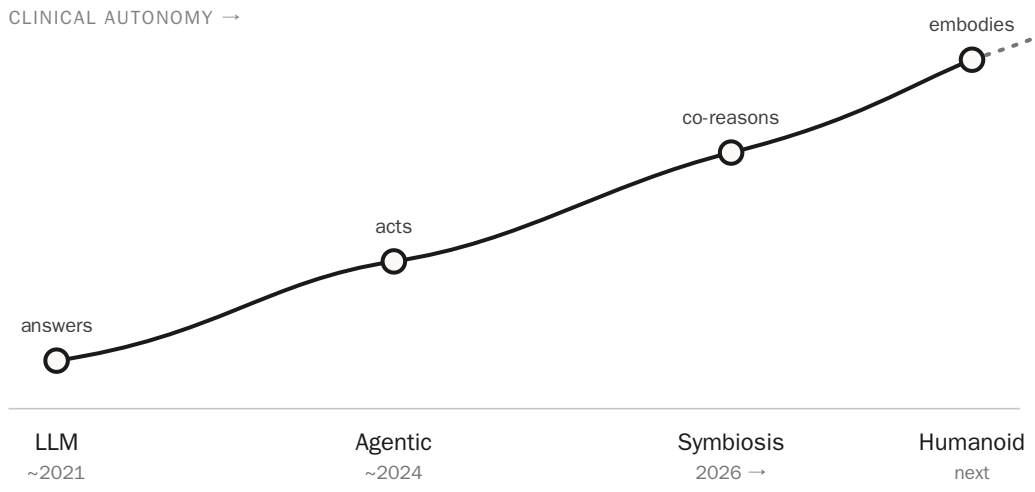
Y học sau Mô hình: Từ Ngôn ngữ đến Cộng sinh

1 tháng 7, 2026 · Nikon Sugar

> Giữa năm 2021 và 2026, trí tuệ nhân tạo trong y học đã vượt qua ba ngưỡng cửa nối tiếp nhau chóng vánh: các mô hình ngôn ngữ có thể trả lời, các tác nhân có thể hành động, và những hệ thống đầu tiên được xây dựng để suy luận song hành cùng người thầy thuốc thay vì phục tùng dưới quyền họ. Bài luận này lần theo sự tăng tốc đó, trình bày phần toán học đã làm nên điều ấy, và lập luận rằng đích đến không phải là tự động hóa hay thay thế mà là cộng sinh — với chăm sóc hiện thân, dạng hình người, là biên giới kế tiếp. Câu hỏi cho thế hệ bác sĩ tương lai không phải là liệu cỗ máy đã đủ giỏi hay chưa, mà là người thầy thuốc tồn tại để làm gì một khi nó đã đủ giỏi.

Trong phần lớn lịch sử của mình, y học điện toán là một câu chuyện về sự hạn hẹp. Một mô hình có thể đọc phim chụp nhũ ảnh (1) hay phân độ một tổn thương da (2) ở trình độ của một chuyên gia, và mỗi hệ thống như thế là một tượng đài cho một tác vụ duy nhất — được huấn luyện với chi phí khổng lồ, triển khai tại một phòng khám duy nhất, mù tịt trước mọi thứ nằm ngoài khung nhìn của nó. Lời hứa hẹn là có thật và trần giới hạn thì thấp. Điều thay đổi, bắt đầu vào khoảng năm 2021, không phải là các mô hình trở nên khá hơn đôi chút. Mà là chúng thôi không còn hạn hẹp nữa.

Bài luận này nói về sự thay đổi ấy và nơi nó dẫn tới. Bản tóm tắt ngắn gọn của lập luận được thâu tóm trong Hình 1: trong năm năm, AI y khoa đã leo từ những hệ thống *trả lời*, qua những hệ thống *hành động*, hướng tới những hệ thống *đồng suy luận* với người thầy thuốc — và độ dốc không hề thoai thoải lại ở chỗ mực vẽ trên sơ đồ cạn đi.

