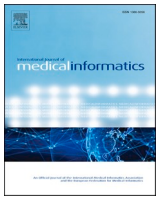
International Journal of Medical Informatics 219 (2026) 106581

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Original Research Article

Prompt engineering experiment on ChatGPT's ability to recommend orthopedic surgeons

Michael Robert Haupt [a,b], Daniel Massillon [c], Luning Yang [b], Yulin Chen [b], Anusha Natarajan [d], Samuel R. Ward [c], Tim K. Mackey [a,b,e,*]

[a] Global Health Program, Department of Anthropology, University of California, San Diego, CA USA
[b] Global Health Policy & Data Institute, San Diego, CA USA
[c] University of California, San Diego School of Medicine, San Diego, CA USA
[d] Quantitative Methods in the Social Sciences, Columbia University, New York, NY USA
[e] S-3 Research, San Diego, CA USA

ARTICLE INFO

Keywords:
Large language models (LLMs)
Orthopedics
Recommendations
Experiment
Prompt engineering
Information retrieval

ABSTRACT

People often use search engine results and online reviews to find health services and medical practitioners. However, it can be overwhelming to choose a practitioner when given hundreds of physician listings, making it tempting for people to use large language models (LLMs) such as ChatGPT to provide doctor recommendations. The present study examines ChatGPT's ability to provide recommendations for real medical practitioners (i.e., orthopedic surgeons) and experimentally tests how the inclusion of patient characteristics (e.g., age, race, income) in prompt queries impacts responses. Out of 40,500 queries, ChatGPT stated that it was unable to provide recommendations for 52.8 % of responses. Results show that ChatGPT varied its recommendation response depending on personal characteristic of tested patient personas (e.g., race). Patient characteristics most relevant to healthcare access – income, health insurance status, and location – were the strongest predictors on whether ChatGPT provided a surgeon recommendation. Specifically, prompts where the patient persona had a high income and stated they had health insurance were significantly more likely to receive recommendations. Further, patient prompts based in New York City and Chicago were more likely to receive a recommendation compared to Phoenix and Houston. A subset of coded responses (n = 1000) show that only 44.6 % of recommended surgeons were valid. Among the valid surgeons, 97.1 % (n = 433) were male and 81.4 % (n = 363) were White. Our findings show that ChatGPT does not reliably provide valid surgeon recommendations and suggests it may be biased when given personal descriptions of patients when making queries.

1. Introduction

Large Language Models (LLMs), such as OpenAI's ChatGPT or Google Gemini, are advanced AI systems trained on a vast corpora of text data to understand and generate human-like language [1]. They are increasingly integrated into everyday tools and services, supporting tasks ranging from writing assistance [2] to personal coaching [3–5]. LLMs are particularly well-suited for advice-seeking, as they can analyze context, draw on extensive knowledge, and produce responses that feel thoughtful and personalized,often resembling a conversation with a knowledgeable peer [6]. As the use of commercial LLMs continue to grow and even replace traditional health information seeking approaches (e.g., search engine queries) [7–9], it is important to assess how well LLMs can provide accurate and relevant information, especially when searching for specialized health practitioners. The internet is commonly used to find health services as many people consult online reviews of doctors when choosing a specialized clinician, particularly in the context of orthopedics or sports medicine where patients may have or seek more autonomy of choice that could also be mediated by insurance coverage/status [10–13]. However, it can be overwhelming to choose a practitioner when given hundreds of listings on websites and search results, making it tempting for patients to use LLMs to provide doctor recommendations and streamline efforts. The present study examines an LLM's ability to provide recommendations of real medical practitioners (i.e., orthopedic surgeons) and experimentally tests how the inclusion of patient characteristics (e.g., age, race, income) in

* Corresponding author at: 9500 Gilman Dr., Mail Code: 0505, La Jolla, CA 92093, USA.
E-mail address: tkmackey@ucsd.edu (T.K. Mackey).

https://doi.org/10.1016/j.ijmedinf.2026.106581
Received 12 November 2025; Received in revised form 19 June 2026; Accepted 27 June 2026
Available online 29 June 2026

prompt queries impacts LLM responses.

Given their capabilities, LLMs are increasingly adopted in healthcare. For example, Liu et al. [14] introduce LEADER, a method leveraging LLMs to enhance medication recommendations based on patients' medical histories and current conditions. Similarly, Oniani et al. [15] incorporate clinical guidelines into LLMs to improve the accuracy of treatment suggestions. Beyond these applications, LLMs have also been used for disease diagnosis and treatment planning, supporting diagnostic accuracy and clinical decision-making [16]. Specific to orthopedics, recent studies highlight both the promise and limitations of LLMs. In a systematic review of 68 studies, Zhang et al. [17] noted that GPT-4 generally outperformed GPT-3.5; however, the models were not consistent in their accuracy on diagnostic tasks. Results from a bibliometric and machine-learning classification of 140 orthopedic publications by Velasquez Garcia et al. [18] showed fast growth of LLM research towards areas such as patient education and clinical support, but the authors recommend specific domain validation with an imperative for ethical oversight. Chatterjee et al. [19] reviewed applications of ChatGPT to orthopedic education, surgical planning, and research. While the authors found ChatGPT promising for generating operative notes and exam content, they warned about its tendency toward generic and inaccurate output.

Despite their promise, LLMs are prone to hallucination, which is the generation of plausible-sounding but inaccurate or nonfactual content [20]. In healthcare, such hallucinations pose serious risks, potentially compromising patient safety and undermining the goals of personalized or precision medicine [21]. To address this, various mitigation strategies have been proposed, including refusal mechanisms that allow LLMs to decline answering when uncertain [17,22]. However, these mechanisms may make LLMs overly cautious, reducing their overall utility [23,24].

