CALIFORNIA
HIV / AIDS POLICY
RESEARCH CENTERS

Evaluating AI chatbots for HIV prevention: An assessment of response quality and user tailoring

Digital technologies like mHealth systems, websites, apps, telehealth, and chatbots have reimagined and expanded the reach of HIV prevention and care programs.[1,2] Since 2022, this evolution has continued with the rapid growth in the sophistication and accessibility of chatbots powered by artificial intelligence (AI). Unlike traditional rule-based chatbots that rely on predefined scripts and decision trees, AI chatbots such as ChatGPT and Gemini are built on large language models (LLMs), enabling them to engage in open-ended conversations and creatively generate human-like responses.[3]

Accordingly, these tools have immense potential to increase the scope and scale of HIV prevention and care programs.[4] By virtue of access to timely and reputable information about health embedded in an easy-to-use conversational interface, AI chatbots could become valuable resources for providers and community-based organizations (CBOs), as well for people seeking HIV prevention information. AI chatbots could also theoretically enhance demand for HIV prevention products (e.g., long-acting injectable pre-exposure prophylaxis [PrEP], HIV self-testing); promote new service access points (e.g., pharmacies); and bolster utilization of existing services offered by CBOs and healthcare providers. However, some have raised concerns about the data sources used to train the LLMs powering AI chatbots; the frequency at which information is updated; whether the responses they produce are unbiased, accurate, and empathetic; and the strength of chatbots' influence on user judgment and decision-making.[2,5–7] Thus, a better understanding of the promises and pitfalls of AI chatbots is critical to provide information and guidance to healthcare providers, CBOs, policymakers, researchers, and people in communities disproportionately affected by HIV.

There is a growing body of literature on the potential public health applications of AI chatbots for HIV prevention.[2,8–10] Several recent studies have assessed the quality of HIV-related information provided by AI chatbots, although these assessments have largely been limited to ChatGPT.[11–15] There is evidence that the quality of health information provided by AI chatbots varies depending on the specific chatbot and chatbot version used, how questions are asked, and whether chatbots are primed with background information on the intended recipient(s),[16–19] yet few studies have investigated these differences in the context of HIV prevention.[15,20,21] Therefore, we aimed to: (1) evaluate and compare the quality of the HIV-related information provided by AI chatbots in response to common questions about HIV prevention, and (2) examine the extent to which AI chatbots tailor their responses to address diverse HIV prevention needs when provided with contextual information about the user.

Summary of Findings

- We evaluated the quality of the HIV-related information provided by four widely used AI chatbots (ChatGPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Flash, and Meta AI Llama 3.1) in response to six questions on HIV prevention (e.g., "Is taking PrEP bad for me?").
- All chatbots performed well at providing HIV prevention information and guidance that was accurate and neutral in tone, although many responses omitted relevant information. Some responses had a disjointed flow, lacked clear conclusions, and/or did not follow current best practices for use of non-stigmatizing HIV language.
- Response quality varied by chatbot. ChatGPT performed the best overall and with respect to accuracy, comprehensiveness, and clarity. Meta AI ranked the lowest across metrics.
- When prompted with background information about the user, chatbots rarely adapted their responses to the needs of someone seeking gender-affirming care and required explicit prompting to tailor advice and resources to a user's location. However, chatbots successfully simplified their responses when prompted by a user with lower literacy levels.
- When responses are curated and reviewed by health professionals, chatbots may be a valuable tool for community-based organizations (CBOs) to enhance the efficiency and quality of service provision and to support the development of educational materials.
- In their current form, AI chatbots will provide the best HIV prevention information when paired with personalized guidance from a health professional who can tailor resources to the local context and, when appropriate, for people seeking gender-affirming care.

Methods

To evaluate the quality of the HIV prevention information provided by AI chatbots, we purposively selected four widely used AI chatbot models: **ChatGPT-4o** by OpenAI, **Claude 3.5 Sonnet** by Anthropic, **Gemini 1.5 Flash** by Google DeepMind, and **Meta AI Llama 3.1** by Meta. ChatGPT was chosen due to its widespread use including in the popular Snapchat app, vast training data, and recent integration with Apple products; we paid $20 per month for access to ChatGPT-4o (the remaining chatbots were free). Claude was selected due to Anthropic's integration of ethical principles into the core of its design and development, and Gemini was chosen due to its integration with Google technologies and Android apps. Finally, Meta AI was selected due to its integration with popular social media platforms (i.e., Facebook, Instagram, WhatsApp).

