

DOI: https://doi.org/10.48009/3_iis_2026_142

From sample selection bias to algorithmic bias: understanding how clinical trial underrepresentation shapes healthcare LLM outputs

Loreen Powell, *Marywood University*, lpowell@marywood.edu

Gwendolyn Powell, *Philadelphia College of Osteopathic Medicine*, gp4345@pcom.edu

Abstract

Large language models (LLMs) are increasingly used to support health information seeking and decision-making. However, these systems are trained on data sources that may reflect historical inequities for minorities. This paper examines how the underrepresentation of racial and ethnic minorities in clinical trials can contribute to algorithmic bias in healthcare LLMs. The study also provides prompt engineering framework to overcome the LLM health information inequities.

Keywords: large language models, bias, health, minorities, prompt engineering framework

Introduction

The digital health information landscape has expanded considerably over the past decade. Individuals increasingly seek health-related information on topics such as nutrition (Powell et al., 2026), disease prevention, symptom assessment, medical conditions, and treatment options through online resources (Yun et al., 2025). More recently, health information has become accessible through artificial intelligence (AI) based large language models (LLMs), including ChatGPT, Google Gemini, and Claude (Nadarzynski et al., 2023; Nadarzynski et al., 2024). Olsen (2026) reports that one in three Americans uses AI for health information. Hence, LLMs are becoming increasingly influential in shaping how individuals' access, interpret, and apply healthcare information (Omiye et al., 2024; Turchioe et al., 2026; Yun et al., 2025). LLMs offer several potential benefits for seeking health information. They provide immediate access to health information, assist users in understanding medical terminology, exploring symptoms and treatment options, and preparing questions for healthcare providers (Costa-Gomes et al., 2026; Singhal et al., 2023). These capabilities may be particularly valuable for individuals residing in medically underserved areas and those with limited access to healthcare services. Additionally, AI-powered chatbots have demonstrated the potential to improve health literacy, support chronic disease management, facilitate patient-provider communication, and promote patient engagement and self-management (Barreda et al., 2025; Kurniawan et al., 2024).

Despite these LLM benefits, concerns remain regarding the reliability, safety, and equity of healthcare-related outputs (Hanna et al., 2025; Omiye et al., 2023; Singhal et al., 2023; Wu et al., 2025; Yun et al., 2025). A systematic review by Yun et al. (2025) found that LLMs may generate inaccurate, incomplete, or potentially unsafe medical information, including responses that omit clinically relevant details, conflict with established clinical guidelines, or present hallucinated information with a high degree of confidence. Because patients may rely on these outputs when making health-related decisions, such inaccuracies may

contribute to inappropriate self-diagnosis, delayed care, poor treatment decisions, and adverse health outcomes.

While concerns regarding factual inaccuracies and hallucinations have received substantial attention, less research has examined whether historical disparities in clinical research participation contribute to knowledge gaps within LLM health-related outputs. In particular, the longstanding underrepresentation of racial and ethnic minority populations in clinical trials may result in output disparities that become embedded within the biomedical literature used to train LLMs, potentially affecting the accuracy and representativeness of health information generated for underserved populations (Hanna et al., 2025; Omiye et al., 2023).

This paper makes two primary contributions. First, it proposes a conceptual explanation linking historic racial and ethnic minority underrepresentation in clinical trials to algorithmic bias in healthcare LLM outputs. Second, it introduces a two-stage prompt engineering recommendation for racial and ethnic minority users seeking health information from LLMs. These recommendations are not intended to replace professional medical advice but rather to encourage critical evaluation of AI-generated information and awareness of potential limitations within LLM health information outputs.

Literature Review

Language Models & Large Language Models

Language models (LMs) are AI-based computational models that can understand, interpret, and generate human language. They are a form of generative AI (GenAI) systems which understand and generate human language as they are trained on extensive text corpora to identify, generate, summarize, translate, and respond to natural language (Chang et al., 2024).

LLMs for Health Information

LLMs were not originally designed, clinically validated, or formally tested for medical decision-making purposes (Markman, 2026) and they were not intended to serve as substitutes for professional medical advice, diagnosis, or treatment provided by qualified healthcare practitioners (Shah et al., 2024; Singhal et al., 2023). However, today, “health is among the most high-stakes domains in which people interact with conversational artificial intelligence (AI)” (Costa-Gomes et al., 2026).

LLMs are popular across numerous domains, including healthcare, for many reasons. For example, Kurniawan et al. (2024) found that AI-powered chatbots can support chronic disease management, facilitate patient-provider communication, deliver health education, and encourage positive behavioral changes. Similarly, Barreda et al. (2025) report that healthcare chatbots can assist individuals with appointment management, medication reminders, health promotion, and disease-specific education while providing information in a conversational and user-friendly format. These capabilities may empower patients to take a more active role in their healthcare by increasing their understanding of medical conditions and treatment recommendations (Barreda et al., 2025; Kurniawan et al., 2024).

Furthermore, for individuals facing financial, geographic, or structural barriers to healthcare access, LLMs such as ChatGPT may serve as a readily available and low-cost source of health information, symptom interpretation, medical explanations, and treatment guidance when professional healthcare services are inaccessible or delayed (Blease et al., 2024; Turchioe et al., 2026). This reliance may be particularly pronounced among underserved and racial and ethnic minority populations, who are more likely to experience disparities in access to traditional healthcare services (Hill et al., 2023).

Additionally, LLMs often provide a low-cost alternative for individuals who cannot easily access traditional healthcare because of cost, lack of insurance, transportation barriers, provider shortages, or distrust of healthcare institutions. This is especially important because recent national data continues to show uneven access to affordable care. For example, the Commonwealth Fund's 2024 State Health Disparities Report found persistent racial and ethnic disparities in health outcomes, healthcare access, and quality of care across the United States (Hill et al., 2023).