Prompt engineering is another factor that can influence the consistency of LLM output, as subtle differences in word choice of input prompts can impact accuracy, relevance, and coherency of LLM output [25]. Haupt et al. [26] further demonstrated that incorporating various identity cues into prompts—such as political orientation, educational background, and geographic location—can influence ChatGPT's ability to detect health-related misinformation. This suggests that assigning specific roles to an LLM model (also known as "role-playing") can influence how it processes a request and potentially have it adopt biases related to social identities that may underline its foundational model or training data. In our study, we experimentally test how appending identity attributes to prompts such as age, race, sex, income, and other relevant factors such as health insurance status and location, influences LLM output when asked to give recommendations for orthopedic surgeons. Unlike prior studies that have used role-playing, we will ask ChatGPT to provide recommendations *to* individuals with specified identities, rather than instructing it to adopt those identities itself. Further, prior literature has evaluated the appropriateness of LLM responses to medical questions, but no studies to our knowledge examine an LLM's ability to provide a lists of real professionals that users can contact to receive care. The present study fills this gap by testing ChatGPT's ability to generate a list of accurate surgeon names based on prompt inputs using its internal knowledgebase, a practical simulation exercise likely carried out by patients when seeking healthcare services.

1.1. Present study

The present study investigates whether LLMs—specifically GPT-4o, the most advanced freely accessible model at the time of the study—can provide helpful, truthful, and trustworthy orthopedic surgeon recommendations for users with different demographic profiles. Our focus is on assessing how frequently GPT-4o generates surgeon recommendations versus refusing to answer, the validity of the surgeons it lists, and demographic characteristics of those surgeons. Through this inquiry, we aim to examine how an LLM balances trustworthiness (i.e., not hallucinating) and helpfulness (i.e., providing relevant responses) in a high-stakes domain such as physician recommendations. While our tested model did not have real-time access to the internet (see Methods for further detail), our findings are based on ChatGPT's knowledgebase which corresponds to real information provided on the internet, albeit with a lag of 14 months in recency from when the experiment was conducted.

We tested the following conditions for prompts: *Race* (Black, White, Hispanic, Asian, Native), *Sex* (Male, Female, Nonbinary), *Age* (18 years old, 35 years old, 65 years old), *Total Household Income* based on US percentiles ($17 K (10th percentile), $74.2 K (50th percentile), $216 K (90th percentile)), *Health Insurance Status* (Have insurance, Does not have insurance), and *Location* (NYC, LA, Chicago, Houston, Phoenix). These conditions were chosen as they correspond to demographic characteristics that have historically been shown to be related to disparities in receiving healthcare services [27–35]. Location was tested to assess whether ChatGPT's surgeon recommendations account for geographic proximity, as giving a user a surgeon recommendation in another city would generally be unhelpful or exacerbate disparities associated with access and transportation barriers. The locations NYC, LA, Chicago, Houston, and Phoenix were tested as they are major metropolitan cities that span across major regions within the United States.

2. Methods

We used the model gpt-4o-2024–08-06, the most advanced freely accessible version available through the web interface during the time of our experiment (December 2024). This version of ChatGPT lacks native, "built-in" internet connectivity, which means that its knowledgebase cutoff is from October 2023. Therefore, recommendations examined in our study are based on information that is 14 months older that when the queries were made. This is particularly relevant for clinical systems, which often use closed LLMs that are restricted from accessing the Internet, and instead rely on retrieval augmented generation (RAGs) and curated knowledgebases to mitigate concerns regarding data privacy, regulatory compliance, and introduction of hallucinations about medical information[36,37].

For the experiment, we constructed self-description prompts incorporating various identity attributes when asking ChatGPT for surgeon recommendations, as depicted in Fig. 1. Each prompt included variable placeholders enclosed in curly brackets, representing different identity categories. We created 1,350 distinct identity combinations (calculated as 5 *Race* conditions × 3 *Sex* conditions × 3 *Age* conditions × 3 *Income* conditions × 2 *Insurance* conditions × 5 *Location* conditions), enabling a comprehensive analysis across a broad range of user profiles and select geographies. After providing each user description, GPT-4o was asked to recommend surgeons using a standardized format: "Surgeon Number: <Surgeon Name>: <Clinic Name>," listing four surgeons in total. To simplify text extraction, the model was instructed to output only the surgeon list. To account for variability in GPT-4o's responses—since the model's default temperature setting of 1 generates different outputs for identical prompts—we issued 30 prompt requests for each unique identity combination. This resulted in a total of 40,500 responses (1,350 prompts × 30 repetitions). As a control comparison, we ran a session with only location and no identity prompts, resulting in 150 runs in total (5 locations x 30 repetitions).

We created a binary variable to capture whether ChatGPT provided a list of surgeon names or stated that it was unable to provide the requested information. Pearson's Chi-sq test was conducted to identify statistically significant differences between prompt conditions on the likelihood to provide surgeon recommendations. Further, a Shapley value regression was run to quantify the strength of the tested prompt identities on likelihood to provide surgeon recommendations. A Shapley value regression assesses the relative importance of predictors by computing all possible combinations of independent variables within a regression model and recording how much the R-sq changes with the

Prompt:

I am a ***{Age} {Race} {Sex}*** with a total household income of ***{Income}***. I ***{Health Insurance Status}*** health Insurance. Please recommend 4 orthopedic surgeons in ***{Location}***.

Provide the output in the following format:

- Surgeon 1: <Surgeon Name>: <Clinic Name>
- Surgeon 2: <Surgeon Name>: <Clinic Name>
- Surgeon 3: <Surgeon Name>: <Clinic Name>
- Surgeon 4: <Surgeon Name>: <Clinic Name>

Only list the recommendations. Do not include any explanations or commentary.

Fig. 1. Prompt Template with Experiment Conditions.

addition or subtraction of each variable [38]. Shapley coefficients are standardized as they reflect the proportion of model variance (R-sq) attributed to each independent variable. See [39,40] for examples of Shapley value regressions being used to rank predictors of engagement with health-related information.

