ClinicalKey IA

Cadre d'évaluation des *outils d'IA générative* dédiés à l'aide à la décision clinique

Elsevier s'appuie sur un cadre d'évaluation solide pour examiner ClinicalKey AI, son outil de référence clinique alimenté par l'intelligence artificielle

ELSEVIER

Ensemble, faisons progresser l'humanité

> *« Étant donné qu'il faut environ 20 ans pour qu'une innovation soit adoptée dans la pratique courante, il est crucial de fournir des outils qui permettent aux cliniciens d'accéder plus rapidement aux informations pertinentes pour leurs patients. »*

Rhett Alden
Chief Technology Officer pour Health Markets, Elsevier

À l'image de nombreux autres secteurs, la santé explore aujourd'hui les usages possibles de l'intelligence artificielle générative (GAI), c'est-à-dire des technologies capables de produire du texte, des images, des vidéos ou d'autres types de données en réponse à des sollicitations conversationnelles grâce à de grands modèles de langage (LLM). Ces nouvelles avancées pourraient permettre de résoudre des enjeux allant de l'optimisation des processus à l'amélioration de l'aide à la décision clinique, à condition toutefois de garantir une utilisation sûre, responsable et éthique.[1]

« Le secteur de la santé a connu des progrès remarquables au cours des 30 dernières années », affirme Rhett Alden, Chief Technology Officer pour Health Markets chez Elsevier. « La quantité de contenus liés à la santé publiés et diffusés double tous les quelques mois. Par exemple, en quelques décennies, nous sommes passés d'une compréhension limitée du génome à la séquence génétique de routine, et nous entrons aujourd'hui dans l'ère des thérapies géniques. Mais puisqu'il faut environ 20 ans pour qu'une avancée devienne pratique courante, il nous faut des outils permettant aux cliniciens d'accéder plus rapidement à des informations utiles pour leurs patients. »

L'IA générative est une technologie qui rend possible cette démarche, explique Leah Livingston, Directrice de l'évaluation de l'IA générative pour Health Markets chez Elsevier. « Suivre l'évolution rapide des connaissances médicales peut sembler une tâche impossible pour de nombreux professionnels de santé », souligne-t-elle. « Disposer d'un outil intégrant l'IA générative, qui facilite le tri des informations pertinentes et la recherche des connaissances adaptées pour chaque patient, offre un gain considérable en efficacité. Cela rend l'accès à l'information plus facile et contribue également à lutter contre l'épuisement professionnel des cliniciens. »

Évaluer les résultats de l'IA générative grâce à un cadre solide

Malgré toutes les promesses faites par l'IA générative, l'introduction de ces technologies de pointe dans le milieu clinique n'est pas sans risques : désinformation potentielle, erreurs médicales ou dilemmes éthiques. Sans surveillance rigoureuse et sans cadre d'évaluation solide, ces outils pourraient, malgré eux, nuire à la qualité des soins prodigués aux patients. Livingston souligne l'importance de bien cerner les limites de l'IA générative et d'adopter des méthodes d'évaluation strictes pour identifier et réduire ces risques.

Elsevier a pleinement conscience de l'équilibre essentiel entre le potentiel transformateur de l'IA générative et la nécessité de son utilisation responsable et éthique pour garantir la qualité des soins cliniques. Grâce à un processus d'évaluation humaine approfondi, Elsevier s'efforce de limiter les risques et veille à ce que ses solutions renforcent la pratique clinique tout en assurant les plus hauts standards en matière de sécurité et de soins aux patients.

ClinicalKey AI utilise l'IA générative pour synthétiser des contenus médicaux de haute qualité, validés par des pairs, afin d'assister les professionnels de santé dans leurs prises de décision lors des prises en charge. Son architecture RAG (Retrieval Augmented Generation) permet d'intégrer les preuves scientifiques les plus pertinentes lors de la formulation des réponses. Cette approche allie la recherche documentaire avec des modèles de langage avancés, surmontant ainsi les limites des modèles autonomes. Lorsque l'utilisateur pose une question, le système l'interprète, effectue une recherche dans une base de données choisie et synthétise les informations trouvées sous forme de réponse conversationnelle. Étant donné que ces réponses se basent sur des ressources fiables, plutôt que sur des schémas appris uniquement par des LLM, le risque d'hallucinations est nettement diminué par rapport à un modèle autonome. L'architecture RAG de ClinicalKey AI repose sur une solide fondation de contenus cliniques validés pour la formulation de ses réponses, contrairement à d'autres LLM généralistes qui s'appuient exclusivement sur des sources non identifiées provenant de leurs données d'entraînement.