Despite the promising benefits of LLMs for healthcare information retrieval, concerns still exist (Wu et al. 2025) especially regarding their ability to provide reliable, safe, equitable and representative health guidance across diverse populations (Hannah et al., 2025; Omiye et al., 2023; Singhal et al., 2023; Yun et al., 2025). In a systematic review of LLM for health information, Yun et al. (2025) found that many studies reported inaccuracies, omissions, and potentially unsafe recommendations generated by these systems. Specifically, LLMs were found to occasionally produce factually incorrect medical information, generate incomplete responses that omitted clinically relevant details, and provide recommendations that were inconsistent with established clinical guidelines. In some cases, models exhibited hallucinations, whereby plausible sounding but inaccurate information was presented with a high degree of confidence. These limitations are particularly concerning for health information because patients may rely on LLM-generated information to make decisions regarding symptom evaluation, treatment options, medication use, and disease management. Consequently, inaccurate or incomplete information may contribute to delayed care, inappropriate self-diagnosis, poor treatment decisions, and adverse health outcomes.

Similarly, Stiennon et al. (2020) noted that as LLM become more advanced and are entrusted with important tasks, the mistakes they generate may become less obvious but more safety critical. Thus, they argue that additional research is needed. Their concern is particularly relevant with health information, where subtle inaccuracies, omissions, or biases in LLM information may influence health-related decisions and potentially exacerbate existing disparities among underserved populations.

Disparities in Medical Research: Minorities Underrepresented in Clinical Trials

The Centers for Disease Control and Prevention (CDC, 2022a, 2022b) states that racial and ethnic minority populations experience disproportionately higher rates of chronic disease compared to White populations. Specifically, 14.7% of American Indians/Alaska Natives, 12.5% of Hispanics, and 11.7% of non-Hispanic Black individuals are diagnosed with diabetes compared to 7.5% of non-Hispanic Whites. Given these disparities in disease burden, one would expect clinical trials to reflect proportionate racial representation. However, research findings consistently report the opposite. For example, Ahmed et al. (2022) conducted a meta-analysis examining racial participation in type 2 diabetes randomized clinical trials. They found that the participation-to-prevalence ratio (PPR) for White participants was 2.19 in industry-funded trials and 1.11 in government-funded trials. In contrast, minority representation ratios were substantially lower (0.33 in industry-funded trials and 0.73 in government-funded trials), indicating persistent and systematic underrepresentation. These findings and many other similar medical research findings suggest that clinical trial enrollment does not align with disease prevalence. Thus, this raises concerns regarding external validity and equitable evidence generation.

Policy and Regulatory Context

To eliminate the bias and help provide more equitable data, the National Institutes of Health (NIH) implemented the Inclusion of Women and Minorities policy (Office of the Law Revision Counsel, 2022). However, it is important to note that this policy does not mandate enforceable quotas or proportional targets. Thus, researchers argued that while inclusion language exists, measurable recruitment goals are often absent

(Andrasik et al., 2021). More importantly, Ahmed et al. (2022) reported that the inclusion requirements do not uniformly apply to industry-funded trials, thereby limiting the policy's systemic impact.

Additional legislative efforts have attempted to address financial barriers for the economic disadvantage to participate in clinical trials. For example, Section 210 of the Consolidated Appropriations Act 2021 mandated Medicaid coverage for routine costs associated with qualifying clinical trials beginning January 1, 2022 (United States Congress, 2020). While this provision may reduce financial disincentives, awareness remains limited.

Research consistently shows that individuals in poor and low-income households face disproportionately worse access to routine health care services compared to those in higher-income groups, including limited insurance coverage, reduced access to ongoing care, and fewer opportunities to build relationships with providers. Reduced access to routine care can in turn limit awareness of clinical trial opportunities and related protections (Chun and Park, 2012; Agency for Healthcare Research and Quality [AHRQ], 2021). Although these policies address direct medical costs, they do not remediate distrust, structural access limitations, or the absence of enforceable diversity benchmarks in industry-sponsored trials.

LLM Algorithm/Learning

LLMs are generally not trained directly on proprietary clinical trial datasets or patient medical records. As illustrated in Figure 1, LLMs learn from large corpora of publicly available text for the web (Kasneci. et al., 2023) which may include biomedical literature, clinical guidelines, textbooks, educational materials, and other publicly accessible sources (Singhal et al., 2023). It is important to note that LLMs often acquire knowledge regarding clinical trial findings indirectly through published research and related educational resources rather than through direct access to raw clinical trial data (Bommasani et al., 2021; Singhal et al., 2023). This indirect learning process has important implications for healthcare equity.

It is important to note that historically, racial and ethnic minority populations have been underrepresented in many clinical trials and biomedical research studies, resulting in evidence-generation disparities and gaps in knowledge regarding minority-specific health outcomes (Ahmed et al., 2017; National Academies of Sciences, Engineering, and Medicine [NASEM], 2022). As a result, the published biomedical literature may disproportionately reflect findings derived from majority populations, which can subsequently influence medical textbooks, clinical guidelines, and other healthcare knowledge resources (Ahmed et al., 2017; Oh et al., 2015). Because LLMs derive much of their healthcare knowledge from these sources, concerns have emerged that existing disparities and omissions within the biomedical evidence base may become reflected in LLM-generated healthcare information (Chauhan, 2025; Omiye et al., 2023). Thus, the concern is not necessarily that LLMs are intentionally biased, but rather that they may reproduce or reinforce biases, limitations, and knowledge gaps that already exist within the underlying healthcare evidence ecosystem (Shah et al., 2024).

