Validación preliminar de MedSearch: un agente conversacional para responder preguntas médicas en tiempo real basadas en evidencia

Authors
Natalia Castaño-Villegas, Maria Camila Villa, Isabella Llano, Jose Zea

Affiliations
Arkangel AI, Bogotá, Colombia

Introducción

La llegada de la inteligencia artificial (IA)a permitido el desarrollo de herramientas avanzadas para la automatización de procesos, apoyo a la decisión, y rápido acceso a la información. En particular los modelos de lenguaje (LLMs), entrenados en grandes fuentes de texto, puede transformar y procesar lenguaje natural para llevar a cabo tareas como traducir, resumir, extraer información, entre otros [1]. En el campo médico se ha evaluado la capacidad de estos agentes para contestar preguntas médicas apoyándose en bases de datos validadas por humanos, con preguntas de selección múltiple (MedQA, MedMCQA) en los que han obtenido puntajes suficientes para pasar exámenes de licenciamiento médico internacionales [2-3]. Sin embargo, uno de los principales retos de su uso en medicina es la característica estática de sus bases de datos y que no suelen estar conectados a la internet, limitando su habilidad para proveer evidencia validada en tiempo real. Para esto diseñamos MedSearch, un agente conversacional especializado en preguntas médicas capaz de buscar en tiempo real en internet, conectando directamente con las referencias de la literatura usada. En este estudio buscamos validar MedSearch con profesionales de la salud en escenarios clínicos y académicos comparando grupos de participantes usando MedSearch y métodos tradicionales. En este póster presentamos algunos resultados preliminares.

Objetivo

Con la investigación buscamos validar MedSearch con profesionales de la salud en escenarios clínicos y académicos comparando grupos de participantes usando MedSearch o métodos tradicionales. En este póster presentamos algunos resultados preliminares.

Metodología

Evaluaremos al agente en tres aspectos principales: 1. Validez de las respuestas, 2. Tiempo promedio de respuesta y número de búsquedas por pregunta (velocidad) y 3. aceptabilidad del modelo por los usuarios. Para esto diseñamos 4 casos clínicos con el apoyo de especialistas, con 4 preguntas cada uno; diagnóstico general, diagnóstico diferencial, investigación o estado del arte y conocimiento general. Reclutamos estudiantes de medicina, médicos generales y especialistas y los asignamos aleatoriamente a uno de dos grupos. El grupo A utilizará la herramienta MedSearch como apoyo en la resolución de los casos y el grupo B usará cualquier método tradicional excluyendo otras plataformas de inteligencia artificial. Se registra para cada participante el tiempo promedio de respuesta y número de búsquedas por pregunta. Para el grupo A también se registra una breve encuesta de aceptabilidad y validez percibida de la herramienta en la escala de 1 a 3 mostrada a continuación.

1. Las respuestas de la herramienta fueron útiles para responder las preguntas
- **(1)** No me ayudaron / **(2)** Me dijeron exactamente lo que ya sabía / **(3)** Si me ayudaron

2. ¿Cuán confiado se siente en la veracidad de las respuestas de la herramienta?
- **(1)** No confiado / **(2)** Ni confiado ni desconfiado / **(3)** Confiado

3. ¿Usaría esta herramienta en su práctica diaria?
- **(1)** Probablemente no / **(2)** No estoy segur@ / **(3)** Probablemente si

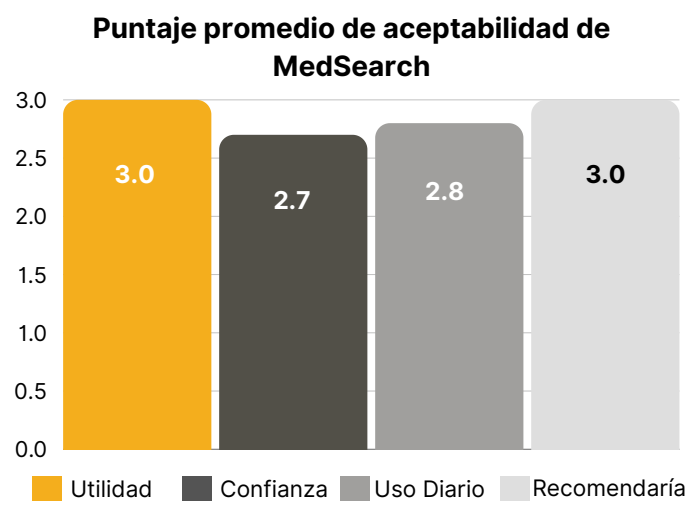
4. ¿Recomendaría esta herramienta a sus compañeros/colegas?
- **(1)** Probablemente no / **(2)** No estoy segur@ / **(3)** Probablemente si

Realizamos un análisis preliminar con 25 participantes (13 Grupo A, 12 Grupo B) que completaron la validación, para la velocidad y aceptabilidad del modelo. los resultados de la validez se evaluarán como un constructo de preguntas dirigidas a evaluar la capacidad real de obtener de MedSearch los objetivos propuestos. Estos serán analizados cuando la prueba haya sido realizada por mínimo 50 participantes.

Resultados

El tiempo promedio de respuesta por pregunta es de 46 segundos para el grupo A y 1.75 minutos para el grupo B que corresponde a una diferencia del 79.8%. El número promedio de búsquedas necesarias por pregunta son 3.81 para Grupo A y 5.38 para Grupo B siendo una diferencia del 34.13%. El puntaje promedio de utilidad y probabilidad de recomendación es de 3 mientras que el promedio de confianza en las preguntas es de 2.7 y de probabilidad de uso diario de 2.8.

Variable	Grupo A (MedSearch)	Grupo B (Tradicional)	Diferencia Porcentual (%)
Tiempo promedio de respuesta por pregunta	00:00:46	00:01:47	79.08
Número promedio de búsquedas por pregunta	3.81	5.38	34.13



Conclusiones

Los resultados preliminares muestra que en promedio los participantes del Grupo A son 79% más rápidos y requieren 34% menos búsquedas para contestar cada pregunta que los participantes del Grupo B. Esto sugiere un posible impacto positivo de usar MedSearch en comparación a métodos tradicionales de busqueda de información. Además, es una herramienta que es percibida como útil con alta probabilidad de ser usada y recomendada. Estos resultados concuerdan con lo hallado por Goh et al, que realizaron una comparación del impacto de usar GPT-4 vs métodos tradicionales y encontraron un tiempo medio de respuesta más rápido y un mejor razonamiento diagnóstico [4].

Bibliografía
1. Liu, N., Chen, L., Tian, X., Zou, W., Chen, K., Cui, M.: From LLM to Conversational Agent: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models, http://arxiv.org/abs/2401.02777, (2024).
2. Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., Chartash, D.: How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med. Educ. 9, e45312 (2023). https://doi.org/10.2196/45312.
3. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nat. Med. 29, 1930–1940 (2023). https://doi.org/10.1038/s41591-023-02448-8.