La médecine après le modèle : du langage à la symbiose

1 juillet 2026 · Nikon Sugar

Entre 2021 et 2026, l'intelligence artificielle en médecine a franchi trois seuils en succession rapide : des modèles de langage capables de répondre, des agents capables d'agir, et les premiers systèmes conçus pour raisonner aux côtés d'un clinicien plutôt qu'en dessous de lui. Cet essai retrace cette accélération, expose les mathématiques qui l'ont rendue possible et soutient que la destination n'est ni l'automatisation ni le remplacement, mais la symbiose — avec des soins incarnés, prodigués par des humanoïdes, comme prochaine frontière. La question qui se pose à la prochaine génération de médecins n'est pas de savoir si la machine est assez bonne, mais à quoi sert un clinicien une fois qu'elle l'est.

Pendant la plus grande partie de son histoire, la médecine computationnelle a été une étude de l'étroitesse. Un modèle pouvait lire une mammographie (1) ou évaluer une lésion cutanée (2) au niveau d'un spécialiste, et chacun de ces systèmes était un monument à une tâche unique — entraîné à grands frais, déployé dans une seule clinique, aveugle à tout ce qui se trouvait hors de son cadre. La promesse était réelle et le plafond était bas. Ce qui a changé, à partir de 2021 environ, n'est pas que les modèles se sont légèrement améliorés. C'est qu'ils ont cessé d'être étroits.

Cet essai porte sur ce changement et sur ce à quoi il mène. La version courte de l'argument est résumée dans la Figure 1 : en cinq ans, l'IA médicale s'est élevée de systèmes qui *répondent*, en passant par des systèmes qui *agissent*, vers des systèmes qui *co-raisonnent* avec un clinicien — et la pente ne s'aplatit pas là où le diagramme finit à court d'encre.

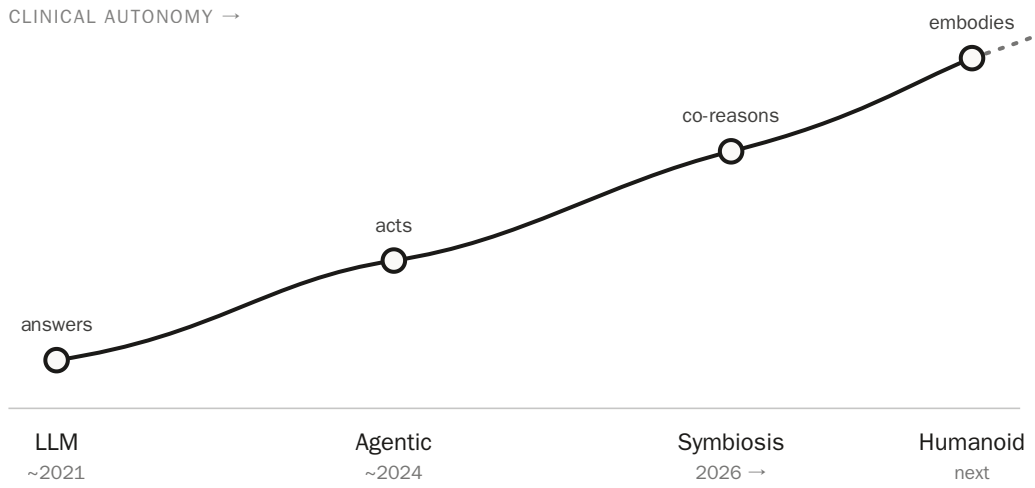
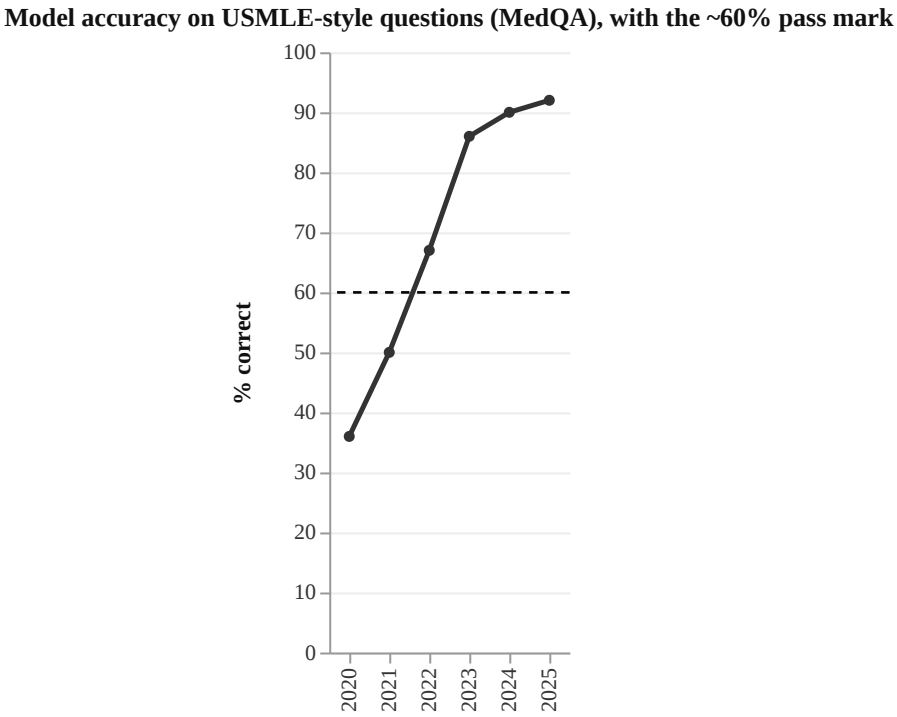