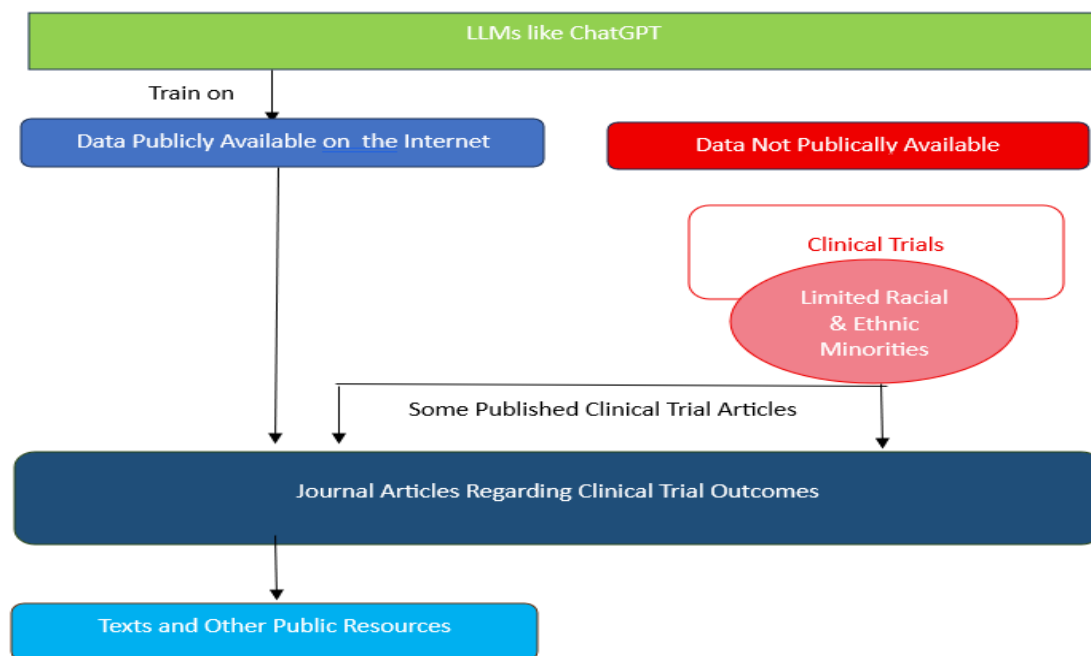


Figure 1. LLM Corpus Training

LLM Barriers

Systemic underrepresentation of minority populations in clinical trials introduces bias into the evidence base that informs biomedical research. This bias does not remain isolated; rather, it becomes encoded in the scientific literature and subsequently embedded within LLMs like ChatGPT, which are trained on and retrieve from these sources. Consequently, even when models generate responses grounded in peer-reviewed clinical evidence, users may encounter outputs that reflect historically skewed data. This raises critical concerns regarding equity, generalizability, and the reliability of LLM health information, particularly for underrepresented populations.

Bias embedded within clinical trial populations has significant implications for LLMs used to seek health information. Systemic underrepresentation of racial and ethnic minority populations in clinical trials introduces bias into the biomedical evidence base, limiting the generalizability of research findings (Oh et al., 2015; U.S. Food and Drug Administration, 2020). As large language models, including ChatGPT, are trained on and retrieve from this literature, these biases may be perpetuated in generated responses (Bender et al., 2021; Birhane et al., 2022). Consequently, individuals querying health-related information may receive outputs that reflect historically skewed data, increasing the risk of inequitable clinical guidance, delayed or missed diagnoses, and reduced treatment effectiveness for underrepresented populations (Seyyed-Kalantari et al., 2021; Gianfrancesco et al., 2018). Moreover, the amplification of biased recommendations at scale raises critical concerns regarding health equity, patient trust, and the ethical deployment of AI in healthcare settings (Rajkomar et al., 2018; World Health Organization, 2021). Collectively, these challenges highlight the need for more inclusive data practices and responsible AI governance to ensure equitable and reliable health information delivery.

Prompt Engineering Strategies

Prompt Engineering Strategies for Racial and Ethnic Minority Health Information Seeking

One common approach to interacting with LLMs is prompt engineering (Clavié et al, 2023; White et al, 2023), where users design and provide specific prompt texts to guide LLMs in generating desired responses or completing specific tasks. “The quality of the output(s) generated by a conversational LLMs like are directly related to the quality of the prompts provided” (White et al, 2023)”.

Prompt engineering refers to the process of designing and structuring prompts to improve the quality, transparency, and relevance of LLM responses (White et al., 2023). Prompt engineering plays a critical role in shaping the quality, specificity, and usefulness of AI-generated responses. Research has demonstrated that prompt design can significantly influence the accuracy, reasoning, and overall performance of LLMs with more detailed and context-rich prompts generally producing more relevant and informative outputs (Wei et al., 2022; White et al., 2023). Because LLMs generate responses based on patterns learned from their training data, underspecified prompts may produce generalized outputs that fail to account for important demographic, cultural, or contextual factors, particularly for populations that have historically been underrepresented in the biomedical literature (Bommasani et al., 2021; Chauhan et al., 2025; Omiye et al., 2023; White et al., 2023). For example, a generic or zero shot prompt such as, "Tell me about Type 2 diabetes," is likely to produce a broad overview of the condition. In contrast, a more detailed context prompt which involves providing demographic, clinical, and situational information (e.g., age, race or ethnicity, comorbidities, and healthcare access factors) to help the model generate more relevant and personalized response (White et al., 2023).

Given concerns regarding evidence-generation disparities and potential bias in LLM health information outputs, racial and ethnic minority users may benefit from employing prompt engineering strategies to obtain more relevant, transparent, and contextualized health information (Esmacilzadeh, 2025). Table 1 provides three foundational prompt engineering strategies and examples that may help racial and ethnic minority users obtain more relevant health information from LLMs.