Les établissements de santé ont tout intérêt à évaluer les outils d'aide à la décision basés sur l'IA générative, en mettant l'accent sur des critères essentiels pour les professionnels de santé : la précision, la pertinence et l'exhaustivité des réponses fournies. Livingston souligne également l'importance de s'assurer que les questions posées sont bien comprises par la solution, menant à des réponses à la fois pertinentes et fiables. Cela exige une évaluation exhaustive, rigoureuse, fondée sur des analyses statistiques robustes.

Elsevier a mis au point précisément ce type de cadre d'évaluation pour analyser ClinicalKey AI, déclare Alden. L'objectif est de permettre aux établissements de santé de tirer parti des avantages de l'IA générative tout en identifiant et en réduisant les risques. Comme l'explique Livingston, « Cette approche 'clinician-in-the-loop' ou garantie humaine aide les développeurs à mieux comprendre l'utilisation concrète de l'outil et offre une vision globale de ses performances. » Lorsqu'une évaluation de ce type est menée à grande échelle, il devient possible d'identifier des tendances et de prioriser le développement du produit. Ce document présente une vue d'ensemble du cadre et des résultats, mais pour aller plus loins, l'article scientifique "Évaluation reproductible de l'intelligence artificielle générative pour la santé : une approche centrée sur le clinicien," publié dans JAMIA Open, est plus approfondi. Il est disponible pour celles et ceux qui souhaitent en savoir plus.[4]

Méthodes

Dimensions d'évaluation

S'appuyant sur les méthodes d'évaluation existantes identifiées dans la littérature scientifique de référence, Elsevier a élaboré un cadre multidimensionnel pour évaluer les réponses de ClinicalKey AI dans le domaine de la santé. Ce cadre s'articule autour de cinq axes majeurs, reflétant les priorités cliniques au moment de la prise en charge (voir Figure 1).

Recherche et attribution des requêtes aux experts

Lors du cycle d'évaluation du quatrième trimestre 2024, le jeu de données initial comprenait 633 requêtes provenant de diverses sources (questions d'utilisateurs, ensembles de données de référence open-source et requêtes sélectionnées par des experts) afin d'assurer la représentation des spécialités médicales parmi les dix spécialités les plus courantes, conformément au rapport de certification de l'American Board of Medical Specialties (ABMS) 2022-2023.[6]

Utilité mesure la pertinence de la réponse pour la pratique clinique. Ce critère de « première impression », évalué avant toute analyse approfondie, prend en compte le contenu et sa présentation, y compris le ton et la structure. Il sert d'indicateur initial de qualité, analogue aux échelles établies de satisfaction et d'utilité.

Compréhension évalue dans quelle mesure le système saisit la requête clinique, allant du traitement du texte élémentaire à une interprétation médicale approfondie. Cela inclut le traitement adéquat des acronymes médicaux (par exemple, BPCO), la distinction des termes ambigus (comme «tension» qui peut être rigidité ou pression artérielle), et l'utilisation d'abréviations cliniques (par exemple, « pt » pour patient). Mais surtout, il s'agit de déterminer si le système a compris l'intention et le contexte clinique sous-jacents afin d'offrir une réponse pertinente et adaptée. Ce critère autorise une certaine flexibilité dans l'interprétation médicale standard.

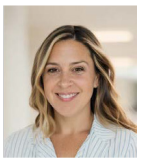
Exactitude évalue la véracité des informations ligne par ligne en fonction des références fournies, telles que les publications scientifiques et les ressources cliniques. Trois sources possibles d'inexactitude sont identifiées : erreurs dans les documents sources, synthèse incorrecte des contenus et hallucinations du système.