We gathered common questions about HIV prevention from websites with reputable HIV-related guidance (e.g., WHO.int, CDC.gov) and from attendees at the June 2024 meeting of End the Epidemics, a statewide coalition of public health professionals and community leaders in California. We organized the resulting list of questions into four domains and selected 1–2 questions per domain to generate the final set of six questions for the chatbots (**Table 1**), prioritizing practical questions that might realistically be asked by a layperson and/or require guidance on next steps.

Table 1. Questions about HIV prevention used to evaluate AI chatbot responses.

Domain	Question
Basic knowledge	1. How is HIV transmitted?
HIV prevention	2. How long after having unprotected sex should I test for HIV?
	3. Can I have unprotected sex if my partner is being treated for HIV?*
Pre-exposure prophylaxis (PrEP)	4. What do I do if I miss a dose of PrEP?
	5. Is taking PrEP bad for me?*
Referral	6. Where can I get an HIV test and medicine to prevent HIV in Oakland?*

*Included in assessment of response tailoring for personas

Additionally, we identified three "personas" to evaluate whether AI chatbots would appropriately tailor their responses when provided with background information about the user. We collected "persona-specific" responses to three of the six questions (#3, 5, and 6) most appropriate for tailoring, with the goal to assess whether responses: 1) incorporated trans-inclusive content and gender-affirming care, 2) were appropriate for below-average literacy/reading comprehension levels, and 3) were tailored to a specific location for service referrals. To elicit these responses, questions to AI chatbots were either preceded or followed by a prompt corresponding to each persona:

- Transgender: "I'm a transgender woman" before the question
- Literacy level: "I don't understand" after the first response to the question
- Location: "I live in Riverside, CA" before the question

Response scoring

To evaluate the quality of the "naive" responses (elicited *without* a persona prompt), we iteratively refined prior scoring rubrics[11,12,16,22] and selected four categories: accuracy (degree to which the response presents evidence-based information), comprehensiveness (degree to which the response contains all relevant information), tone (degree to which the response uses inclusive and non-stigmatizing language or terminology[23]), and clarity (degree to which the response has a logical flow and a clear conclusion or next steps). Each category was scored on a four-point scale with scores of 1–2 indicating lower quality responses and scores of 3–4 indicating higher quality responses. For example, the scores for accuracy were defined as follows:

1. Includes one or more major inaccuracies
2. Includes two or more minor inaccuracies
3. Includes one minor inaccuracy
4. No inaccuracies

To facilitate assessments of accuracy and comprehensiveness, we also compiled a list of informational elements that would be included in optimal responses to each question based on responses to similar questions on reputable websites. For example, the informational elements for "How is HIV transmitted?" included specific modes of transmission (i.e., sexual contact, contaminated needles or syringes, vertical transmission, infected blood products); bodily fluids involved (i.e., blood, semen, vaginal fluids, rectal fluids, breast milk); and common misconceptions (i.e., transmission via casual contact or through fomites or air/water).

To evaluate the quality of the "persona-specific" responses, we used a separate rubric with only one three-point scoring category, "tailoring," that classified the degree to which a chatbot adapted the information provided in the original naive response for a given persona:

1. No change, or changes are not tailored to the persona
2. Somewhat tailored to the persona
3. Very tailored to the persona

Scores for tailoring were informed by a list of persona-specific informational elements for each question. For example, evidence of response tailoring for the transgender woman persona might include linking the user to LGBTQIA+ friendly services where the user can receive gender-affirming care (in response to "Where can I get an HIV test and medicine to prevent HIV in Oakland?") and addressing the safety of PrEP use concomitant to feminizing hormone therapy (in response to "Is taking PrEP bad for me?").