Our first strategy, demographic context, encourages users to include characteristics such as age, race or ethnicity, and gender within their prompts. Because health risks, disease prevalence, treatment responses, and preventive care recommendations frequently differ across demographic groups, this strategy is essential to include this type of context prompting as it encourages the model to generate responses that are more specific to the population being described rather than relying on generalized health information. Nevertheless, demographic prompting cannot compensate for all biases presented within the underlying training data. If racial and ethnic minority populations were historically underrepresented in clinical trials or biomedical literature, LLM outputs may still reflect incomplete evidence or majority-centered recommendations despite highly specific prompts (Bender et al., 2021; Omiye et al., 2023).

Our second strategy expands the prompt by incorporating clinical context, including existing diagnoses and comorbid conditions. This strategy aligns with the prompt engineering principle of contextual conditioning, whereby additional domain-specific information improves the relevance of generated responses (White et al., 2023). Rather than producing generic educational information about a single disease, clinically enriched prompts encourage the model to consider interactions among multiple chronic conditions and generate more comprehensive health information. For example, asking about Type 2 diabetes together with hypertension encourages discussion of cardiovascular risk, medication interactions, and lifestyle modifications that may not appear in a more general prompt. However, LLMs remain educational tools rather than diagnostic systems. Specifically, general LLMs like ChatGPT and Gemini lack access to patient histories, laboratory

results, medication lists, and physical examinations necessary for individualized clinical decision-making (Bender et al., 2021).

Finally, our third strategy incorporates constraint-based prompting in which users specify practical constraints that guide the model toward recommendations that are feasible within the user's real-world circumstances (White et al., 2023). For example, instead of assuming ideal access to healthcare resources, the model is prompted to discuss lower-cost treatment options, community resources, patient assistance programs, or questions that patients can raise with their healthcare providers. This strategy is particularly important because healthcare disparities often result not only from medical conditions but also from social determinants.

Table 1. Initial Prompt Engineering Strategies and Examples for Racial and Ethnic Minorities Seeking Health Information

Prompt Strategy & Example

Purpose	AI Principle	Supporting Literature	Limitation
Demographic Context			
"I am a 52-year-old African American male. What are my risk factors for Type 2 diabetes?" or "What health disparities affect African Americans males with Type 2 diabetes?"			
Provide age, race/ethnicity, and gender	Context enrichment and prompt specificity	Prompt specificity reduces ambiguity and generally improves response quality by narrowing the model's search space (White et al., 2023).	Not all highly specific prompts can compensate for biased or incomplete training data. If minority populations were underrepresented in clinical trials or biomedical literature, model outputs may still reflect those underlying biases (Bender et al., 2021; Omiye et al., 2023).
Clinical Context			
"I am a 52-year-old African American male with Type 2 diabetes and hypertension. What should I know about managing these conditions?"			
Provide clinically relevant diagnoses and comorbidities	Contextual conditioning	More complete contextual prompts may improve relevance and reduce overly generic responses (White et al., 2023).	LLMs do not perform individualized medical diagnosis and may omit important clinical factors making professional (Bender et al., 2021).
Healthcare Access Context			
"I have Type 2 diabetes, limited transportation, and no health insurance. What management options should I discuss with my doctor?"			

Describe healthcare access barriers and social determinants affecting care.	Constraint-based prompting	Constraint-based prompting provides better responses regarding the user's real-world circumstances rather than ideal clinical scenarios (White et al., 2023).	Training datasets often underrepresent social determinants of health, reducing the relevance of recommendations for underserved populations (Omiye et al., 2023).
---	----------------------------	---	---

Follow-up/Conversational Prompt Engineering Strategies for Racial and Ethnic Minority Health Information Seeking

It is important to note that current LLMs are designed to support multi-turn dialogue which enables users to engage in iterative conversations by asking follow-up questions that clarify recommendations, identify knowledge gaps, and request supporting evidence (Kwan et al., 2024). Rather than accepting the initial response as complete or authoritative, users can employ conversational prompt engineering techniques to improve transparency, encourage critical evaluation, and better understand the strengths and limitations of AI-generated information. This iterative process is particularly important when individuals from racial and ethnic minority populations seek health information because the underlying evidence supporting model responses may itself reflect historical underrepresentation in clinical research. As a result, users may benefit from going beyond passive consumption of AI-generated information and instead engaging in active, evidence-informed dialogue with LLMs by asking follow-up questions regarding research representation, supporting evidence, assumptions, and limitations. Table 2 provides suggested follow-up/conversational strategies and examples which may help racial and ethnic minorities seeking health information from LLM.

Our first follow-up strategy uses reflective prompting for understanding research gaps and clinical trial representation. among the output. Users may benefit from explicitly questioning whether the available scientific evidence adequately represents populations similar to themselves rather than accepting them at face value (Jeevarathinam, 2026). For example, by asking whether African Americans or other minority populations were sufficiently represented in the clinical trials supporting a recommendation, users encourage the model to acknowledge potential limitations in the evidence base rather than presenting findings as universally applicable. This approach aligns closely with the central premise of the present study that algorithmic bias may originate from sample selection bias within clinical research.

It is important to note; this follow-up strategy is limited because excessive or repetitive follow-up questioning may lead the model to overemphasize uncertainty or generate increasingly speculative responses. As a result, this may lead to prompt overfitting and diminished response stability (Jeevarathinam, 2026).