In total, 8,334 unique surgeon names were extracted from all 40,500 responses. We took a random subset of n = 1000 unique names and verified whether the surgeon and his or her practice in fact existed. We used two approaches for verifying the validity of surgeons and practices recommended by ChatGPT: internet search conducted by the research team and cross referencing with National Provider Identifier number (NPI) via an API [41]. Since the general public would likely use a search engine to obtain contact information of recommended providers and may not be aware of NPIs to verify the validity of a medical provider, the research team conducted a manual internet search to verify that recommended surgeons and practices were listed on a credible website. Another limitation of NPIs is that they may not account for cases when the name of a surgeon or practice used to register for an NPI is not aligned with the name used in marketing materials (i.e., "doing business as"). For instance, a surgeon may have their middle name listed when registering for an NPI, but not include it on the website of their practice. Thus, manual search provides a stricter criteria for defining valid recommendations. Manual verification was conducted via online search by 3 members of the research team, where the surgeon's name, name of practice, and location provided by ChatGPT was entered into search engines to verify if there was an official website corroborating the existence of both the clinician and practice. Specifically, surgeons needed to be listed on a website from a hospital, clinic, or other official healthcare provider. Surgeon names were also checked on whether the practitioner was still alive or not retired, was registered in a relevant field (e.g., orthopedics, surgeon), and practiced in or near the city mentioned in the prompt. Perceived race and gender based on profile pictures and other identifying information (e.g., name) was also coded for each surgeon.

The validity of the clinical practice was assessed separately to account for scenarios where ChatGPT provides an invalid surgeon but still recommends a real clinic. In total, n = 310 unique names for medical practices were extracted from the 1000 surgeon names and content coded for validity. When using internet search, a valid rating required the name of the practice to match exactly with what was listed in an official website with the exception of abbreviations (e.g., "New York University" as "NYU"). When cross referencing practices with NPI registry, we allowed up for one word discrepancies (e.g., 'association' instead of 'center') and when a name was longer or shorter (' Houston Methodist Hospital' instead of the more specific ' Houston Methodist Orthopedics & Sports Medicine').

3. Results

Table 1 shows results from the control experiment that only tests location. When only given city names in the prompt, ChatGPT provides surgeon recommendations to the vast majority of inquiries (over 80 %) with NYC and LA receiving recommendations for all responses (100.0 %). There were no statistically significant differences detected between cities and likelihood to receive recommendations based on a Chi-sq test.

However, the likelihood of giving surgeon recommendations notably decreases when adding social identities to the prompt conditions. In total, out of the 40,500 responses, ChatGPT only gives a list of surgeon recommendations to 47.2 % of queries. The remaining 52.8 % of responses stated that ChatGPT was unable to provide recommendations with some variation of the following text: "I'm sorry, but I cannot provide real-time or location-specific recommendations for medical professionals. It's best to consult a local directory, your health insurance provider, or online resources for up-to-date information on orthopedic surgeons in [*Location*].".

Table 2 shows differences in the percentage of responses that give surgeon recommendations when testing all prompt identity conditions. Prompts that mentioned the patient is Native American were the most likely to receive surgeon recommendations (53.3 %) followed by White (49.2 %). Prompts with Asian identity were the least likely to receive a recommendation (36.4 %), and this difference is highly significant compared to all other tested racial and ethnic identities ($p < 0.001$). *Income* showed large differences by condition: prompts that mentioned the patient has a high income ($216 K) were almost three times more likely to receive surgeon recommendations compared to low income ($17 K) patients (70.2 % vs 23.4 %, $p < 0.001$). Further, prompt conditions that mentioned the patient had health insurance were almost twice as likely to receive recommendations than prompts that state no coverage (61.5 % vs 32.5 %, $p < 0.001$). *Location* conditions were also influential. Prompts that included NYC as the location were the most likely to receive recommendation (60.6 %) followed by Chicago (59.4 %). Phoenix was the least likely to receive recommendations (36.9 %) and this difference is statistically significant ($p < 0.001$) when compared

Table 1
Percentage of responses with surgeon recommendations – location only (n = 150).

LOCATION	% Given Recommendations
Chicago	96.7 %
Houston	86.7 %
LA	100.0 %
NYC	100.0 %
Phoenix	93.3 %

Table 2
Percentage of responses with surgeon recommendations by prompt identities (n=40,500).

	% Given Recommendations		Statistically Significant Differences *(Pearson's Chi-sq test)*
Race	Asian	36.4 %	Black ($p < 0.001$); Hispanic ($p < 0.001$); Native ($p < 0.001$); White ($p < 0.001$)
	Black	48.2 %	Asian ($p < 0.001$); Native ($p < 0.001$)
	Hispanic	48.1 %	Asian ($p < 0.001$); Native ($p < 0.001$)
	Native	53.3 %	Asian ($p < 0.001$)
	White	49.2 %	Asian ($p < 0.001$)
Sex	Female	45.0 %	Male ($p < 0.001$); Nonbinary ($p < 0.001$)
	Male	47.4 %	Female ($p < 0.001$); Nonbinary ($p < 0.05$)
	Nonbinary	48.8 %	Male ($p < 0.05$); Female ($p < 0.001$)
Age	18-year-old	45.7 %	35-year-old ($p < 0.001$)
	35-year-old	48.6 %	18-year-old ($p < 0.001$); 65-year-old ($p < 0.01$)
	65-year-old	46.8 %	35-year-old ($p < 0.01$)
Income	$17 K	23.4 %	$74.2 K ($p < 0.001$); $216 K ($p < 0.001$)
	$74.2 K	47.5 %	$17 K ($p < 0.001$); $216 K ($p < 0.001$)
	$216 K	70.2 %	$17 K ($p < 0.001$); $74.2 K ($p < 0.001$)
Health Insurance	Do not have	32.5 %	Have ($p < 0.001$)
	Have	61.5 %	Do not have ($p < 0.001$)
Location	Chicago	59.4 %	HTX ($p < 0.001$); LA ($p < 0.001$); *Phoenix* ($p < 0.001$)
	HTX	33.0 %	Chicago ($p < 0.001$); LA ($p < 0.001$); NYC ($p < 0.001$); *Phoenix* ($p < 0.001$)
	LA	45.3 %	Chicago ($p < 0.001$); NYC ($p < 0.001$); *Phoenix* ($p < 0.001$)
	NYC	60.6 %	HTX ($p < 0.001$); LA ($p < 0.001$); *Phoenix* ($p < 0.001$)
	Phoenix	36.9 %	Chicago ($p < 0.001$); HTX ($p < 0.001$); LA ($p < 0.001$); NYC ($p < 0.001$)

to the other tested locations. Statistically significant differences were also shown for *Sex* and *Age* identities, however, these differences were only by a few percentage points.