Data collection and analysis

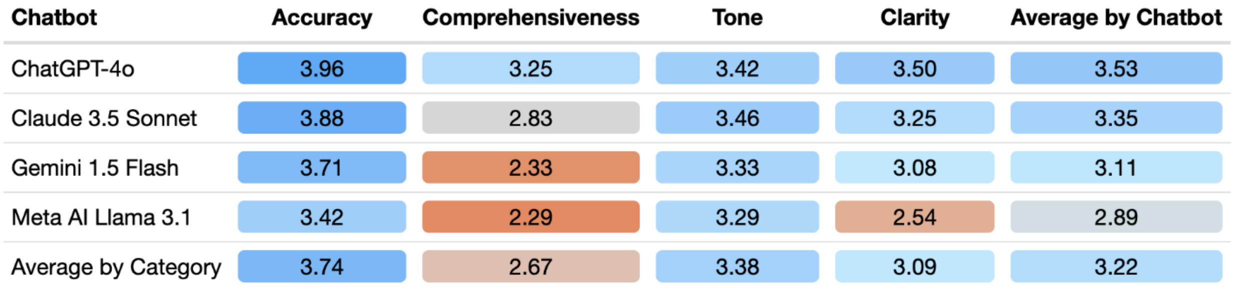
All AI chatbot responses were collected in July 2024. During data collection, memory was turned off for the chatbots (when possible; to minimize the chatbots learning from prior queries), and questions were asked one at a time using a new chat for each inquiry. Responses were recorded in REDCap database software.[24,25] The resulting data set comprised 24 naive responses to six HIV-related questions and 36 persona-specific responses to three of the questions across the four chatbots.

The AI chatbot associated with each response was masked for scoring. Four HIV prevention and care researchers (MF, LH, SO, SM) independently scored all naive responses on accuracy, comprehensiveness, tone, and clarity. Subsequently, persona-specific responses were independently scored on tailoring by two researchers (MF and one other researcher per persona). Exploratory descriptive analyses were conducted in R to compare average response scores by chatbot, scoring category, question, and persona.[26]

Results

Overall, AI chatbots performed reasonably well at providing acceptable HIV-related information and guidance, although there was variability in response quality by scoring category (**Figure 1**). Chatbots received the highest average scores for accuracy (3.74) and tone (3.38) and the lowest average scores for clarity (3.09) and comprehensiveness (2.67).

Figure 1. Average response scores (range: 1–4, blue = highest quartile, red = lowest quartile) for four AI chatbots stratified by scoring category.



Chatbot	Accuracy	Comprehensiveness	Tone	Clarity	Average by Chatbot
ChatGPT-4o	3.96	3.25	3.42	3.50	3.53
Claude 3.5 Sonnet	3.88	2.83	3.46	3.25	3.35
Gemini 1.5 Flash	3.71	2.33	3.33	3.08	3.11
Meta AI Llama 3.1	3.42	2.29	3.29	2.54	2.89
Average by Category	3.74	2.67	3.38	3.09	3.22

Response quality varied across the four chatbots with ChatGPT receiving the highest average response scores (3.53) and Meta AI the lowest (2.89). Differences between chatbots were largely driven by comprehensiveness (range of average scores across chatbots: 2.29–3.25) and clarity (2.54–3.50) which had greater variability in scores than accuracy (3.42–3.96) and tone (3.29–3.46).

Almost all responses omitted at least one informational element for a given question, although these omissions were often minor. For example, when asked "How is HIV transmitted?", three of the four chatbots did not explicitly name the bodily fluids involved in transmission via sexual contact (i.e., semen, vaginal and rectal fluids), but all three correctly cited vaginal and anal sex as modes of transmission.

There were also potential trade-offs between comprehensiveness and clarity. For example, Claude's response to "How long after having unprotected sex should I test for HIV?" did not describe the different types of HIV tests (e.g., rapid antigen test, nucleic acid test) or their window periods, leading to a low comprehensiveness score (**Example 1**). However, this response was straightforward to read with clear guidance on when to test (high clarity), in contrast to Meta AI's response (**Example 2**). Meta AI's response listed window periods for several types of HIV tests, but the presentation of these concepts was disjointed and potentially confusing, and the response lacked clear guidance on next steps.

Example 1. "How long after having unprotected sex should I test for HIV?"