Figure 1. Quatre ères de l'IA médicale sur une courbe ascendante d'autonomie clinique : des modèles de langage qui répondent, des systèmes agentiques qui agissent, une symbiose humain–IA qui co-raisonne, et des soins incarnés prodigués par des humanoïdes.

Le bond de cinq ans

Il est facile d'oublier à quel point cela a été rapide. Encore en 2020, un grand modèle tentant l'examen de licence médicale des États-Unis (United States Medical Licensing Examination) obtenait un score situé dans la fourchette haute des trente pour cent — en deçà du hasard pour un étudiant motivé, loin du seuil de réussite d'environ 60 %. À la fin de 2022, un modèle affiné pour la médecine franchissait la barre pour la première fois (3) ; en l'espace d'un an, des systèmes à usage général obtenaient des scores dans la fourchette haute des quatre-vingts pour cent sur les mêmes banques de questions (4), et la courbe n'a cessé depuis de se rapprocher de son plafond.



Pourquoi si vite ? La réponse honnête, c'est que la médecine a bénéficié des retombées d'une découverte plus générale. L'architecture qui sous-tend tout cela — le transformeur — a remplacé la récurrence par une opération unique, l'attention, qui permet à chaque jeton de pondérer directement tous les autres (5) :

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

Ce que la Équation 1 a apporté, ce n'était pas du savoir médical mais de la *scalabilité* — un modèle dont la qualité s'améliore de manière prévisible à mesure que l'on ajoute des paramètres et des données. Cette prévisibilité a elle-même été mesurée : sur de nombreux ordres de grandeur, la perte décroît selon une loi de puissance en fonction de la taille du modèle,

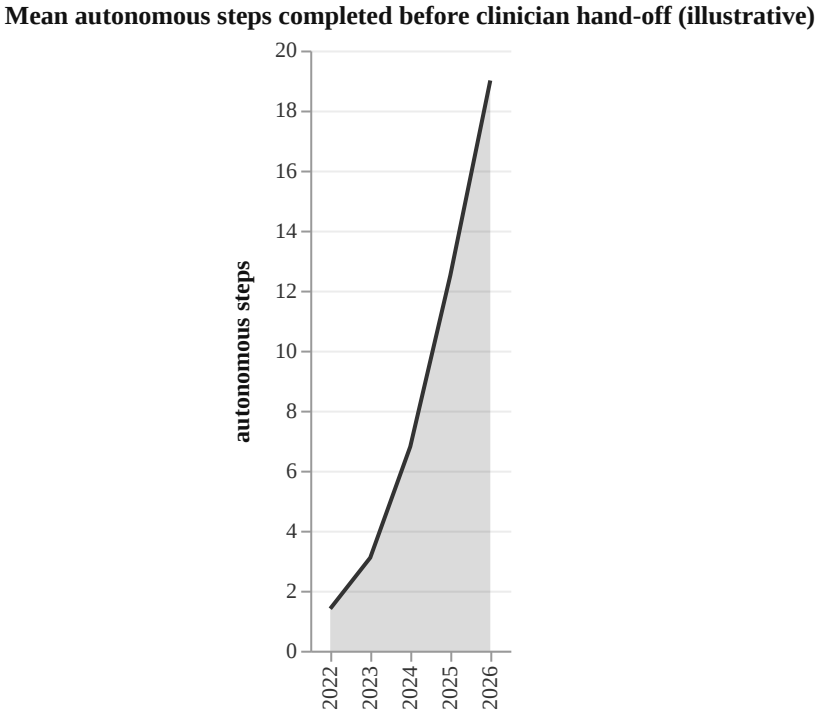
$$L(N) \approx \left(\frac{N_c}{N}\right)^{\alpha_N} \tag{2}$$