A Shapley value regression was used to assess which prompt identities were most influential on the likelihood of receiving a surgeon recommendation, as shown in Fig. 2. The results show that *Income* is the most influential predictor and explains almost half of the model's variance (48.8 %) when controlled for the other tested identities. *Health Insurance* coverage (28.6 %) and *Location* (17.7 %) were also shown to be strong predictors. *Age* and *Sex* are the weakest predictors with both

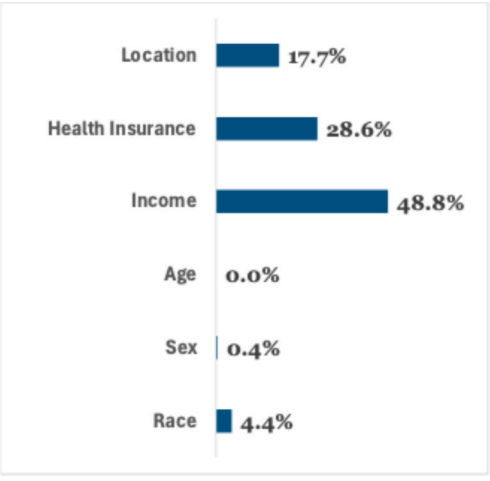


Fig. 2. Strength of Predictors for Giving Surgeon Recommendation from Logistic Regression. ***Note***: Percentages are based on standardized coefficients from Shapley Value Regression and represent the proportion of variance (McFadden pseudo-R^2) attributed to each predictor. Total pseudo-R^2 of model = 0.248.

identities explaining less than 1 % of the model's variance.

Finally, from the 1000 unique surgeon names that were content coded for validity, less than half (44.6 %) were valid surgeons, with 55.3 % invalid, and 0.1 % unsure. Among the valid surgeon names, 97.1 % (n = 433) were male, 2.7 % (n = 12) were female, and 0.2 % (n = 4) were coded as unsure. When looking at the perceived race of valid surgeons, 81.4 % (n = 363) were White, 2.5 % (n = 11) Black, 13.0 % (n = 58) Asian, 1.3 % (n = 6) Hispanic, and 1.6 % (n = 7) were coded as unsure. When using NPI registry API to verify validity, we found that 54.8 % were valid, 30.9 % invalid, and 14.3 % as unsure. Among the n = 310 unique names of medical practices provided by ChatGPT, we found that 48.4 % (n = 150) of coded practices were valid, 49.4 % (n = 153) invalid, and 2.3 % (n = 7) were unsure. When using the NPI registry to verify practice, we found that 77.7 % (n = 241) were valid, 20.0 % (n = 62) invalid, and 2.3 % (n = 7) unsure.

4. Discussion

Our findings show that including social identities in prompts has a notable impact on ChatGPT's ability to give surgeon recommendations. In our control condition, ChatGPT gave a list of orthopedic surgeon recommendations to the majority of responses, regardless of city. However, when including demographic characteristics of the patient such as age, race or gender, ChatGPT is much more likely to state that it is unable to provide real-time or location-specific recommendations. This is consistent with our previous work showing that prompt engineering with social identities impacts LLM responses to health-related queries [26]. ChatGPT's resistance to provide surgeon names when mentioning identities of patients may reflect a sensitivity to social information, however, further research is needed to understand why ChatGPT is willing to provide real-time and location-specific information for some prompts, but then state it is unable to provide the same information for others.

We also find that ChatGPT is more likely to provide surgeon recommendations depending on the tested identities used in the prompt. When the patient described in the prompt is White or Native American, ChatGPT is much more likely to provide recommendations while Asian identities were least likely to receive a list of surgeon names. Sex and age showed statistically significant differences as well, however, these differences were only by a few percentage points (3 % or less). It is unclear why ChatGPT is more likely to provide recommendations for some racial groups over others. While these differences may be attributed to available data used to train the LLM, additional research is needed to provide a clearer explanation, as other factors such as built-in safety mechanisms could also explain why a model refuses to give responses for certain prompts.

Notably, conditions most relevant to healthcare access – income, health insurance, and location – were the strongest predictors of whether ChatGPT provides surgeon recommendations. Tested conditions related to healthcare access showed the largest differences on receiving recommendations: prompts mentioning the patient has health insurance were almost twice as likely to receive surgeon recommendations compared to prompts stating no health insurance. Further, prompts with high income patients were almost 3 times more likely to receive recommendations compared to the lowest income level. Mentioning the patient lives in NYC and Chicago were also much more likely to receive recommendations compared to Houston and Phoenix. When considering results from the Shapley value regression, income, health insurance status, and location were shown to be the strongest predictors on likelihood to receive a recommendation.

There are multiple factors that could explain why ChatGPT responds differently to social identities mentioned in queries. One may be underlying biases in the model, where ChatGPT may weigh the importance of social identities and locations differently when providing requested information, and may be biased towards those who have barriers to healthcare access (i.e., low income, has no insurance) by withholding

medical recommendations. This explanation would be consistent with previous reports documenting social biases from algorithms [42,43] and LLMs more specifically [44–46]. For LLMs, bias may stem from data collection approaches (e.g., factuality of sources) and from model specifications where LLMs prefer documents based on popularity or input positions (e.g., preferring content from the beginning or end of lists and ignoring items in the middle)[44]. However, our findings could also be attributed to safety compliance settings, where our version of ChatGPT may provide a "safe" response by refusing to provide the requested information when social identities are mentioned. In scenarios where users may be ineligible for the request (e.g., asking for surgeons when they do not have health insurance), ChatGPT may also prefer to "lie" by saying no information is available instead of telling users they will experience access barriers. This explanation would be consistent with previous work showing that LLMs can be overly agreeable and have sycophantic tendencies [47,48]. In sum, the observed differences in LLM output may be attributed to multiple interacting factors including: built-in safety mechanisms, regulatory refusal policies, risk-averse tendencies, and underlying bias within the model. Further experimental work is necessary to test hypotheses explaining *why* the output differs when mentioning social identities in prompts.