Our second follow-up strategy incorporates evidence-based prompting and metacognitive prompting to encourage the model to explain its reasoning and explicitly communicate the confidence and limitations associated with its responses (Esmailzadeh, 2025) rather than simply accepting the initial output. Transparency is particularly valuable in healthcare because users often assume that confident responses indicate factual accuracy. Requesting supporting evidence encourages greater critical evaluation and may reduce the likelihood of unquestioningly accepting incorrect or incomplete information.

However, this type of prompting will still require the user to use their judgement with regards to understanding that the LLM may provide outdated, incomplete, or not directly applicable to the user's clinical situation (Xu et al., 2025). Additionally, the LLM may sometimes confuse generated explanations

with actual supporting evidence, producing poor scientific literature (Huo et al., 2023). Therefore, users should independently verify cited guidelines and research findings through trusted clinical sources and healthcare professionals before making healthcare decisions. Often this requires asking for the digital object identifier (DOI) and reading the resource.

As illustrated in Figure 2, Tables 1 and 2 create two-step prompt engineering framework designed to improve health information seeking among racial and ethnic minority populations. The framework is intended to be applied sequentially.

Prompt Engineering Strategies

Prompt Engineering Strategies for Racial and Ethnic Minority Health Information Seeking

One common approach to interacting with LLMs is prompt engineering (Clavié et al, 2023; White et al, 2023), where users design and provide specific prompt texts to guide LLMs in generating desired responses or completing specific tasks. "The quality of the output(s) generated by a conversational LLMs like are directly related to the quality of the prompts provided" (White et al, 2023)".

Prompt engineering refers to the process of designing and structuring prompts to improve the quality, transparency, and relevance of LLM responses (White et al., 2023). Prompt engineering plays a critical role in shaping the quality, specificity, and usefulness of AI-generated responses. Research has demonstrated that prompt design can significantly influence the accuracy, reasoning, and overall performance of LLMs with more detailed and context-rich prompts generally producing more relevant and informative outputs (Wei et al., 2022; White et al., 2023). Because LLMs generate responses based on patterns learned from their training data, underspecified prompts may produce generalized outputs that fail to account for important demographic, cultural, or contextual factors, particularly for populations that have historically been underrepresented in the biomedical literature (Bommasani et al., 2021; Chauhan et al., 2025; Omiye et al., 2023; White et al., 2023). For example, a generic or zero shot prompt such as, "Tell me about Type 2 diabetes," is likely to produce a broad overview of the condition. In contrast, a more detailed context prompt which involves providing demographic, clinical, and situational information (e.g., age, race or ethnicity, comorbidities, and healthcare access factors) to help the model generate more relevant and personalized response (White et al., 2023).

Given concerns regarding evidence-generation disparities and potential bias in LLM health information outputs, racial and ethnic minority users may benefit from employing prompt engineering strategies to obtain more relevant, transparent, and contextualized health information (Esmailzadeh, 2025). Table 1 provides three foundational prompt engineering strategies and examples that may help racial and ethnic minority users obtain more relevant health information from LLMs.

Our first strategy, demographic context, encourages users to include characteristics such as age, race or ethnicity, and gender within their prompts. Because health risks, disease prevalence, treatment responses, and preventive care recommendations frequently differ across demographic groups, this strategy is essential to include this type of context prompting as it encourages the model to generate responses that are more specific to the population being described rather than relying on generalized health information. Nevertheless, demographic prompting cannot compensate for all biases presented within the underlying training data. If racial and ethnic minority populations were historically underrepresented in clinical trials or biomedical literature, LLM outputs may still reflect incomplete evidence or majority-centered recommendations despite highly specific prompts (Bender et al., 2021; Omiye et al., 2023).

Our second strategy expands the prompt by incorporating clinical context, including existing diagnoses and comorbid conditions. This strategy aligns with the prompt engineering principle of contextual conditioning, whereby additional domain-specific information improves the relevance of generated responses (White et al., 2023). Rather than producing generic educational information about a single disease, clinically enriched prompts encourage the model to consider interactions among multiple chronic conditions and generate more comprehensive health information. For example, asking about Type 2 diabetes together with hypertension encourages discussion of cardiovascular risk, medication interactions, and lifestyle modifications that may not appear in a more general prompt. However, LLMs remain educational tools rather than diagnostic systems. Specifically, general LLMs like ChatGPT and Gemini lack access to patient histories, laboratory results, medication lists, and physical examinations necessary for individualized clinical decision-making (Bender et al., 2021).

Finally, our third strategy incorporates constraint-based prompting in which users specify practical constraints that guide the model toward recommendations that are feasible within the user's real-world circumstances (White et al., 2023). For example, instead of assuming ideal access to healthcare resources, the model is prompted to discuss lower-cost treatment options, community resources, patient assistance programs, or questions that patients can raise with their healthcare providers. This strategy is particularly important because healthcare disparities often result not only from medical conditions but also from social determinants.

Table 1. Initial Prompt Engineering Strategies and Examples for Racial and Ethnic Minorities Seeking Health Information

Prompt Strategy & Example

Purpose	AI Principle	Supporting Literature	Limitation
Demographic Context			
"I am a 52-year-old African American male. What are my risk factors for Type 2 diabetes?" or "What health disparities affect African Americans males with Type 2 diabetes?"			
Provide age, race/ethnicity, and gender	Context enrichment and prompt specificity	Prompt specificity reduces ambiguity and generally improves response quality by narrowing the model's search space (White et al., 2023).	Not all highly specific prompts can compensate for biased or incomplete training data. If minority populations were underrepresented in clinical trials or biomedical literature, model outputs may still reflect those underlying biases (Bender et al., 2021; Omiye et al., 2023).
Clinical Context			
"I am a 52-year-old African American male with Type 2 diabetes and hypertension. What should I know about managing these conditions?"			
Provide clinically relevant	Contextual conditioning	More complete contextual prompts may improve relevance and reduce	LLMs do not perform individualized medical diagnosis and may omit important clinical factors

diagnoses and comorbidities	overly generic responses (White et al., 2023).	making professional (Bender et al., 2021).
Healthcare Access Context		
"I have Type 2 diabetes, limited transportation, and no health insurance. What management options should I discuss with my doctor?"		
Describe healthcare access barriers and social determinants affecting care.	Constraint-based prompting provides better responses regarding the user's real-world circumstances rather than ideal clinical scenarios (White et al., 2023).	Training datasets often underrepresent social determinants of health, reducing the relevance of recommendations for underserved populations (Omiye et al., 2023).