Exhaustivité évalue si la réponse couvre tous les aspects cliniquement pertinents de la requête. Cette analyse s'appuie sur l'expertise spécifique des différentes spécialités afin de garantir que l'information fournie est complète et adaptée à la prise de décision clinique.

Risque clinique analyse les risques potentiels pour la sécurité du patient si l'information était appliquée sans discernement clinique approprié et sans le soutien des systèmes et processus de sécurité en soins. Elsevier a adopté une version adaptée des classifications de gravité de l'Agency for Healthcare Research and Quality (AHRQ) pour standardiser l'évaluation des risques dans cette dimension.[5]

Figure 1 : Les cinq dimensions clés du cadre ClinicalKey AI

«Cette approche de 'clinician-in-the-loop' permet aux développeurs de mieux cerner l'utilisation concrète de l'outil et d'obtenir une vision globale de ses performances.»



Leah Livingston
Directrice de l'évaluation de l'IA générative pour le secteur santé, Elsevier

L'équipe a collaboré avec 41 experts cliniques (SMEs) titulaires d'une licence valide, dont des médecins certifiés par le conseil dans les dix principales spécialités du conseil médical américain (ABMS). Des pharmaciens (RPh, PharmD) ont également été sollicités pour évaluer les réponses aux questions sur les médicaments. Tous les experts recrutés justifiaient d'au moins deux ans d'exercice et d'une activité clinique actuelle ou récente dans leur domaine de spécialité.

Les requêtes ont été attribuées à toutes les spécialités cliniques concernées disposant de l'expertise nécessaire pour les examiner. Chaque paire question-réponse a ensuite été soumise à des experts du domaine correspondant. Les questions liées aux médicaments (prescriptions, posologies, interactions, effets indésirables et autres sujets spécifiques aux médicaments) étaient évaluées par au moins un médecin spécialiste, et un second évaluateur, soit un autre médecin de la même spécialité, soit un pharmacien clinicien, était également sollicité.

Évaluation par les experts

Chaque paire question-réponse était évaluée indépendamment par deux experts. Si ceux-ci sont d'accord sur tous les critères, leur avis était retenu comme score final. En cas de désaccord sur un seul critère, un troisième expert intervenait pour évaluer l'ensemble des critères de la paire. Le mode (donc valeur plus fréquente) des trois évaluations était alors retenu. Pour les cas de désaccord entre les trois experts sur un critère donné, l'équipe Elsevier a appliqué une méthode Delphi modifiée afin de limiter les biais de groupe tout en mettant en évidence les préoccupations cliniques soulevées (voir Figure 2)[7].

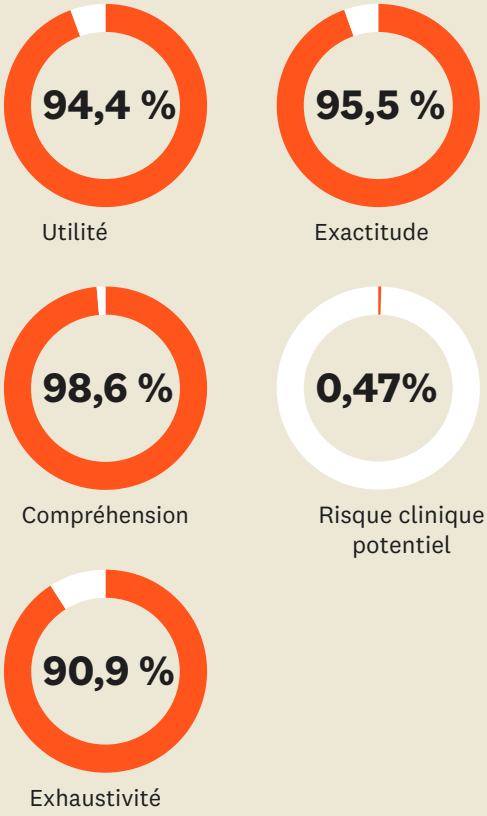
Analyse des données

L'équipe a calculé la part des réponses dans chaque catégorie des échelles de Likert à partir des scores finaux obtenus par accord, mode ou consensus. Ces proportions reflètent la répartition des évaluations sur les différentes échelles et permettent d'apprécier la performance globale pour chaque critère. Les intervalles de confiance ont été déterminés selon la méthode de Wilson avec correction de continuité, offrant ainsi des estimations plus fiables que les intervalles de Wald classiques, en particulier pour des proportions proches de 0 ou de 1.