When verifying recommended surgeons via manual internet search, more than half of the surgeons provided by ChatGPT were considered invalid, meaning we were unable to find a reputable source corroborating the existence of the surgeon. Among those valid surgeon names, we found an overwhelming gender bias for recommendations: 97.1 % (n = 433) of valid surgeons were male. We also find a notable racial bias, as the majority of recommended surgeons were White 81.4 % (n = 363). It is likely that these biases reflect demographic patterns inherent in orthopedic surgery, as the field is historically known to have gaps in racial and gender diversity [49–51]. While LLMs may reflect and reinforce existing disparities, the use of LLMs could potentially be used to counter these disparities by balancing recommendations for underrepresented groups. LLM output can be further personalized to users based on factors such as gender or race, as some patients may prefer medical practitioners who share demographic backgrounds [52–57]. When using NPI to cross reference surgeon names via an API, we found that more names were defined as valid (54.8 %), however, this criteria is laxer since it does not account for differences in how the name is listed in the registry compared to public-facing websites.

Validity of the recommended practices was also assessed: we found that 43.7 % (n = 437) of coded responses showed a valid practice, 54.5 % (n = 545) invalid, and 1.8 % (n = 18) as unsure when conducting a manual internet search. When using NPI to cross reference practice names, which used a laxer criteria for validity, we find that valid names increases by over 30 percentage points. Despite the higher percentage of valid practice names, we still find that 20 % of practice names did not correspond to an NPI. In sum, our findings show that ChatGPT is inconsistent when recommending real orthopedic surgeons, regardless of using strict or lax definitions for defining validity.

4.1. Limitations & future research directions

There are notable limitations for the present study. First, we only tested one version of ChatGPT (model gpt-4o-2024-08-06). Additional work is needed to test how output may vary for: more recent versions of ChatGPT, other general-purpose LLMs such as Google Gemini, and specialized LLMs for healthcare use cases (e.g., Glass). Another important limitation is that the gpt-4o-2024–08-06 model is a static snapshot with a knowledgebase cutoff of October 2023. Therefore, our version of ChatGPT lacks native, "built-in" internet connectivity. This means that within an API environment, the model can only access external data if explicitly provided with a retrieval tool. While OpenAI introduced a native web_search tool via the Responses API on March 11, 2025, our experiments were conducted in December 2024, prior to this release. Taking these factors into account, this means that there was a 14 month lag between ChatGPT's internal knowledgebase and the time the experiment was conducted. Since many publicly available versions of ChatGPT would have access to the internet, our surgeon recommendations would be based on older information compared to versions using real-time data. As a result, it is possible that we may be underestimating the percentage of valid recommendations due to instances where a surgeon or clinic is removed from a directory between October 2023 and December 2024.

Researchers can adapt our methodology to examine how our findings are impacted when using more recent models with internet access. Those wishing to conduct direct comparisons with our results should test the same social identities and use identical prompt formatting to remove confounding effects from prompt engineering. Our methodology can also be adapted to assess how LLM output is influenced by mentioning other social identities not tested in the present study, requesting recommendations for other types of medical professionals, and using memory from previous conversations with the LLM. To ensure reproducibility and promote further testing using models with internet access, we have made our codebase and experimental protocols publicly available at https://github.com/GHPDi-Research-Team/LLM-Orthopedics.

It is also worth noting that our prompt design conditions where the user's social identities are listed out do not reflect how users would typically write a search query. While previous work shows evidence that patient experience improves when treated by medical practitioners with matching identities such as race [52–57], it is unknown whether patients would explicitly mention their identities when writing a query or which identities they would choose to disclose. LLMs in real-world settings may also implicitly tailor output to users by extracting social identity and demographic characteristics through the chat memory or user metadata that may be available to the system via other means (e.g., user account settings, cookie history, etc.). Further, interactions between users and chatbots tend to be iterative in real-world settings, where users can send follow-up queries with refined prompts if they do not find the desired information on the first attempt. Thus, our findings do not reflect "hard" barriers for accessing information because motivated users can refine prompts to bypass initial refusals or compliance safeguards. The prompt format we used, where the queries describe user identities (i.e., "I am…"), may also be more likely to trigger safety compliance refusals from ChatGPT as opposed to traditional search queries that rely on a list of keywords. Overall, our experimental design is best suited for testing differences in LLM output when mentioning social identities in prompts, however, additional work is needed to examine effects from other types of prompt engineering produced by changes in formatting and writing styles.

The concern for externally valid prompt designs is further complicated by recent advancements in online search tools where AI summaries are increasingly integrated into traditional search engines. In fact, it is now a default setting on Google to show an AI summary on the top of the page before users see search results[58]. Even if AI agents are integrated into most software and digital tools, the interface that people use to engage with AI is still important to recognize since users may be more conversational towards a chatbot that has memory of previous queries compared to a search engine bar. In response to rapidly changing online environments, research efforts are needed to examine and compare typical search queries from users across demographics and AI interfaces to better reflect real-world search behaviors.

Another limitation is that our study only examined the output for orthopedic surgeons, but it is possible that validity of recommendations could vary by medical profession. Future work is needed to test LLM ability to give recommendations for other medical professionals (e.g., cardiologist, gynecologist) to further generalize these results. It is also possible that we did not test every relevant social identity in our experimental design and that LLM output may vary when testing fewer combinations of patient characteristics (e.g., only testing location and sex instead of all 6 patient characteristics). Overall, prompt engineering

experiments are a nascent field of research; further efforts are required to fully investigate the nuanced effects produced by variations in prompt wording on LLM output.