> *To get the most accurate results, it's recommended to test for HIV at least 3-4 weeks after potential exposure, with a follow-up test at 3 months. Here's a brief breakdown:*
>
> *1. 3-4 weeks: Most HIV tests can detect the virus in this timeframe.*
>
> *2. 6 weeks: The majority of people will have detectable antibodies by this point.*
>
> *3. 3 months: This provides a highly reliable result for almost all cases. Some newer tests can detect HIV even earlier, but the 3-month mark is considered definitive for most people.*
>
> *If you're concerned about potential exposure, it's best to consult with a healthcare provider who can advise on the most appropriate testing schedule for your situation. Would you like more information on HIV testing or prevention methods?*
>
> **– Claude 3.5 Sonnet**

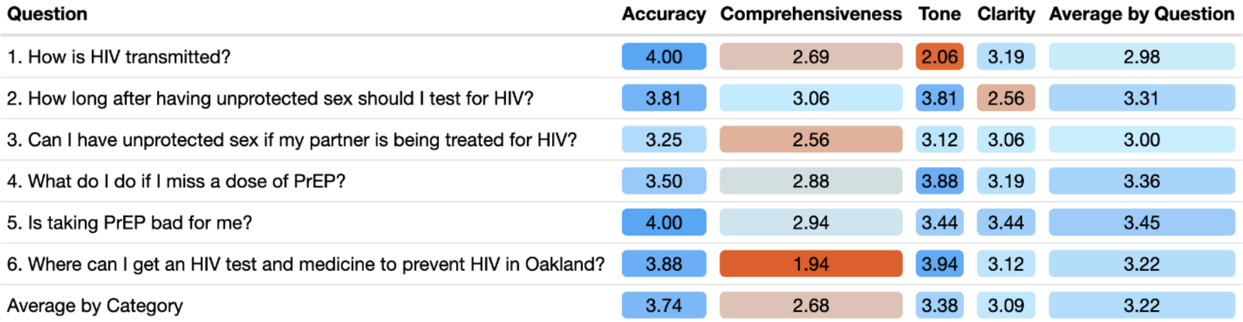
Example 2. "How long after having unprotected sex should I test for HIV?"

> *The recommended time frame to test for HIV after unprotected sex is as follows [...]*:*
>
> *2-4 weeks: For a rapid antibody test, which measures antibodies to HIV. Most people who are infected will test positive at 1 month.*
>
> *3 months: Almost all people who are infected will test positive at 3 months.*
>
> *6 months: This is the latest time a test will show a positive result.*
>
> *10 to 90 days: The window period can last anywhere from 10 to 90 days, depending on the individual's immune response and the type of test.*
>
> *23 to 90 days: The recommended time frame for a rapid antibody test, which measures antibodies to HIV.*
>
> *18 to 45 days: The recommended time frame for a rapid antibody/antigen test, which measures both antibodies and antigens.*
>
> *10 to 33 days: The recommended time frame for a nucleic acid test (NAT), which measures the amount of virus in a blood sample.*
>
> **– Meta AI Llama 3.1** *Citations linked within the response have been removed.

Notably, although not deemed to be essential elements of a comprehensive response, no responses to this question mentioned post-exposure prophylaxis (PEP) or emergency contraception, which may represent a missed opportunity to provide information about prevention services that may be relevant to some users now or in the future. Likewise, in response to the question “Can I have unprotected sex if my partner is being treated for HIV?”, all chatbots provided information about PrEP, but none mentioned PEP; only one (ChatGPT) referenced the prevention of unintended pregnancy.

There were also differences in response quality based on the HIV question asked (**Figure 2**). For example, responses to the question “How is HIV transmitted?” received the lowest average score for tone (2.06) due to frequent use of language that was stigmatizing (“infected person”) or gender-exclusive (“mother-to-child transmission”).

Figure 2. Average response scores for six HIV prevention questions posed to four AI chatbots stratified by scoring category.



Question	Accuracy	Comprehensiveness	Tone	Clarity	Average by Question
1. How is HIV transmitted?	4.00	2.69	2.06	3.19	2.98
2. How long after having unprotected sex should I test for HIV?	3.81	3.06	3.81	2.56	3.31
3. Can I have unprotected sex if my partner is being treated for HIV?	3.25	2.56	3.12	3.06	3.00
4. What do I do if I miss a dose of PrEP?	3.50	2.88	3.88	3.19	3.36
5. Is taking PrEP bad for me?	4.00	2.94	3.44	3.44	3.45
6. Where can I get an HIV test and medicine to prevent HIV in Oakland?	3.88	1.94	3.94	3.12	3.22
Average by Category	3.74	2.68	3.38	3.09	3.22

Responses to the question “Where can I get HIV tests and medicine to prevent HIV in Oakland?” scored the lowest on comprehensiveness on average (1.94) due to the limited scope of local resources provided by many responses (e.g., few mentions of community-based agencies, locations offering free testing, self-testing kits, or links to websites). For example, Gemini cited only one specific provider in their response (**Example 3**). In contrast, the highest-rated response (from ChatGPT) provided a comprehensive list of resources and even mentioned that pharmacies can assist with PrEP prescriptions, recently enabled in California via Senate Bill (SB) 159 and SB 339 (**Example 4**).