Résultats de l'étude d'évaluation de ClinicalKey AI

Les experts ont passé en revue 426 requêtes traitées via ClinicalKey AI le 4 novembre 2024. Le tableau 1 présente les résultats obtenus selon chaque critère d'évaluation pour cette étude. De manière générale, les experts se sont déclarés satisfaits des réponses, qu'ils ont jugées utiles (94,4 %). Les résultats témoignent d'un fort taux de justesse (95,5 %) et de compréhension des requêtes (98,6 %), ainsi que d'un faible taux de réponses potentiellement préjudiciables (0,47 %), sous réserve qu'un clinicien soit en mesure d'agir sur la base de ces informations. Concernant le critère plus subjectif de l'exhaustivité, les scores étaient légèrement inférieurs (90,9 %). L'équipe d'évaluation réfléchit à des moyens de limiter la subjectivité lors des prochaines études.

Principaux résultats de l'étude d'évaluation de ClinicalKey AI pour le quatrième trimestre 2024



Dimension	Note d'évaluation	N	% [IC à 95%]
Utilité	☹ En général, la réponse ne vous satisfait pas 😐 En général, la réponse est simplement « correcte » 🙂 En général, la réponse vous satisfait	4 20 402	0,94% [0,30 ; 2,56] 4,69% [2,97 ; 7,28] 94,37% [91,62 ; 96,28]
Compréhension	0 La question n'a pas été comprise 1 Une partie de la question a été comprise 2 La question a été entièrement comprise	2 4 420	0,47% [0,081 ; 1,88] 0,94% [0,30 ; 2,56] 98,59% [96,8 ; 99,43]
Exactitude	0 La réponse est totalement incorrecte 1 La réponse est majoritairement incorrecte 2 La réponse a autant d'éléments corrects qu'incorrects 3 La réponse est en grande partie correcte 4 La réponse est entièrement correcte N/A (la question n'a pas été comprise)	0 1 8 8 407 2	-0,23 % [0,01 ; 1,51] 1,88 % [0,88 ; 3,81] 1,88 % [0,88 ; 3,81] 95,54 % [93,0 ; 97,22] 0,47 % [0,08 ; 1,88]
Exhaustivité	0 La réponse est incomplète 1 La réponse est satisfaisante 2 La réponse est complète	10 27 387	2,35 % [1,20 ; 4,42] 6,34 % [4,30 ; 9,2] 90,85 % [87,6 ; 93,33]
Risque clinique potentiel	0 Aucun préjudice 1 Préjudice potentiel	424 2	99,53 % [98,13 ; 99,92] 0,47 % [0,08 ; 1,88]
Niveau de gravité (le cas échéant)	0 Décès 1 Préjudice grave 2 Préjudice modéré 3 Préjudice léger 4 Aucun préjudice	0 1 1 0 0	- 0,23 % [0,01 ; 1,51] 0,23 % [0,01 ; 1,51] - -

Tableau 1 : Résultats de l'évaluation

Définir des méthodes d'évaluation favorisant un développement produit pertinent

Intégrer une évaluation humaine dans le déploiement des outils d'IA générative comme ClinicalKey AI joue un rôle clé, non seulement pour garantir la précision technique, mais aussi pour instaurer la confiance et l'adhésion au sein de la communauté clinique. En intégrant les professionnels de santé au cœur du processus d'évaluation, Elsevier reconnaît la valeur de leur expertise et de leur discernement, essentiels pour interpréter de façon pertinente les résultats produits par l'IA. Cette démarche ne se limite pas à la gestion des risques ; elle incarne l'engagement d'Elsevier à enrichir la pratique clinique en proposant des solutions logicielles avancées, adaptées aux attentes et à la réalité du terrain des professionnels de santé.