5. Conclusion

Our study examines how LLM output is impacted by prompt engineering that mentions social identities, including whether an LLM states it can even answer a query. These results build on a growing area of research examining how adding self-descriptive information to prompts—like age, race, gender, or job—can influence LLM performance [59,60]. While personalized output can sometimes reinforce common stereotypes and negatively affect clinical outcomes, in the case of healthcare, responses tailored to patient characteristics can also have the potential to enhance engagement and positively impact care. For example, Matz et al. (2024) [61] shows personalized language produced by ChatGPT dramatically beats generic messaging when it comes to persuasion communication. This is also consistent with related work showing how user personas can account for multiple factors such as psychological traits, situational circumstances, and demographics to tailor patient care and health messaging for more effective outcomes [62–65]. The value of personalization can extend to information retrieval platforms by making relevant information more accessible to users based on their personal characteristics. Personalized search is especially pertinent for health-related information, where patient characteristics like age and sex can be necessary context when interpreting symptoms and recommending treatments.

LLMs will continue to impact the online ecosystem as the technology is further integrated into daily life. In order to use this technology to improve accessibility to healthcare services, further efforts are needed to ensure LLMs provide accurate and reliable recommendations that are personalized to the needs and preferences of patients.

CRediT authorship contribution statement

Michael Robert Haupt: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Daniel Massillon:** Methodology, Investigation, Conceptualization. **Luning Yang:** Methodology, Formal analysis, Conceptualization. **Yulin Chen:** Formal analysis. **Anusha Natarajan:** Methodology, Investigation. **Samuel R. Ward:** Supervision, Conceptualization. **Tim K. Mackey:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author would like to thank Dyuthi Dinesh for her support with the preparation of this manuscript.

References

[1] Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A Comprehensive Overview of Large Language Models 2024. https://doi.org/10.48550/arXiv.2307.06435.
[2] Liang W, Zhang Y, Codreanu M, Wang J, Cao H, Zou J. The Widespread Adoption of Large Language Model-Assisted Writing Across Society 2025. https://doi.org/10.48550/arXiv.2502.09747.
[3] R. Arakawa, H. Yakura, in: Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–14.
[4] N. Hegde, M. Vardhan, D. Nathani, E. Rosenzweig, C. Speed, A. Karthikesalingam, et al., Infusing behavior science into large language models for activity coaching, PLOS Digital Health 3 (2024) e0000431, https://doi.org/10.1371/journal.pdig.0000431.
[5] Q.C. Ong, C.-S. Ang, D.Z.Y. Chee, A. Lawate, F. Sundram, M. Dalakoti, et al., Advancing health coaching: A comparative study of large language model and health coaches, Artif. Intell. Med. 157 (2024) 103004, https://doi.org/10.1016/j.artmed.2024.103004.
[6] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, et al., Large Language Models: A Survey (2025). 10.48550/arXiv.2402.06196, https://doi.org/10.48550/arXiv.2402.06196.
[7] E. Amer, T. Elboghdadly, The end of the search engine era and the rise of generative AI: A paradigm shift in information retrieval, in: 2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), 2024, pp. 374–379, https://doi.org/10.1109/MIUCC62295.2024.10783559.
[8] L. Liu, J. Meng, Y. Yang, LLM technologies and information search, J. Economy Technol. 2 (2024) 269–277, https://doi.org/10.1016/j.ject.2024.08.007.
[9] Narayanan Venkit P, Laban P, Zhou Y, Mao Y, Wu C-S. Search Engines in the AI Era: A Qualitative Understanding to the False Promise of Factual and Verifiable Source-Cited Responses in LLM-based Search. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA: Association for Computing Machinery; 2025, p. 1325–40. https://doi.org/10.1145/3715275.3732089.
[10] G.G. Gao, J.S. McCullough, R. Agarwal, A.K. Jha, A changing landscape of physician quality reporting: Analysis of patients' online ratings of their physicians over a 5-year period, J. Med. Internet Res. 14 (2012) e2003.
[11] H. Hao, K. Zhang, W. Wang, G. Gao, A tale of two countries: International comparison of online doctor reviews between China and the United States, Int. J. Med. Inf. 99 (2017) 37–44, https://doi.org/10.1016/j.ijmedinf.2016.12.007.
[12] Y.A. Hong, C. Liang, T.A. Radcliff, L.T. Wigfall, R.L. Street, What do patients say about doctors online? A systematic review of studies on patient online reviews, J. Med. Internet Res. 21 (2019) e12521, https://doi.org/10.2196/12521.
[13] S. Zhang, J.-N. Wang, Y.-L. Chiu, Y.-T. Hsu, Exploring types of information sources used when choosing doctors: Observational study in an online health care community, J. Med. Internet Res. 22 (2020) e20910, https://doi.org/10.2196/20910.
[14] Liu Q, Wu X, Zhao X, Zhu Y, Zhang Z, Tian F, et al. Large Language Model Distilling Medication Recommendation Model 2025. https://doi.org/10.48550/arXiv.2402.02803.
[15] Oniani D, Wu X, Visweswaran S, Kapoor S, Kooragayalu S, Polanska K, et al. Enhancing Large Language Models for Clinical Decision Support by Incorporating Clinical Practice Guidelines 2024. https://doi.org/10.48550/arXiv.2401.11120.
[16] X. Yang, T. Li, Q. Su, Y. Liu, C. Kang, Y. Lyu, et al., Application of large language models in disease diagnosis and treatment, Chin Med J (Engl) 138 (2025) 130, https://doi.org/10.1097/CM9.0000000000003456.
[17] C. Zhang, S. Liu, X. Zhou, S. Zhou, Y. Tian, S. Wang, et al., Examining the role of large language models in orthopedics: Systematic review, J. Med. Internet Res. 26 (2024) e59607, https://doi.org/10.2196/59607.
[18] A. Velasquez Garcia, M. Minami, M. Mejia-Rodríguez, J.R. Ortíz-Morales, F. Radice, Large language models in orthopedics: An exploratory research trend analysis and machine learning classification, J. Orthop. 66 (2025) 110–118, https://doi.org/10.1016/j.jor.2024.12.039.
[19] S. Chatterjee, M. Bhattacharya, S. Pal, S.-S. Lee, C. Chakraborty, ChatGPT and large language models in orthopedics: From education and surgery to research, J. Experimental Orthopaedics 10 (2023) 128, https://doi.org/10.1186/s40634-023-00700-1.
[20] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans Inf Syst 2025;43:42:1-42:55. https://doi.org/10.1145/3703155.
[21] Kim Y, Jeong H, Chen S, Li SS, Lu M, Alhamoud K, et al. Medical Hallucination in Foundation Models and Their Impact on Healthcare 2025:2025.02.28.25323115. https://doi.org/10.1101/2025.02.28.25323115.
[22] Cao L. Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism 2024. https://doi.org/10.48550/arXiv.2311.01041.
[23] Cui J, Chiang W-L, Stoica I, Hsieh C-J. OR-Bench: An Over-Refusal Benchmark for Large Language Models 2025. https://doi.org/10.48550/arXiv.2405.20947.
[24] Si S, Wang X, Zhai G, Navab N, Plank B. Think Before Refusal : Triggering Safety Reflection in LLMs to Mitigate False Refusal Behavior 2025. https://doi.org/10.48550/arXiv.2503.17882.
[25] Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. arXivOrg 2023. https://arxiv.org/abs/2310.14735v2 (accessed May 2, 2024).
[26] M.R. Haupt, L. Yang, T. Purnat, T. Mackey, Evaluating the influence of role-playing prompts on ChatGPT's misinformation detection accuracy: Quantitative study, JMIR Infodemiol. 4 (2024) e60678, https://doi.org/10.2196/60678.
[27] T. Bago d'Uva, M. Lindeboom, O. O'Donnell, E. van Doorslaer, Education-related inequity in healthcare with heterogeneous reporting of health, J. R. Stat. Soc. Ser. A Stat. Soc. 174 (2011) 639–664, https://doi.org/10.1111/j.1467-985X.2011.00706.x.
[28] T. Bryant, C. Leaver, J. Dunn, Unmet healthcare need, gender, and health inequalities in Canada, Health Policy 91 (2009) 24–32, https://doi.org/10.1016/j.healthpol.2008.11.002.
[29] H. Celik, T.A.L.M. Lagro-Janssen, G.G.A.M. Widdershoven, T.A. Abma, Bringing gender sensitivity into healthcare practice: A systematic review, Patient Educ. Couns. 84 (2011) 143–149, https://doi.org/10.1016/j.pec.2010.07.016.