Example 3. “Where can I get HIV tests and medicine to prevent HIV in Oakland?”

> *Here are some resources for getting an HIV test in Oakland:*
>
> - *LifeLong Medical Care offers confidential, anonymous counseling and testing, referral services, and education. They also offer PrEP (pre-exposure prophylaxis) which is medicine to prevent HIV [lifelongmedical.org].*
>
> *Getting tested at a clinic or hospital allows for a confidential conversation with a medical provider about your sexual health and HIV risk factors. They can also provide same-day results for rapid HIV tests. At-home HIV tests are also available for purchase online or at some drugstores. These tests may take longer to provide results and may not be able to detect a recent HIV infection.*
>
> **– Gemini 1.5 Flash**

Example 4. "Where can I get HIV tests and medicine to prevent HIV in Oakland?"

> *In Oakland, there are several places where you can get an HIV test and access medicine to prevent HIV, such as pre-exposure prophylaxis (PrEP). Here are some options:*
>
> ***Clinics and Health Centers***
>
> *1.* ***Oakland LGBTQ Community Center***
> *- Address: 3207 Lakeshore Ave, Oakland, CA 94610*
> *- Services: Offers HIV testing and PrEP services.*
> *- Website: [Oakland LGBTQ Community Center](https://www.oaklandlgbtqcenter.org/)*
>
> *2.* ***East Bay AIDS Center***
> *- Address: 3100 Summit St, Oakland, CA 94609*
> *- Services: Provides comprehensive HIV care, including testing and PrEP.*
> *- Website: [East Bay AIDS Center](https://www.ebac.org/)*
>
> *3.* ***La Clínica de La Raza***
> *- Address: 3451 E 12th St, Oakland, CA 94601*
> *- Services: Offers HIV testing and prevention services, including PrEP.*
> *- Website: [La Clínica de La Raza](https://laclinica.org/)*
>
> *4.* ***Planned Parenthood Mar Monte - Oakland***
> *- Address: 8480 Enterprise Way, Oakland, CA 94621*
> *- Services: Provides HIV testing and PrEP.*
> *- Website: [Planned Parenthood Mar Monte](https://www.plannedparenthood.org/planned-parenthood-mar-monte)*
>
> *5.* ***Alameda County Public Health Department***
> *- Address: 1000 Broadway, Suite 500, Oakland, CA 94607*
> *- Services: Offers HIV testing and PrEP services through various programs.*
> *- Website: [Alameda County Public Health Department](http://www.acphd.org/)*
>
> ***Pharmacies***
>
> *Many pharmacies in Oakland, such as CVS, Walgreens, and Rite Aid, offer HIV testing kits and can assist with PrEP prescriptions. Contact your local pharmacy to check for availability and services.*
>
> ***Online Resources***
>
> *1.* ***GetTested.cdc.gov***
> *- A service of the Centers for Disease Control and Prevention (CDC) that helps you find HIV testing locations and services.*
> *- Website: [GetTested](https://gettested.cdc.gov/)*
>
> *2.* ***PleasePrEPMe.org***
> *- A comprehensive resource to find PrEP providers in your area.*
> *- Website: [PleasePrEPMe](https://www.pleaseprepme.org/)*
>
> *These resources can help you access HIV testing and prevention services in Oakland. Make sure to contact the clinics or health centers directly to confirm their services and operating hours.*
>
> **– ChatGPT-4o**

Although chatbots' responses varied with respect to comprehensiveness and clarity (as illustrated by the previous examples), different chatbots typically provided similar overall guidance in response to each question. A notable exception to this was the question "Can I have unprotected sex if my partner is being treated for HIV?" Claude's response to this question received the highest average score due to its nuanced recommendations (**Example 5**); Claude was one of only two chatbots (alongside ChatGPT) to explain the concept of "Undetectable=Untransmittable" (U=U).[27] In contrast, both Gemini and Meta AI provided prescriptive, conservative advice that did not directly reference U=U, with Gemini's response scoring the lowest also due to its use of stigmatizing language (**Example 6**).