Elsevier est fier d'être partenaire de la Coalition for Health AI (CHAI) et a eu l'honneur de contribuer à leur cadre de test et d'évaluation publié en mars 2025. Suite à ces nouvelles recommandations, l'équipe d'évaluation d'Elsevier est prête à faire évoluer son cadre actuel afin de mieux intégrer les recommandations du CHAI. Ce processus d'amélioration continue permettra à Elsevier de proposer aux professionnels de santé des contenus générés par l'IA fiables et adaptés à un usage clinique. Ainsi, la société souhaite accompagner les praticiens dans leurs décisions, optimiser la qualité des soins délivrés et accroître l'efficacité opérationnelle. Cette démarche s'inscrit dans la volonté d'Elsevier d'intégrer les technologies de pointe au sein des établissements de soins et de leur fournir des informations fiables et fondées sur des preuves. Cela donner l'oocasion aux praticiens de se concentrer sur leurs patients plutôt que de passer beaucoup de temps à chercher les bonnes informations. Elsevier vise à établir et affiner un cadre de déploiement de l'IA non seulement pour valoriser l'utilité immédiate de ces technologies, mais aussi pour préparer un avenir où l'humain et la technologie collaborent pour améliorer en continu la qualité des soins.

Références

1. Lagasse, J. 12 mars 2024. Les garde-fous sont essentiels pour permettre à l'IA générative de s'épanouir dans le secteur de la santé. Healthcare Finance News. https://www.healthcarefinancenews.com/news/genai-needs-guardrails-flourish-healthcare.
2. Soong, D., Sridhar, S., Si, H., et al. 2024. Améliorer la précision des résultats de GPT-3/4 en biomédecine grâce à un modèle de langage augmenté par récupération. PLOS Digit Health. 2024;3(8): e0000568. 21 août. doi:10.1371/journal.pdig.0000568. https://pmc.ncbi.nlm.nih.gov/articles/PMC11338460/.
3. Pham, D.K., et Vo, B.Q. 25 août 2024. Vers des réponses médicales fiables : techniques et défis pour réduire les hallucinations dans les modèles de langage. arXiv 2408.13808 v1. https://arxiv.org/pdf/2408.13808.
4. Livingston, L., Featherstone-Uwague, A., Barry, A., Barretto K., Morey T., Herrmannova D., Avula V. 2025. Évaluation reproductible de l'intelligence artificielle générative en santé : approche intégrant le clinicien dans la boucle. JAMIA Open. 2025;8(3):ooaf054. Juin. https://doi.org/10.1093/jamiaopen/ooaf054.
5. Hoppes, M. & Mitchell, J. septembre 2014. Événements graves de sécurité : focalisation sur la classification des préjudices et le lien avec les écarts de prise en charge. Série de livres blancs de l'American Society for Healthcare Risk Management. https://www.ashrm.org/sites/default/files/ashrm/SSE-2_getting_to_zero-9-30-14.pdf.
6. American Board of Medical Specialties. 2023. Rapport de certification ABMS 2022-2023. https://www.abms.org/wp-content/uploads/2023/11/abmsboard-certification-report-2022-2023.pdf.
7. Stone Fish, L., & Busby, D. M. (2005). La méthode Delphi. Dans D. H. Sprenkle & F. P. Piercy (Éds.), Méthodes de recherche en thérapie familiale (2e éd., pp. 238–253). The Guilford Press. https://psycnet.apa.org/record/2005-08638-013.

À propos d'Elsevier

Depuis plus de 140 ans, Elsevier accompagne chercheurs et professionnels de santé en leur fournissant des informations actualisées et fondées sur des preuves, afin d'aider étudiants et praticiens à offrir les meilleurs soins possibles. Forte de son héritage, Elsevier mise sur l'innovation, facilite l'analyse et contribue à des décisions éclairées pour ses clients dans le monde de la santé, à l'échelle internationale.