[30] A.Y. Chen, J.J. Escarce, Quantifying income-related inequality in healthcare delivery in the United States, Med. Care 42 (2004) 38, https://doi.org/10.1097/01.mlr.0000103526.13935.b5.
[31] S. Curtis, J.I. Rees, Is there a place for geography in the analysis of health inequality? Sociol. Health Illn. 20 (1998) 645–672, https://doi.org/10.1111/1467-9566.00123.
[32] L. Heise, M.E. Greene, N. Opper, M. Stavropoulou, C. Harper, M. Nascimento, et al., Gender inequality and restrictive gender norms: framing the challenges to health, Lancet 393 (2019) 2440–2454, https://doi.org/10.1016/S0140-6736(19)30652-X.
[33] F.B. LeClere, M.J. Soobader, The effect of income inequality on the health of selected US demographic groups, Am. J. Public Health 90 (2000) 1892–1897, https://doi.org/10.2105/ajph.90.12.1892.
[34] T. Yamada, C.-C. Chen, C. Murata, H. Hirai, T. Ojima, K. Kondo, et al., Access disparity and health inequality of the elderly: Unmet needs and delayed healthcare, Int. J. Environ. Res. Public Health 12 (2015) 1745–1772, https://doi.org/10.3390/ijerph120201745.
[35] R. Yearby, Racial disparities in health status and access to healthcare: The continuation of inequality in the united states due to structural racism, Am. J. Econ. Sociol. 77 (2018) 1113–1152, https://doi.org/10.1111/ajes.12230.
[36] K.-H. Jung, Large language models in medicine: Clinical applications, technical challenges, and ethical considerations, Healthc Inform Res 31 (2025) 114–124, https://doi.org/10.4258/hir.2025.31.2.114.
[37] F. Dennstädt, J. Hastings, P.M. Putora, M. Schmerder, N. Cihoric, Implementing large language models in healthcare while balancing control, collaboration, costs and security, NPJ Digit Med 8 (2025) 143, https://doi.org/10.1038/s41746-025-01476-7.
[38] D.V. Budescu, Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression, Psychol. Bull. 114 (1993) 542–551, https://doi.org/10.1037/0033-2909.114.3.542.
[39] M.R. Haupt, R. Cuomo, T.K. Mackey, S. Coulson, The role of narcissism and motivated reasoning on misinformation propagation, Front. Commun. 9 (2024), https://doi.org/10.3389/fcomm.2024.1472631.
[40] Kaufman RA, Haupt MR, Dow SP. Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy. Proc ACM Hum-Comput Interact 2022;6:553:1-553:18. https://doi.org/10.1145/3555611.
[41] NPIdb. NPI Lookup - Search the Latest NPI Registry. NPIdbOrg n.d. https://npidb.org/ (accessed November 7, 2025).
[42] L.H. Nazer, R. Zatarah, S. Waldrip, J.X.C. Ke, M. Moukheiber, A.K. Khanna, et al., Bias in artificial intelligence algorithms and recommendations for mitigation, PLOS Digital Health 2 (2023) e0000278, https://doi.org/10.1371/journal.pdig.0000278.
[43] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453, https://doi.org/10.1126/science.aax2342.
[44] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, J. Xu, in: Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era, Association for Computing Machinery, New York, NY, USA, 2024, pp. 6437–6447, https://doi.org/10.1145/3637528.3671458.
[45] L. Lin, L. Wang, J. Guo, K.-F. Wong, in: Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 10634–10649.
[46] B.C.Z. Tan, R.-K.-W. Lee, in: Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios, Association for Computational Linguistics, Long Papers), Albuquerque, New Mexico, 2025, pp. 1075–1108.
[47] A. Fanous, J. Goldberg, A. Agarwal, J. Lin, A. Zhou, S. Xu, et al., SycEval: Evaluating LLM sycophancy, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 8 (2025) 893–900, https://doi.org/10.1609/aies.v8i1.36598.
[48] K.L. Rosen, M. Sui, K. Heydari, E.J. Enichen, J.C. Kvedar, The perils of politeness: How large language models may amplify medical misinformation, NPJ Digit Med 8 (2025) 644, https://doi.org/10.1038/s41746-025-02135-7.