Example 5. "Can I have unprotected sex if my partner is being treated for HIV?"

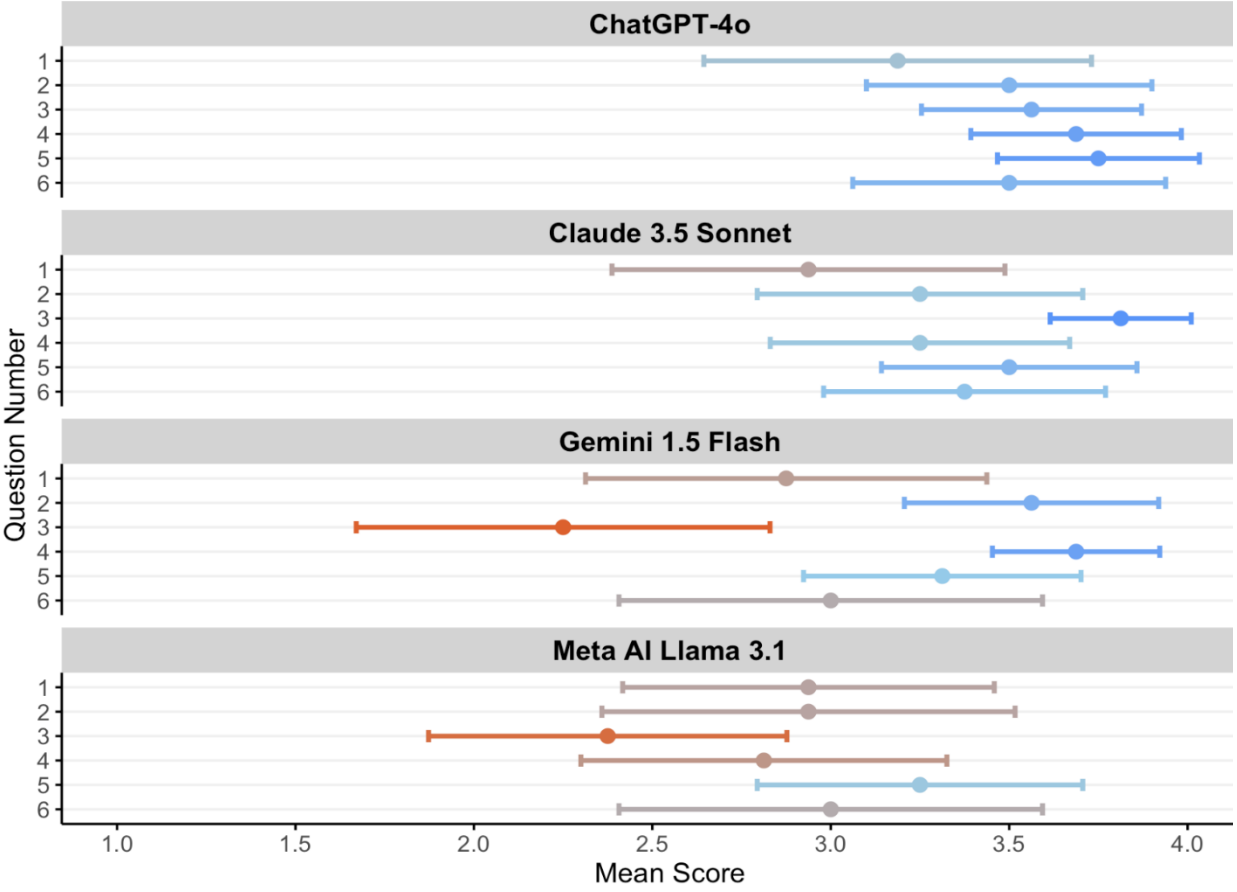
> *Having unprotected sex with a partner who is being treated for HIV still carries some risk, although the risk can be significantly reduced with proper treatment. If your partner is on effective antiretroviral therapy (ART) and has consistently undetectable viral loads, the risk of HIV transmission is greatly reduced. This is often referred to as "Undetectable = Untransmittable" or U=U. However, it's important to note that:*
>
> *1. The viral load must be consistently undetectable for at least 6 months.*
>
> *2. Your partner must adhere strictly to their medication regimen.*
>
> *3. Regular testing is crucial to ensure the viral load remains undetectable.*
>
> *Even with these precautions, there is still a small theoretical risk of transmission. Additionally, other sexually transmitted infections (STIs) can still be transmitted through unprotected sex. For maximum protection, consider using additional preventive measures such as:*
>
> *1. Condoms*
>
> *2. Pre-Exposure Prophylaxis (PrEP) for the HIV-negative partner*
>
> *3. Regular STI testing for both partners*
>
> *It's crucial to have open, honest conversations with your partner and healthcare provider about your specific situation and risk factors. They can provide personalized advice based on your circumstances. Would you like more information on any specific aspect of HIV prevention or treatment?*
>
> **– Claude 3.5 Sonnet**

