Can Artificial Intelligence Direct Patients Towards a Complaint-Specific Provider?

Ethan C Gazan, Colin M Emrich, Alexander James Baur, Jenna Alysse Bernstein, David C Landy

INTRODUCTION:

: Many patients use internet search engines to identify a healthcare provider to see for a specific musculoskeletal complaint. With the integration of artificial intelligence (AI) models into search engines as well as the direct use of large language models (LLMs) like ChatGPT to obtain recommendations, we expect these models to increasingly influence patients to specific healthcare providers based on their complaints, whether intentionally or unintentionally. If these models perform well and help optimize the initial triage process, this could increase patient satisfaction as well as the efficient use of healthcare resources. This study sought to evaluate the ability of LLMs to suggest a relevant healthcare provider after being given a representative musculoskeletal patient query.

METHODS:

We selected 3 popular LLMs, ChatGPT, DeepSeek, and Gemini, and evaluated their responses when prompted with a representative patient query. The same prompts were used for each LLM, but requested physicians from two different American cities: Lynchburg, VA, and Trumbull, CT. These cities were selected given the authors' knowledge of providers in these areas. The three prompts were, "*I have low back pain that sometimes travels down my leg,*", "*I hurt my shoulder while lifting a heavy box over my head,*" and "*I keep waking up at night because my thumb and pointer finger are numb.*". Each prompt concluded with, "*I live in (city, state), which doctor should I see?*". If the response did not include physician's names or included less than three names, the prompt "*Can you give me the names of specific doctors in (city)?*" was given. Recommendations were considered appropriate if the provider specialized in the appropriate area and was still practicing. Contact information provided was also assessed for accuracy. Descriptive statistics were used to summarize the results.

RESULTS:

Across all queries and LLMs tested, the proportion of accurate names varied: ChatGPT, 100% (17/17); Gemini, 9/21 (43%); DeepSeek 4/10 (40%). Of the 18 total inappropriate names, 13 (72%) were local providers in non-relevant specialties and 5 (28%) were hallucinations. DeepSeek was the only LLM to hallucinate, inventing half (5) of the names it provided across all prompts. A hallucination occurs when it fabricates the names of providers who do not exist. The proportion of accurate phone numbers also varied: Gemini, 5/6 (83%); ChatGPT, 6/9 (67%); DeepSeek, 2/15 (13%). Most of the incorrect phone numbers provided were for local, non-orthopedic medical offices. The accuracy of the names or phone numbers did not vary based on city.

DISCUSSION AND CONCLUSION:

AI has been shown to recommend appropriate local providers after being given a representative patient query. ChatGPT displayed greater accuracy than both Gemini and DeepSeek across all prompts. Although not requested, each LLM provided multiple phone numbers with each query though these were often inaccurate with Gemini having the best accuracy. DeepSeek hallucinated multiple provider names and phone numbers which is a known issue with AI models. It appears we are nearing a point where AI models can triage patients based on their musculoskeletal complaints. Practices should be aware of this and work to ensure that their contact information is easily accessible and accurate as best possible.