Follow-up/Conversational Prompt Engineering Strategies for Racial and Ethnic Minority Health Information Seeking

It is important to note that current LLMs are designed to support multi-turn dialogue which enables users to engage in iterative conversations by asking follow-up questions that clarify recommendations, identify knowledge gaps, and request supporting evidence (Kwan et al., 2024). Rather than accepting the initial response as complete or authoritative, users can employ conversational prompt engineering techniques to improve transparency, encourage critical evaluation, and better understand the strengths and limitations of AI-generated information. This iterative process is particularly important when individuals from racial and ethnic minority populations seek health information because the underlying evidence supporting model responses may itself reflect historical underrepresentation in clinical research. As a result, users may benefit from going beyond passive consumption of AI-generated information and instead engaging in active, evidence-informed dialogue with LLMs by asking follow-up questions regarding research representation, supporting evidence, assumptions, and limitations. Table 2 provides suggested follow-up/conversational strategies and examples which may help racial and ethnic minorities seeking health information from LLM.

Table 2. Follow-up/Conversational Prompt Engineering Strategies and Examples for Racial and Ethnic Minorities Seeking Health Information from LLMs

Prompt Strategy & Example			
Purpose	AI Principle	Supporting Literature	Limitation
Research Gaps & Clinical Trial Representation			
"What research gaps exist regarding Type 2 diabetes among African Americans?" or "Were African Americans adequately represented in the clinical trials supporting these treatments?"			
Identify limitations in available research	Reflective prompting	Reflective prompting encourages thoughtful evaluation of AI-generated responses, helping identify missing information and reduce errors (Jeevarathinam, 2026).	Overfitting may occur ((Jeevarathinam, 2026).

Evidence Request			
"What evidence or clinical guidelines support these recommendations?"			
or			
"What assumptions, limitations, or potential biases should I consider when interpreting this information?"			
Request supporting evidence & identify uncertainty or limitations	Evidenced-Based and/ or Metacognitive prompting	Metacognitive prompting aids in acknowledging uncertainties and limitations and can prevent errors (Esmaeilzadeh, 2025).	Supporting literature may be limited in scope, outdated, irrelevant, and/or inaccurate (Xu et al., 2025). Also, LLMs can confuse the answer with evidence (Huo et al., 2023).

Our first follow-up strategy uses reflective prompting for understanding research gaps and clinical trial representation. among the output. Users may benefit from explicitly questioning whether the available scientific evidence adequately represents populations similar to themselves rather than accepting them at face value (Jeevarathinam, 2026). For example, by asking whether African Americans or other minority populations were sufficiently represented in the clinical trials supporting a recommendation, users encourage the model to acknowledge potential limitations in the evidence base rather than presenting findings as universally applicable. This approach aligns closely with the central premise of the present study that algorithmic bias may originate from sample selection bias within clinical research.

It is important to note; this follow-up strategy is limited because excessive or repetitive follow-up questioning may lead the model to overemphasize uncertainty or generate increasingly speculative responses. As a result, this may lead to prompt overfitting and diminished response stability (Jeevarathinam, 2026).

Our second follow-up strategy incorporates evidence-based prompting and metacognitive prompting to encourage the model to explain its reasoning and explicitly communicate the confidence and limitations associated with its responses (Esmaeilzadeh, 2025) rather than simply accepting the initial output. Transparency is particularly valuable in healthcare because users often assume that confident responses indicate factual accuracy. Requesting supporting evidence encourages greater critical evaluation and may reduce the likelihood of unquestioningly accepting incorrect or incomplete information.

However, this type of prompting will still require the user to use their judgement with regards to understanding that the LLM may provide outdated, incomplete, or not directly applicable to the user's clinical situation (Xu et al., 2025). Additionally, the LLM may sometimes confuse generated explanations with actual supporting evidence, producing poor scientific literature (Huo et al., 2023). Therefore, users should independently verify cited guidelines and research findings through trusted clinical sources and healthcare professionals before making healthcare decisions. Often this requires asking for the digital object identifier (DOI) and reading the resource.

As illustrated in Figure 2, Tables 1 and 2 create two-step prompt engineering framework designed to improve health information seeking among racial and ethnic minority populations. The framework is intended to be applied sequentially. In the first step, users develop context-rich initial prompts by incorporating demographic, clinical, and healthcare access information to improve the relevance and personalization of LLM-generated responses. In the second step, users engage in iterative follow-up prompting to critically evaluate those responses by requesting supporting evidence, identifying research gaps, recognizing uncertainty, and examining potential bias. Together, these complementary steps integrate contextual and constraint-based prompt construction with reflective and evidence-based conversational refinement to provide a more transparent and contextually appropriate LLM output.