avec un petit exposant α_N (6). La Équation 2 est le moteur discret qui se cache derrière la ligne abrupte ci-dessus : la médecine n'avait pas besoin d'une percée sur mesure chaque année, seulement d'une part d'une courbe que tout le domaine gravissait. Un modèle entraîné sur le corpus ouvert du langage humain avait, presque incidemment, lu assez de médecine pour raisonner à son sujet — et un substrat généraliste, légèrement affiné, a commencé à surpasser les spécialistes étroits qu'il était censé se contenter d'assister (7).

Des réponses aux actions

Un examen récompense un système qui *sait*. Une clinique récompense un système qui *fait* — qui commande l'analyse, réconcilie la liste des médicaments, rédige l'orientation, signale l'interaction et relance le résultat trois jours plus tard. La deuxième vague de l'IA médicale, grosso modo à partir de 2023, a été le passage d'un modèle que l'on interroge à un **agent** qui poursuit un objectif à travers de nombreuses étapes, en appelant des outils et en lisant leurs sorties au fur et à mesure.

La capacité qui définit l'ère agentique n'est pas l'éloquence mais la *longueur de la tâche* : quelle part d'un flux de travail réel un système peut mener avant de devoir repasser la main à un humain. Ce nombre a fortement augmenté.



C'est aussi là que les enjeux changent. Un modèle qui répond mal à une question produit une mauvaise phrase ; un agent qui agit mal produit un mauvais *événement* — une commande passée, un message envoyé.[1] La discipline d'ingénierie de la clinique agentique porte donc moins sur l'exactitude brute que sur l'*autonomie bornée* : des permissions d'outils explicites, des actions réversibles et un point de contrôle humain placé exactement là où le coût de l'erreur s'envole.

Ce que la machine change à propos du diagnostic

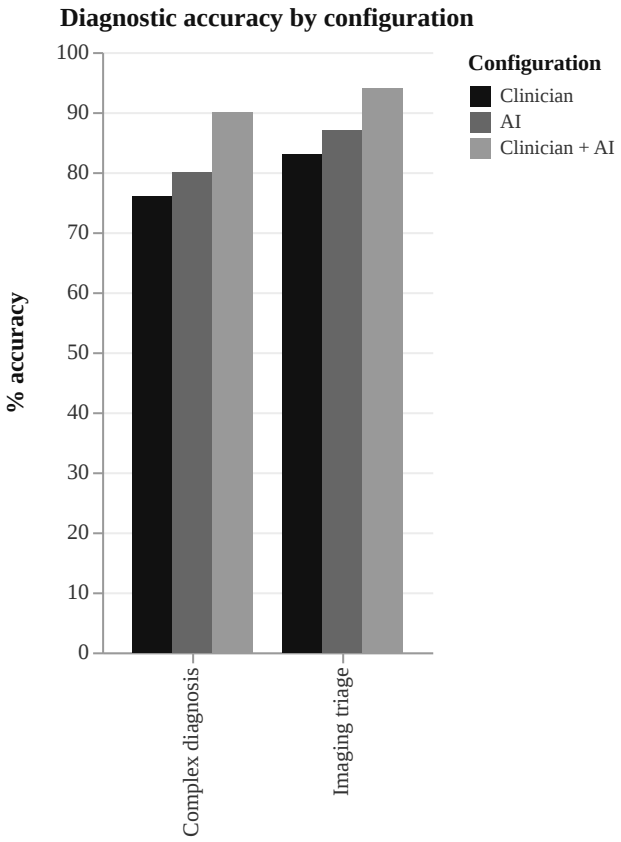
Le diagnostic a toujours été probabiliste, même lorsque les cliniciens faisaient l'arithmétique dans leur tête. La règle de Bayes sous forme de cotes en résume tout : un test met à jour une croyance pré-test par son rapport de vraisemblance,

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

Le problème n'a jamais été la Équation 3 elle-même ; c'est qu'un humain jonglant avec un diagnostic différentiel ne peut pas tenir vingt a priori concurrents et les mettre tous à jour fidèlement à mesure que chaque résultat tombe. Un modèle de niveau langagier le peut — et, sollicité comme un partenaire de conversation plutôt que comme un oracle, il peut aussi *poser la question suivante*. Les premiers systèmes conçus explicitement pour recueillir une anamnèse de cette manière ont égalé ou dépassé les médecins de premier recours en matière de précision diagnostique et sur les mesures plus subtiles d'une bonne consultation, dans des études contrôlées (8). La leçon n'est pas que la machine est un meilleur médecin. C'est que la machine est un bayésien infatigable, et que c'est là une vertu étroite et complémentaire.