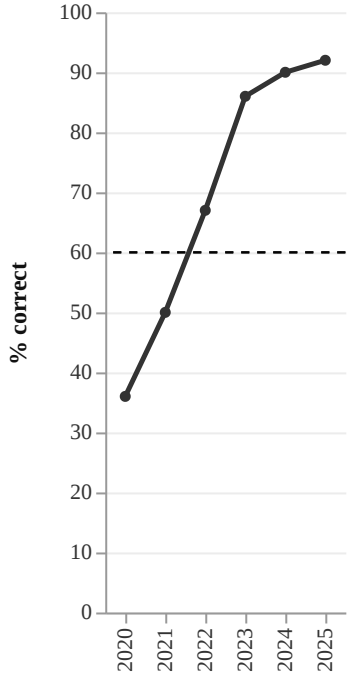


Hình 1. Bốn kỷ nguyên của AI y khoa trên một đường cong đi lên của mức độ tự chủ lâm sàng: các mô hình ngôn ngữ trả lời, các hệ thống tác nhân hành động, sự cộng sinh người–AI đồng suy luận, và chăm sóc hiện thân dạng hình người.

Bước nhảy năm năm

Thật dễ quên điều này đã diễn ra nhanh đến nhường nào. Muộn nhất là năm 2020, một mô hình lớn khi thử làm Kỳ thi Cấp phép Hành nghề Y khoa Hoa Kỳ (USMLE) chỉ đạt điểm ở ngưỡng ba mươi mấy — dưới mức may rủi đối với một sinh viên có động lực, còn xa mới chạm mức đậu ~60%. Đến cuối năm 2022, một mô hình được tinh chỉnh cho y khoa lần đầu tiên vượt qua ngưỡng ấy (3); trong vòng một năm, các hệ thống đa dụng đã đạt điểm ở ngưỡng tám mươi mấy cao trên cùng những ngân hàng câu hỏi đó (4), và đường cong từ bấy đến nay vẫn miệt mài tiến về trần giới hạn của nó.

Model accuracy on USMLE-style questions (MedQA), with the ~60% pass mark



Vì sao lại nhanh đến thế? Câu trả lời trung thực là y học được hưởng phần lan tỏa từ một khám phá mang tính tổng quát hơn. Kiến trúc nằm bên dưới tất cả những điều này — transformer — đã thay thế cơ chế hồi quy bằng một phép toán duy nhất, cơ chế chú ý (attention), cho phép mỗi token cân nhắc trực tiếp mọi token khác (5):

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

Cái mà Phương trình 1 mua được không phải là tri thức y khoa mà là *khả năng mở rộng quy mô* — một mô hình mà chất lượng cải thiện một cách có thể dự đoán được khi ta thêm tham số và dữ liệu. Bản thân tính dự đoán được ấy cũng đã được đo lường: trải qua nhiều bậc độ lớn, độ mất mát (loss) giảm theo một hàm lũy thừa của kích thước mô hình,

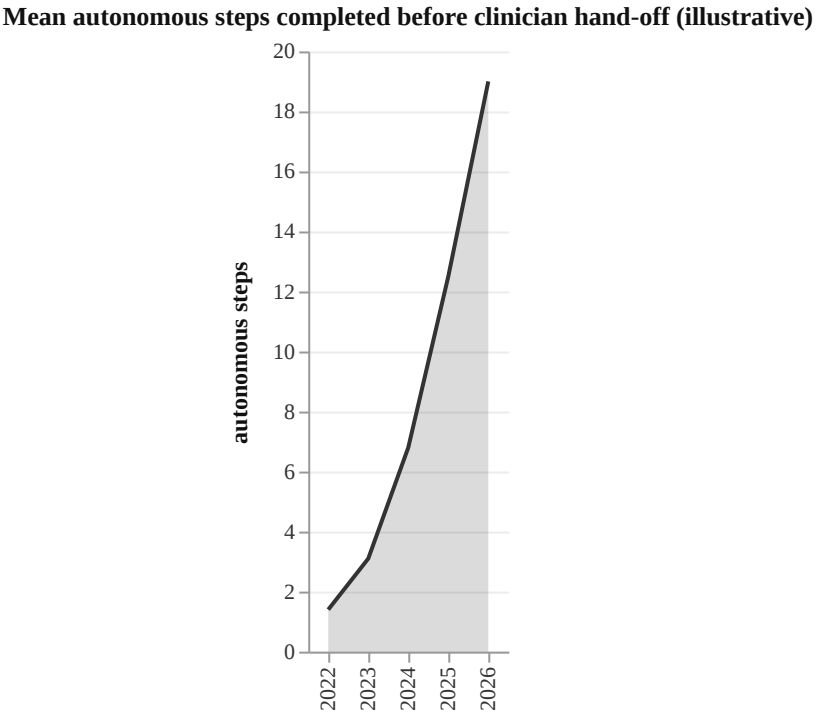
$$L(N) \approx \left(\frac{N_c}{N}\right)^{\alpha_N} \tag{2}$$