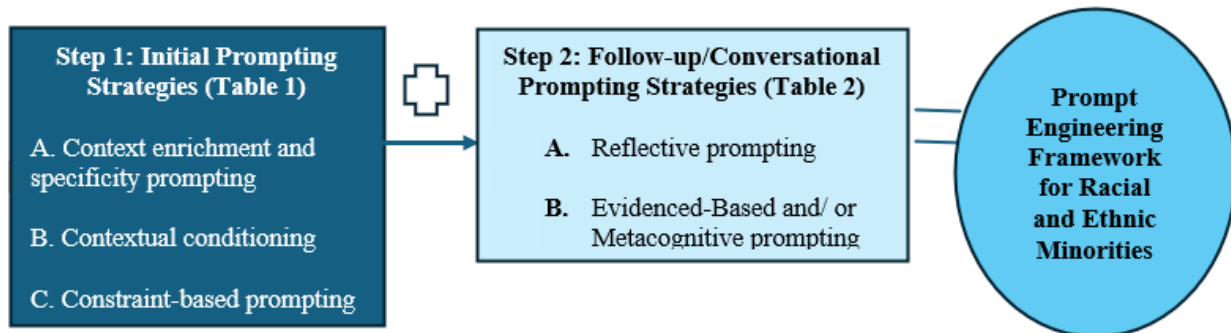


Figure 2. Two-Stage Prompt Engineering Framework for Racial and Ethnic Minorities Seeking Health Information via LLMs

Conclusion

LLMs are increasingly being used by the public as a readily accessible source of health information (Olsen, 2026). However, LLM healthcare information outputs may be dependent upon the evidence base from which the system learns (Chauhan, 2025; Omiye et al., 2023). As this review demonstrates, longstanding underrepresentation of racial and ethnic minority populations in clinical trials has contributed to evidence-generation disparities that may ultimately influence healthcare-related LLM outputs.

While continued efforts are needed to improve diversity and inclusion within clinical research and LLM health information, minorities must also be equipped with the digital literacy skills necessary to critically use LLM for health information. Prompt engineering represents one practical approach for encouraging more contextualized, transparent, and evidence-informed interactions with LLMs (Clavié et al, 2023; White et al, 2023). By incorporating demographic, clinical, and health equity considerations into prompts, users may obtain responses that better reflect their individual circumstances while increasing awareness of potential research limitations.

It is important to note that this paper is not without limitations. First, this research is limited in that it is theoretical in nature. Second, this study did not test any of the techniques described. Future research should address the above limitations and test the suggested prompts provided in Tables 1 and 2 against multiple LLMs. Finally, this paper has practical implications which include an overview of the literature and knowledge for health-seeking minority individuals, health educators, LLM developers, health equity researchers, IT practitioners designing health information systems, practitioners, and academics as a teaching resource regarding using generic LLMs for health information and understanding how clinical trial underrepresentation shapes healthcare LLM outputs.

References

- Agency for Healthcare Research and Quality. (2021, December). 2021 National Healthcare Quality and Disparities Report (AHRQ Publication No. 21(22)-0054-EF). U.S. Department of Health and Human Services. <https://www.ncbi.nlm.nih.gov/books/NBK578529/>
- Ahmed, R., de Souza, R. J., Li, V., & Anand, S. S. (2022). Assessing non-white ethnic participation in type 2 diabetes mellitus randomized clinical trials: A meta-analysis. *medRxiv* <https://doi.org/10.1101/2022.06.28.22275821>
- Andrasik, M. P., Broder, G. B., Wallace, S. E., Michaels, S., Kolb, M., Allen, M. A., Corbie-Smith, G., & COVID-19 Prevention Network Community Engagement Working Group. (2021). Increasing Black, Indigenous and people of Color participation in clinical trials through community engagement and recruitment goal establishment. *PLOS ONE*, 16(10), e0258858. <https://doi.org/10.1371/journal.pone.0258858>
- Barreda, M. (2025). Transforming healthcare with chatbots: Uses and opportunities. *Digital Health*, 11. <https://doi.org/10.1177/20552076251319174>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Birhane, A., Prabhu, V. U., & Kahembwe, E. (2022). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2110.01963>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J. Q., Demszky, D., ... Liang, P. (2021). *On the opportunities and risks of foundation models*. Stanford Center for Research on Foundation Models. <https://doi.org/10.48550/arXiv.2108.0725>
- Centers For Disease Control and Prevention. (2022a). *National Diabetes Statistics Report*. Centers For Disease and Control and Prevention. www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

- Centers for Disease Control and Prevention. (2022b). *By the numbers: Diabetes in America*. <https://www.cdc.gov/diabetes/health-equity/diabetes-by-the-numbers.html>
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45. <https://doi.org/10.1145/3641289>
- Chun, N., & Park, M. (2012). The impact of clinical trial insurance coverage mandates on racial minorities and low-income individuals. *Applied Economics*, 44(15), 1977–1984. <https://doi.org/10.1080/00036846.2011.556595>
- Clavié, B., Ciceu, A., Naylor, F., Soulié, G., & Brightwell, T. (2023, June). Large language models in the workplace: A case study on prompt engineering for job type classification. In *International conference on applications of natural language to information systems* (pp. 3-17). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35320-8_1
- Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., Liu, X., Nour, M. M., Spielman, S., Way, S. F., Shah, Y., Bhaskar, M., Nori, H., Kelly, C., Hames, P., Gross, B., Suleyman, M., & King, D. (2026). Public use of a generalist LLM chatbot for health queries. *Nature Health*. <https://doi.org/10.1038/s44360-026-00117-x>
- Esmailzadeh, P. (2025). Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. *BMC Medical Informatics and Decision Making*. 25(342). <https://doi.org/10.1186/s12911-025-03182-6>
- Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11), 1544–1547. <https://doi.org/10.1001/jamainternmed.2018.3763>
- Hanna, J. J., Wakene, A. D., Johnson, A. O., Lehmann, C. U., Medford, R. J. (2025). Assessing racial and ethnic bias in text generation by large language models for health care-related tasks: Cross-sectional study. *Journal of Medical Internet Research*; 27: e57257 <https://www.doi.org/10.2196/57257>
- Hill, L., Ndugga, N., & Artiga, S. (2023, August 17). Key data on health and health care by race and ethnicity. Kaiser Family Foundation (KFF). <https://www.kff.org/racial-equity-and-health-policy/report/key-data-on-health-and-health-care-by-race-and-ethnicity/>
- Huo, S., Arabzadeh, N., & Clarke, C. L. A. (2023). Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '23)* (pp. 11–20). Association for Computing Machinery. <https://doi.org/10.1145/3624918.3625336>
- Jeevarathinam, V. (2026, June 29). Reflection prompting explained: Why one prompt is not enough. <https://www.f22labs.com/blogs/reflection-prompting-explained-why-one-prompt-is-not-enough/>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