Symbiose, et non substitution

Voici le résultat qui devrait façonner la prochaine génération de la pratique : tâche après tâche, le tandem clinicien-plus-IA surpasse aussi bien le clinicien que l'IA pris isolément. Le rappel et la constance de la machine compensent la fatigue et l'ancrage de l'humain ; le contexte, la responsabilité et le jugement au chevet du patient de l'humain compensent les erreurs assurées de la machine. Topol a annoncé cette convergence des années avant qu'elle ne soit déployable (9) ; nous pouvons désormais la mesurer.



Les ères ne se remplacent pas les unes les autres, elles s'accumulent plutôt en couches, chacune ajoutant une capacité sans écarter la précédente. Le Tableau 1 les présente côte à côte.

Tableau 1. Les ères cumulatives de l'IA médicale — chaque couche ajoute une capacité et redéfinit, sans l'effacer, le rôle du clinicien.

Ère	Le système…	Exemple clinique	Le clinicien est…
LLM	répond	résume un dossier, rédige une note	l'éditeur
Agentique	agit	commande, réconcilie, assure le suivi	le superviseur
Symbiose	co-raisonne	mène le différentiel *avec* vous	le partenaire responsable
Humanoïde	s'incarne	examine, assiste, exécute au chevet	le directeur des soins

Le tournant incarné

Tout ce qui précède vit derrière un écran. La dernière colonne du Tableau 1 est là où se trouve la prochaine frontière — et c'est celle vers laquelle la courbe de la Figure 1 pointe sans encore l'atteindre. La médecine est irréductiblement physique : un pouls se palpe, une plaie se panse, un patient se retourne. Un système de raisonnement qui ne peut pas toucher le monde est un consultant, pas un soignant.

Le tournant humanoïde comble cet écart. Associez un modèle capable de raisonner sur un cas à un corps capable de prendre les signes vitaux, d'aller chercher et de positionner l'équipement, d'aider à un retournement ou à un transfert, et de tenir la longue veille au chevet qui épuise le personnel humain, et la symbiose de la section précédente acquiert des mains. La réalité à court terme est peu glorieuse et précieuse : le système incarné comme l'aide-soignant infatigable et le troisième bras du chirurgien, non le médecin autonome. Les contraintes ne sont plus principalement cognitives — elles sont mécaniques, tactiles et, avant tout, une affaire de sécurité dans un cadre qui ne pardonne pas un patient qu'on laisse tomber.

À quoi ressemble la prochaine génération de la pratique

Assemblez les pièces et la forme de la clinique de demain devient lisible. Le modèle supprime l'impôt documentaire qui pousse les cliniciens à l'épuisement professionnel. L'agent prend en charge la longue et fastidieuse traîne de coordination qui consomme un tiers d'une journée clinique. Le système symbiotique fait du clinicien un meilleur diagnosticien que chacune des deux parties seule. Et le système incarné, quand il arrivera, restituera au chevet du patient les heures que la paperasse et la logistique lui avaient dérobées.

Rien de tout cela ne met le médecin à la retraite. Cela relocalise le médecin — vers le haut, de l'exécutant au directeur ; de celui qui se souvient de chaque recommandation à celui qui décide quand en enfreindre une ; de l'expert rare rationné sur un panel de milliers de patients à l'humain responsable au centre d'un système qui passe enfin à l'échelle. La machine est devenue un bayésien infatigable, un agent inlassable et, bientôt, une paire de mains sûres. Ce qu'elle ne peut pas devenir, c'est la personne qui s'assied auprès du patient effrayé et assume la décision. Voilà, de plus en plus, le métier — et c'en est un meilleur.

Références

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. C'est pourquoi le centre de gravité réglementaire se déplace de l'*approbation du modèle* vers l'*approbation du flux de travail* — certifier la boucle dans laquelle un modèle agit, et non les seuls poids pris isolément. ↩