

Compte-rendu

Le diagnostic médical par IA mis à l'épreuve du grand public

Mots-clés

IA, LLM, diagnostic médical

Reliability of LLMs as medical assistants for the general public: a randomized preregistered study

Bean A.M, et al., Nature Medicine, 09.02.2026

<https://doi.org/10.1038/s41591-025-04074-y>

Introduction

Les grands modèles de langage (LLMs) suscitent un intérêt croissant pour leur potentiel à démocratiser l'accès aux connaissances médicales. Bien que ces modèles obtiennent d'excellents scores aux examens médicaux standardisés, leur performance dans des situations réelles avec des utilisateurs non experts reste insuffisamment explorée. Cette étude a évalué la capacité des LLMs à aider le grand public à identifier des conditions médicales sous-jacentes et à choisir une conduite appropriée dans des scénarios cliniques réalistes.

Méthode

Étude contrôlée randomisée à quatre bras parallèles, réalisée en Angleterre. **Population** : 1'298 participant·es adultes recrutés via la plateforme en ligne Prolific (grand public, sans expertise médicale). **Critères d'exclusion** : non-complétion de l'étude (pas de différence significative de taux d'attrition entre les groupes). **Intervention** : Dix scénarios médicaux fictifs, allant de situations bénignes à des urgences médicales, ont été élaborés par trois médecins jusqu'à consensus unanime sur la prise en charge appropriée, allant de l'automédication à l'appel des secours. Quatre médecins supplémentaires ont fourni des diagnostics différentiels pour constituer des listes gold-standard de conditions médicales pertinentes. Chaque participant·e était assigné·e aléatoirement à l'un des quatre groupes : GPT-4o (OpenAI), Llama 3 (Meta), Command R+ (Cohere) ou Contrôle (ressources libres : internet, connaissances personnelles). Les groupes expérimentaux conversaient avec le LLM assigné pour évaluer deux scénarios tirés aléatoirement, jusqu'à l'obtention de 600 réponses par groupe. **Issue primaire** : identification correcte d'au moins une condition médicale pertinente (correspondance approximative sur texte libre vs. listes gold standard). **Issue secondaire** : choix correct du recours aux soins sur une échelle à cinq niveaux (automédication à appel de l'ambulance), comparé au consensus médical de référence.

Résultats

Les LLMs seuls ont montré une excellente performance pour l'identification des conditions médicales (94.9 % des cas), et une performance plus modeste pour la prise en charge appropriée (56.3 % des cas). Lorsque ces mêmes LLMs ont été utilisés par les participant·es de l'étude, les performances ont chuté de manière significative : identification correcte des conditions dans moins de 34.5 % des cas et de la conduite appropriée dans moins de 44.2 % des cas. Résultats non supérieurs à ceux du groupe témoin.

Discussion

Les résultats révèlent un décalage important entre les performances théoriques des LLMs et leur utilité réelle pour le grand public. L'étude identifie deux points problématiques : les utilisateurs transmettent des informations incomplètes aux LLMs, et les LLMs, bien que suggérant parfois de bonnes réponses, ne parviennent pas à les

transmettre efficacement. Les mécanismes précis de cet échec (mauvais prompting, absence de cadre critique, biais d'automatisation) restent toutefois difficiles à identifier à partir de ces données. **Points forts** : design rigoureux (étude randomisée, préenregistrée, avec sondages pré- et post-intervention), scénarios cliniques validés par des médecins, comparaison avec groupe témoin. **Limitations** : recrutement via plateforme en ligne, potentiellement surreprésentant des utilisateurs à l'aise avec les outils numériques ; scénarios fictifs (role-playing) pouvant modifier les comportements par rapport à une vraie inquiétude médicale ; résultats limités aux trois modèles testés (susceptibles d'évoluer avec des versions ultérieures ou d'autres LLMs).

Conclusion

La performance technique d'un modèle d'IA ne garantit pas son utilité pratique. Ces résultats plaident pour l'intégration systématique de tests avec de vrais utilisateurs avant tout déploiement public de LLMs dans le domaine de la santé, et invitent à la prudence face à l'enthousiasme actuel pour ces technologies comme substitut au premier recours médical.

Date de publication	Auteurs
10.03.2026	Benoit Delabays