- Kurniawan, M. H., Handiyani, H., Nuraini, T., Hariyati, R. T. S., & Sutrisno. (2024). A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Annals of Medicine*, 56(1), Article 2302980. <https://doi.org/10.1080/07853890.2024.2302980>
- Kwan, W.-C., Zeng, X., Jiang, Y., Wang, Y., Li, L., Shang, L., Jiang, X., Liu, Q., & Wong, K.-F. (2024). MT-Eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 20153–20177). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.1124>
- Markman, J. (2026). Everything about ChatGPT health you need to know. *Forbes*. <https://www.forbes.com/sites/jonmarkman/2026/01/08/everything-about-chatgpt-health-you-need-to-know/>
- Nadarzynski, T., Knights, N., Husbands, D., et al. (2024). Achieving health equity through conversational AI: A roadmap for design and implementation of inclusive chatbots in healthcare. *PLOS Digital Health*, 3(5), e0000492. <https://doi.org/10.1371/journal.pdig.0000492>
- Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2023). User intentions to use ChatGPT for self-diagnosis and health-related purposes: Cross-sectional survey study. *Journal of Medical Internet Research - Human Factors*, 10, e47564. <https://doi.org/10.2196/47564>
- National Academies of Sciences, Engineering, and Medicine. (2022). *Improving representation in clinical trials and research: Building research equity for women and underrepresented groups*. National Academies Press. <https://doi.org/10.17226/26479>
- Office of the Law Revision Counsel. (2022). 42 U.S.C. § 289a-2: *Inclusion of women and minorities in clinical research*. United States Code. <https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title42-section289a-2>
- Oh, S. S., Galanter, J., Thakur, N., Pino-Yanes, M., Barcelo, N. E., White, M. J., ... & Burchard, E. G. (2015). Diversity in clinical and biomedical research: A promise yet to be fulfilled. *PLOS Medicine*, 12(12), e1001918. <https://doi.org/10.1371/journal.pmed.1001918>
- Olsen, E. (2026, March). 1 in 3 adults use AI for health information: Poll. *Healthcare Drive*. <https://www.healthcaredrive.com/news/one-in-three-adults-use-artificial-intelligence-health-information-kff/815722/>
- Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *npj Digital Medicine*, 6(195). <https://doi.org/10.1038/s41746-023-00939-z>
- Powell, L.M., Powell, G., Rebman Jr., C.M., Wimmer, H., (2026). Using a Large Language Model to Evaluate Content Quality in Nutritional YouTube Shorts. *Journal of Information Systems Applied Research and Analytics* 19(4), 53-69. <https://doi.org/10.62273/HFNS3900>
- Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12), 866–872. <https://doi.org/10.7326/M18-1990>

- Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs. *Nature Medicine*, 27(12), 2176–2182. <https://doi.org/10.1038/s41591-021-01595-0>
- Shah, N. H., Entwistle, D., & Pfeffer, M. A. (2023). *Creation and adoption of large language models in medicine*. *JAMA*, 330(9), 866–869. <https://doi.org/10.1001/jama.2023.14217>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize from human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems*, 33, 3008–3021. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html
- Turchioe, M. R., Shamnath, A., Mayfield, J., Xu, O., Slotwiner, D., & Benda, N. (2026). *Are older adults using ChatGPT for medical advice? Results from a cross-sectional survey study*. medRxiv [Preprint]. <https://doi.org/10.64898/2026.01.01.25342468>
- U.S. Food and Drug Administration. (2020). *Enhancing the diversity of clinical trial populations—Eligibility criteria, enrollment practices, and trial designs*. <https://www.fda.gov/media/127712/download>
- United States Congress, House. (2020, December 27). *Consolidated Appropriations Act, 2021* (H.R. 133, 116th Cong., 2nd Sess.). <https://www.congress.gov/bill/116th-congress/house-bill/133>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*. <https://doi.org/10.48550/arXiv.2201.11903>
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*. <https://doi.org/10.48550/arXiv.2302.11382>
- Wu, H., Li, M., & Zhang, L. (2025). Comparing physician and large language model responses to influenza patient questions in the online health community. *International Journal of Medical Informatics*, 197, 105836, <https://doi.org/10.1016/j.ijmedinf.2025.105836>.
- Xu, Z., Zhao, Y., Patwardhan, M., Vig, Lovekesh & Cohan, A. (2025). Can LLMs identify critical limitations within scientific research? A systematic evaluation on ai research papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics 1: Long Papers*, 20652–20706, Vienna, Austria. Association for Computational Linguistics. <https://aclanthology.org/2025.acl-long.1009.pdf>

Yun H, & Bickmore, T. (2025). Online health information-seeking in the era of large language models: Cross-Sectional web-based survey study *Journal of Medical Internet Research*, 27: e68560
<https://www.jmir.org/2025/1/e68560> <https://www.doi.org/10.2196/68560>