với một số mũ nhỏ α_N (6). Phương trình 2 chính là cỗ máy thầm lặng đứng sau đường dốc phía trên: y học không cần một đột phá riêng biệt mỗi năm, chỉ cần một phần chia sẻ của đường cong mà cả lĩnh vực đang cùng leo lên. Một mô hình được huấn luyện trên kho tàng mở của ngôn ngữ con người đã, gần như tình cờ, đọc đủ y học để suy luận về nó — và một nền tảng đa dụng, được tinh chỉnh nhẹ nhàng, bắt đầu vượt trội hơn chính những chuyên gia hạn hẹp mà lẽ ra nó chỉ được kỳ vọng hỗ trợ (7).

Từ câu trả lời đến hành động

Một kỳ thi tưởng thưởng cho hệ thống *biết*. Một phòng khám tưởng thưởng cho hệ thống *làm* — thứ chỉ định xét nghiệm, đối chiếu danh mục thuốc, soạn giấy giới thiệu, cảnh báo tương tác, và theo dõi kết quả ba ngày sau đó. Làn sóng thứ hai của AI y khoa, đại khái từ năm 2023 trở đi, là bước chuyển từ một mô hình mà bạn truy vấn sang một **tác nhân** (agent) theo đuổi một mục tiêu qua nhiều bước, gọi các công cụ và đọc đầu ra của chúng khi tiến hành.

Năng lực định nghĩa cho kỷ nguyên tác nhân không phải là sự lưu loát mà là *độ dài tác vụ*: một hệ thống có thể gánh vác bao nhiêu phần của một quy trình công việc thực trước khi phải trao lại cho con người. Con số đó đã tăng lên mạnh mẽ.



Đây cũng là chỗ mà mức độ rủi ro thay đổi. Một mô hình trả lời sai một câu hỏi thì tạo ra một câu văn tồi; một tác nhân hành động sai thì tạo ra một *sự kiện* tồi — một chỉ định được đặt ra, một tin nhắn được gửi đi.[1] Do đó, kỷ luật kỹ thuật của phòng khám tác nhân ít nằm ở độ chính xác thô mà nằm ở *sự tự chủ có giới hạn*: quyền sử dụng công cụ được nêu rõ ràng, các hành động có thể đảo ngược, và một điểm kiểm soát của con người được đặt đúng nơi mà cái giá của sai lầm vọt lên.