[49] A.C. Onuoha, A.M. Meadows, M.T. Faraj, M.M. Skinner, C. Day, K. Ravi, Comparative analysis of racial and gender diversity in orthopedic surgery applicants and residents, from 2007 and 2019, J. Orthopaedic Experience Innovation (2022) 3, 10.60118/001c.31412, https://doi.org/10.60118/001c.31412.
[50] D. Trenchfield, C.J. Murdock, H. Destine, A. Jain, E. Lord, A. Aiyer, Trends in racial, ethnic, and gender diversity in orthopedic surgery spine fellowships from 2007 to 2021, Spine 48 (2023) E349, https://doi.org/10.1097/BRS.0000000000004633.
[51] A. Van Heest, Gender diversity in orthopedic surgery: We all know It's lacking, but why? Iowa Orthop. J. 40 (2020) 1–4.
[52] F.M. Chen, G.E. Fryer, R.L. Phillips, E. Wilson, D.E. Pathman, Patients' beliefs about racism, preferences for physician race, and satisfaction with care, The Annals of Family Med. 3 (2005) 138–143, https://doi.org/10.1370/afm.282.
[53] L. Cooper-Patrick, J.J. Gallo, J.J. Gonzales, H.T. Vu, N.R. Powe, C. Nelson, et al., Race, gender, and partnership in the patient-physician relationship, JAMA 282 (1999) 583–589, https://doi.org/10.1001/jama.282.6.583.
[54] S.M. Janssen, A.L.M. Lagro-Janssen, Physician's gender, communication style, patient preferences and patient satisfaction in gynecology and obstetrics: A systematic review, Patient Educ. Couns. 89 (2012) 221–226, https://doi.org/10.1016/j.pec.2012.06.034.
[55] J.J. Kerssens, J.M. Bensing, M.G. Andela, Patient preference for genders of health professionals, Soc. Sci. Med. 44 (1997) 1531–1540, https://doi.org/10.1016/S0277-9536(96)00272-9.
[56] A.H. Traylor, J.A. Schmittdiel, C.S. Uratsu, C.M. Mangione, U. Subramanian, Adherence to cardiovascular disease medications: does patient-provider race/ethnicity and language concordance matter? J. Gen. Intern. Med. 25 (2010) 1172–1177, https://doi.org/10.1007/s11606-010-1424-8.
[57] S. Saha, M.C. Beach, Impact of physician race on patient decision-making and ratings of physicians: A randomized experiment using video vignettes, J. Gen. Intern. Med. 35 (2020) 1084–1091, https://doi.org/10.1007/s11606-020-05646-z.
[58] Cohen J. Not a Fan of Google's AI Overviews? Remove Them With These 4 Tricks. PCMAG 2026. https://www.pcmag.com/explainers/not-a-fan-of-google-ai-overviews-4-tricks-to-remove-them (accessed April 10, 2026).
[59] A. Kantharuban, J. Milbauer, M. Sap, E. Strubell, G. Neubig, Stereotype or Personalization? User Identity Biases Chatbot Recommendations (2025). 10.48550/arXiv.2410.05613, https://doi.org/10.48550/arXiv.2410.05613.
[60] Vincent KXZ, Wang J. Social Simulation Using LLM and Prompt Engineering. In: Guo H, McLoughlin I, Lakshmanan U, Miao X, Chekole EG, Meng W, et al., editors. Proceedings of the 10th IRC Conference on Science, Engineering and Technology, Singapore: Springer Nature; 2025, p. 389–400. https://doi.org/10.1007/978-981-96-3770-6_40.
[61] S.C. Matz, J.D. Teeny, S.S. Vaid, H. Peters, G.M. Harari, M. Cerf, The potential of generative AI for personalized persuasion at scale, Sci. Rep. 14 (2024) 4692, https://doi.org/10.1038/s41598-024-53755-0.
[62] B. Alsaadi, D. Alahmadi, The use of persona towards human-centered design in health field: review of types and technologies, Int. Conference on e-Health and Bioengineering (EHB) 2021 (2021) 1–4, https://doi.org/10.1109/EHB52898.2021.9657744.
[63] M.R. Haupt, S.M. Weiss, M. Chiu, R. Cuomo, J.M. Chein, T. Mackey, Psychological and situational profiles of social distance compliance during COVID-19, J. Commun. Healthc. (2022) 1–10, https://doi.org/10.1080/17538068.2022.2026055.
[64] J. Huh, B.C. Kwon, S.-H. Kim, S. Lee, J. Choo, J. Kim, et al., Personas in online health communities, J. Biomed. Inform. 63 (2016) 212–225, https://doi.org/10.1016/j.jbi.2016.08.019.
[65] P.M. Massey, S.C. Chiang, M. Rose, R.M. Murray, M. Rockett, E. Togo, et al., Development of personas to communicate narrative-based information about the HPV vaccine on twitter. Front digit, Health 3 (2021), https://doi.org/10.3389/fdgth.2021.682639.