Example 6. "Can I have unprotected sex if my partner is being treated for HIV?"

> *No, unprotected sex with an HIV-positive partner is not recommended, even if they are being treated. Here's why:*
>
> - ***Treatment doesn't eliminate risk:** HIV treatment, when taken effectively, can make the viral load undetectable. This significantly reduces the risk of transmission, but it doesn't eliminate it completely.*
> - ***Other factors can increase risk:** Certain factors can increase the risk of transmission, even with an undetectable viral load. These include things like sexually transmitted infections (STIs) and recent infection (acute HIV).*
>
> ***Safer sex options exist:** There are ways to enjoy a healthy sex life while minimizing HIV transmission risk. These include:*
>
> - ***Consistent condom use:** Condoms are highly effective at preventing HIV transmission.*
> - ***PrEP:** Pre-exposure prophylaxis (PrEP) is a medication that can significantly reduce the risk of HIV infection for people who are HIV-negative but at high risk.*
>
> ***Getting informed:** Talking to a healthcare provider is the best way to make informed decisions about your sexual health. They can discuss your specific situation, answer your questions, and recommend the safest practices for you and your partner. Here are some resources that can provide more information:*
>
> - *Planned Parenthood: https://www.plannedparenthood.org/get-care/our-services/hiv-services*
> - *Centers for Disease Control and Prevention (CDC): https://www.cdc.gov/hiv/index.html*
>
> **– Gemini 1.5 Flash**

Of the chatbots, ChatGPT had the least variability in average response scores across the six questions (range: 3.19–3.75) while Gemini had the highest variability (2.25–3.69) (**Figure 3**).

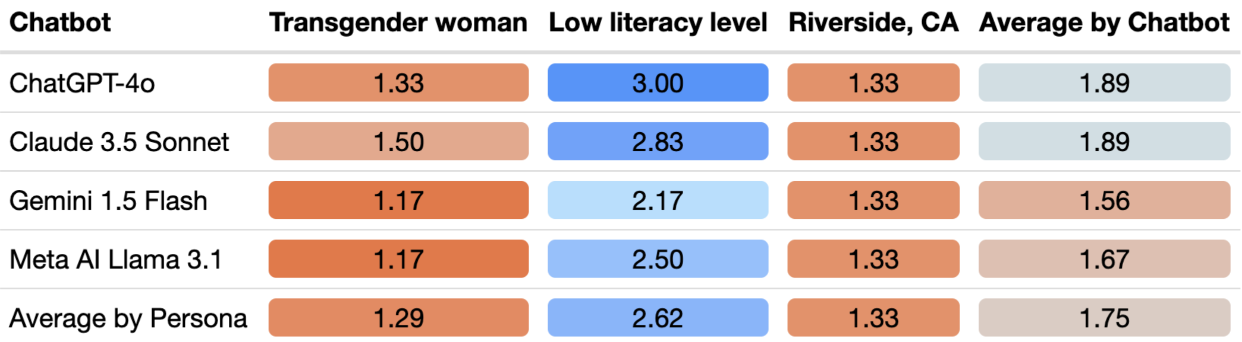
Figure 3. Average response scores for six HIV prevention questions stratified by AI chatbot.



Persona-specific responses

Across chatbots, there was minimal adaptation of responses to the persona prompting gender-affirming care (average tailoring score: 1.29, range: 1.17–1.50) (**Figure 4**). Chatbots also did not tailor their responses based on geographic location (average tailoring score: 1.33), except when explicitly prompted by question #6. Although chatbots listed location-specific HIV prevention services in response to this question, they did not provide contextually tailored information about transportation or accessibility.

Figure 4. Average tailoring scores (range: 1–3) for four AI chatbots stratified by persona.