Điều cỗ máy làm thay đổi về chẩn đoán

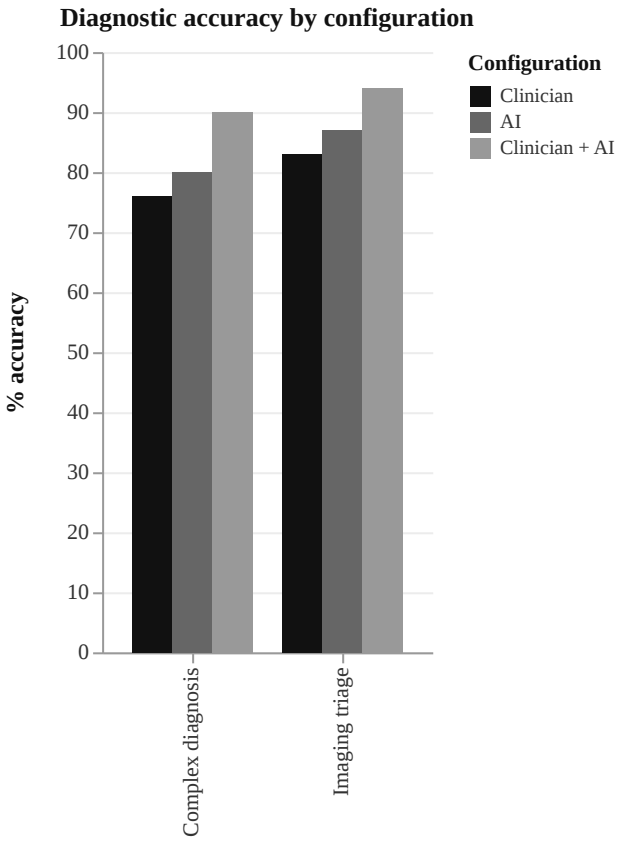
Chẩn đoán vốn luôn mang tính xác suất, ngay cả khi người thầy thuốc làm phép tính trong đầu. Quy tắc Bayes ở dạng tỷ số chênh (odds) là toàn bộ câu chuyện: một xét nghiệm cập nhật niềm tin trước-xét-nghiệm bằng tỷ số khả dĩ của nó,

$$O(D \mid +) = O(D) \times \mathrm{LR}_{+}, \qquad \mathrm{LR}_{+} = \frac{\text{sensitivity}}{1 - \text{specificity}} \tag{3}$$

Vấn đề chưa bao giờ nằm ở chính Phương trình 3; mà là một con người đang xoay xở với một loạt chẩn đoán phân biệt không thể giữ trong đầu hai mươi niềm tin tiên nghiệm cạnh tranh nhau và cập nhật tất cả một cách trung thành khi từng kết quả rơi vào. Một mô hình cỡ ngôn ngữ thì có thể — và, khi được nhắc lệnh như một người đối thoại thay vì một đấng tiên tri, nó còn có thể *hỏi câu hỏi kế tiếp*. Những hệ thống đầu tiên được xây dựng rõ ràng để khai thác bệnh sử theo cách này đã ngang bằng hoặc vượt các bác sĩ chăm sóc ban đầu về độ chính xác chẩn đoán và về những thước đo mềm hơn của một cuộc hội chẩn tốt trong các nghiên cứu có đối chứng (8). Bài học rút ra không phải là cỗ máy là một bác sĩ giỏi hơn. Mà là cỗ máy là một nhà Bayes không biết mệt mỏi, và rằng đây là một phẩm chất hạn hẹp và mang tính bổ trợ.

Cộng sinh, chứ không thay thế

Đây là phát hiện đáng lẽ phải định hình thế hệ thực hành y khoa kế tiếp: qua hết tác vụ này đến tác vụ khác, cặp đôi thầy-thuốc-cộng-AI thắng cả người thầy thuốc lẫn AI đơn độc. Khả năng ghi nhớ và tính nhất quán của cỗ máy che chắn cho sự mệt mỏi và định kiến neo bám của con người; bối cảnh, trách nhiệm giải trình và phán đoán bên giường bệnh của con người che chắn cho những sai lầm đầy tự tin của cỗ máy. Topol đã gọi tên sự hội tụ này nhiều năm trước khi nó có thể triển khai được (9); giờ đây chúng ta có thể đo lường nó.



Các kỷ nguyên không phải là những sự thay thế lẫn nhau cho bằng là những lớp tầng tích lũy, mỗi lớp thêm vào một năng lực mà không vứt bỏ lớp trước. Bảng 1 đặt chúng cạnh nhau.

Bảng 1. Các kỷ nguyên tích lũy của AI y khoa — mỗi lớp thêm vào một năng lực và định nghĩa lại, nhưng không xóa bỏ, vai trò của người thầy thuốc.

