

Compte-rendu

IA en médecine : le spécialiste ne bat pas le généraliste

Mots-clés

Intelligence artificielle, grand modèle de langage, aide à la décision clinique
General-purpose large language models outperform specialized clinical AI tools on medical benchmarks

K. Vishwanath et al., Nature Medicine, 12.06.2026

DOI : [10.1038/s41591-026-04431-5](https://doi.org/10.1038/s41591-026-04431-5)

Introduction

Les grands modèles de langage (Large Language Models, LLM), tels que ChatGPT ou Gemini, sont de plus en plus utilisés par les cliniciens pour interroger la littérature médicale ou aider à la décision. Des outils dits « spécialisés » ont émergé en parallèle, vendus comme supérieurs aux LLM généralistes parce qu'ils s'appuient sur une couche de récupération documentaire (le Retrieval-Augmented Generation, RAG) qui ancre les réponses dans une base de données médicale épurée. Deux d'entre eux dominent le marché : OpenEvidence (gratuit pour les médecins certifiés, financé par la publicité) et UpToDate Expert AI (~ 700 USD/an). Leurs architectures et bases d'entraînement ne sont pas rendues publiques et aucune évaluation indépendante ne les a comparés aux LLM généralistes de dernière génération. Cette étude s'y attelle.

Méthode

Étude comparative indépendante menée à NYU Langone Health (grand centre hospitalier universitaire à New York, USA) opposant, sur 3 étapes, 6 systèmes : 3 LLM généralistes (GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.6), 2 outils IA cliniques (OpenEvidence, UpToDate Expert AI), ainsi que Google Search AI Overview (résumé IA automatique des résultats Google). La 1ère étape « MedQA » teste les connaissances médicales encyclopédiques (500 QCM de type US Medical Licensing Examination); la 2ème étape « HealthBench » mesure l'alignement avec les avis d'experts (500 questions ouvertes standardisées évaluées par un panel de LLM juges); et la 3ème étape « Real Clinical Queries (RCQ) » teste les systèmes en situation réelle: 100 vraies requêtes anonymisées posées par des médecins du NYU Langone Health à une instance GPT interne, évaluées en aveugle par 12 cliniciens sur 4 dimensions (justesse clinique, complétude, sécurité, clarté) via une échelle de 1 à 4, avec signalement pour contenu délétère et hallucination (invention d'un fait). 3 évaluateurs indépendants par réponse, soit 1'800 annotations au total. Le RCQ constitue le seul benchmark exempt de « fuite » (les modèles ne peuvent pas l'avoir vu à l'entraînement) et est le plus proche de la pratique réelle.

Résultats

Sur les 3 étapes (benchmarks), les LLM généralistes surpassent les outils cliniques. MedQA (exactitude): Gemini 97.4 % > GPT 94.2 % > Claude 90.2 % ≈ OpenEvidence 89.6 % ≈ UpToDate 88.4 %. HealthBench (score /100): GPT 88 > Gemini 79.3 ≈ Claude 77 > OpenEvidence 62.6 ≈ UpToDate 61.3. Sur le RCQ (score moyen agrégé /4), deux niveaux de performance ressortent clairement : LLM généralistes en tête (Gemini 3.62 ; GPT 3.54 ; Claude 3.52) puis, significativement en dessous, les outils cliniques et Google (OpenEvidence 3.24 ; UpToDate 3.17 ; Google AI Overview 3.27), sans différence entre ces trois derniers. L'écart tenait sur les 4 dimensions, avec un déficit particulièrement marqué en clarté pour OpenEvidence (score moyen 2.84). UpToDate Expert AI refusait de répondre

à 19 % des requêtes (vs 1-3 % pour les LLM généralistes). Aucun modèle ne se distinguait sur les taux de contenu délétère (0-3 %) ou d'hallucination (0-1 %). Les 12 cliniciens classaient les modèles de manière très concordante, plaçant systématiquement les LLM généralistes au-dessus des outils cliniques.

Discussion

Première évaluation indépendante à comparer des outils IA cliniques commerciaux à des LLM généralistes de dernière génération sur de vraies requêtes de médecins. Le message est clair : la spécialisation clinique par RAG ne se traduit ni par un gain de justesse, ni de complétude, ni de sécurité, ni de clarté. Les outils cliniques ne font pas mieux que le résumé IA gratuit de Google Search. Explication probable : les LLM généralistes bénéficient de corpus d'entraînement massifs, d'itérations rapides et d'un alignement soigné, tandis que le RAG peut nuire à la performance si des documents non pertinents sont récupérés ou mal intégrés par le modèle sous-jacent. Points forts : évaluation indépendante, benchmark RCQ construit à partir de vraies requêtes cliniques, double aveugle par 12 cliniciens, analyses statistiques rigoureuses. Limites : outils cliniques interrogés manuellement via navigateur (pas d'application programming interface (API) disponible donc moins standardisé), possible « fuite » de MedQA/HealthBench durant l'entraînement des LLM généralistes, HealthBench développé par OpenAI (potentiel biais favorisant GPT-5.2), absence d'évaluation de la latence ou de la qualité des citations. Enfin, ce paysage évolue vite : ces résultats pourraient déjà ne plus être valables avec l'apparition de nouveaux modèles.

Conclusion

Un outil « spécialisé clinique » (souvent onéreux et labellisé médical) ne garantit pas une performance supérieure à un LLM généraliste grand public (voire à une simple recherche Google) pour des requêtes cliniques courantes. L'adoption d'une IA en pratique médicale devrait reposer sur une évaluation indépendante par des cliniciens, plutôt que sur les benchmarks fournis par les fournisseurs eux-mêmes. Quel que soit l'outil retenu, aucun n'atteint aujourd'hui une performance sans faille : ils restent des aides à la réflexion, jamais des substituts au jugement clinique.

Date de publication	Auteurs
04.08.2026	Benoit Delabays