Kỷ nguyên	Hệ thống…	Ví dụ lâm sàng	Người thầy thuốc là…
LLM	trả lời	tóm tắt bệnh án, soạn ghi chú	người biên tập
Tác nhân	hành động	chỉ định, đối chiếu, theo dõi	người giám sát
Cộng sinh	đồng suy luận	chạy chẩn đoán phân biệt *cùng* bạn	người cộng sự chịu trách nhiệm
Hình người	hiện thân	thăm khám, hỗ trợ, thực hiện tại giường	người điều hành chăm sóc

Bước ngoặt hiện thân

Mọi thứ cho đến giờ vẫn sống sau một màn hình. Cột cuối cùng của Bảng 1 là nơi biên giới kế tiếp nằm — và đó là biên giới mà đường cong trong Hình 1 chỉ về hướng đó nhưng chưa chạm tới. Y học mang tính vật chất không thể giản lược: một mạch đập được sờ nắn, một vết thương được băng bó, một bệnh nhân được trở mình. Một hệ thống biết suy luận nhưng không thể chạm vào thế giới thì là một chuyên gia hội chẩn, không phải một người chăm sóc.

Bước ngoặt hình người khép lại khoảng cách đó. Ghép một mô hình có thể suy luận về một ca bệnh với một cơ thể có thể đo các dấu hiệu sinh tồn, lấy và bố trí thiết bị, hỗ trợ trở mình hay chuyển bệnh nhân, và đứng canh dài bên giường bệnh — công việc làm kiệt sức nhân viên là người — thì sự cộng sinh của phần trước sẽ có được đôi bàn tay. Thực tế trong ngắn hạn thì không hào nhoáng mà có giá trị: hệ thống hiện thân như một phụ tá điều dưỡng không biết mệt và cánh tay thứ ba của bác sĩ phẫu thuật, chứ không phải một bác sĩ tự chủ. Các ràng buộc không còn chủ yếu mang tính nhận thức nữa — chúng mang tính cơ học, xúc giác, và trên hết là vấn đề an toàn trong một bối cảnh không tha thứ cho việc để rơi một bệnh nhân.

Thế hệ thực hành y khoa kế tiếp trông ra sao

Ghép các mảnh lại và hình hài của phòng khám kế tiếp hiện lên rõ ràng. Mô hình gỡ bỏ gánh nặng ghi chép vốn đẩy người thầy thuốc đến kiệt sức. Tác nhân gánh vác cái đuôi dài và nhàm chán của công tác phối hợp vốn ngốn hết một phần ba ngày làm việc lâm sàng. Hệ thống cộng sinh biến người thầy thuốc thành một nhà chẩn đoán giỏi hơn so với từng bên đơn độc. Và hệ thống hiện thân, khi nó đến, sẽ trả lại cho giường bệnh những giờ đồng hồ mà giấy tờ và hậu cần đã đánh cắp mất.

Chẳng điều nào trong số này cho người bác sĩ về hưu. Nó dời chỗ người bác sĩ — lên cao hơn, từ người thực thi thành người điều hành; từ người ghi nhớ mọi hướng dẫn thành người quyết định khi nào cần phá vỡ một hướng dẫn; từ chuyên gia khan hiếm được phân phối nhỏ giọt qua một danh sách hàng nghìn bệnh nhân thành con người chịu trách nhiệm ở trung tâm của một hệ thống rốt cuộc đã mở rộng được quy mô. Cỗ máy đã trở thành một nhà Bayes không biết mệt, một tác nhân không biết chồn chân, và chẳng bao lâu nữa là một đôi bàn tay vững vàng. Điều nó không thể trở thành là con người ngồi bên bệnh nhân đang sợ hãi và gánh lấy quyết định. Đó, ngày càng rõ, mới là công việc — và đó là một công việc tốt đẹp hơn.

Tài liệu tham khảo

1. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
4. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv preprint arXiv:230313375. 2023;
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In: Advances in Neural Information Processing Systems. 2017.
6. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling Laws for Neural Language Models. arXiv preprint arXiv:200108361. 2020;

7. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–65.
8. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642:442–50.
9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019;25(1):44–56.

Footnotes

1. Đây là lý do trọng tâm của quản lý pháp quy đang dịch chuyển từ *phê duyệt mô hình* sang *phê duyệt quy trình công việc* — chứng nhận vòng lặp mà trong đó một mô hình hành động, chứ không chỉ riêng các trọng số một cách biệt lập. ↩