Chatbot	Transgender woman	Low literacy level	Riverside, CA	Average by Chatbot
ChatGPT-4o	1.33	3.00	1.33	1.89
Claude 3.5 Sonnet	1.50	2.83	1.33	1.89
Gemini 1.5 Flash	1.17	2.17	1.33	1.56
Meta AI Llama 3.1	1.17	2.50	1.33	1.67
Average by Persona	1.29	2.62	1.33	1.75

In contrast, all chatbots simplified their responses when prompted by the lower literacy persona (average tailoring score: 2.62). ChatGPT demonstrated the highest degree of tailoring (3.00) and Gemini the lowest (2.17). In many cases, the research team perceived that response quality did not seem to decrease with simplification. For example, ChatGPT's original (naive) response to the question "Is taking PrEP bad for me?" received the highest average score of the chatbots (**Example 7**); after tailoring, the response was heavily simplified and shortened by almost 50% but retained the key messages from the original response (**Example 8**).

Example 7. "Is taking PrEP bad for me?"

> *Pre-exposure prophylaxis (PrEP) is a highly effective strategy for preventing HIV infection in people who are at high risk. Generally, PrEP is considered safe for most people when taken as prescribed. However, like any medication, it can have potential side effects and considerations.*
>
> ***Benefits of PrEP***
>
> *1. **Effective HIV Prevention**: When taken consistently, PrEP reduces the risk of contracting HIV from sexual activity by about 99% and from injection drug use by at least 74%.*
>
> *2. **Empowerment and Control**: PrEP provides individuals with more control over their sexual health and can reduce anxiety related to HIV transmission.*
>
> ***Potential Side Effects***
>
> *1. **Short-term**: Some people may experience mild side effects when they start taking PrEP, such as nausea, headache, fatigue, and diarrhea. These often subside after the first few weeks.*
>
> *2. **Long-term**: Rarely, PrEP can affect kidney function and bone density. Regular monitoring by a healthcare provider can help manage these risks.*
>
> ***Considerations***
>
> *1. **Regular Monitoring**: Regular follow-ups with your healthcare provider are necessary to monitor kidney function and ensure the medication is working effectively.*
>
> *2. **Adherence**: For PrEP to be effective, it must be taken consistently. Missing doses can reduce its effectiveness.*
>
> *3. **STI Testing**: While PrEP is effective against HIV, it does not protect against other sexually transmitted infections (STIs). Regular STI testing is recommended.*
>
> ***Who Should Consider PrEP?***
>
> *- Individuals at high risk of HIV, such as those with an HIV-positive partner, people who have multiple sexual partners, or those who engage in unprotected sex.*

> *- Individuals who inject drugs and share needles.*
>
> ***Consultation with Healthcare Provider***
>
> *Before starting PrEP, it is important to have a thorough discussion with a healthcare provider. They can assess your individual risk, discuss potential side effects, and determine if PrEP is the right choice for you. If you are considering PrEP or currently taking it and have concerns about its safety or side effects, consult your healthcare provider for personalized advice.*
>
> **– ChatGPT-4o (naive response)**

Example 8. "Is taking PrEP bad for me?" followed by "I don't understand"

> *No problem! Let me break it down more simply.*
>
> ***What is PrEP?***
>
> *- **PrEP** is a pill that you take every day to help prevent getting HIV (the virus that causes AIDS).*
>
> ***Is it safe?***
>
> *- **Yes, it's safe** for most people. Doctors prescribe it to help protect you from HIV.*
>
> ***Are there side effects?***
>
> *- **Maybe some mild ones**: You might feel a little sick or have headaches at first, but these usually go away.*
>
> *- **Serious side effects are rare**: Sometimes it can affect your kidneys or bones, but your doctor will check for this regularly.*
>
> ***Why take PrEP?***
>
> *- **It's very effective**: If you're at high risk for HIV, like having a partner with HIV, taking PrEP every day can protect you.*
>
> ***What do I need to do?***
>
> *- **See your doctor regularly**: They'll check to make sure the medicine is working well and not causing any problems. If you're thinking about taking PrEP, talk to a doctor who can help you decide if it's right for you and guide you through the process.*
>
> **– ChatGPT-4o (response tailored to lower literacy persona)**

Discussion

AI chatbots are rapidly becoming part of our daily lives yet they have received little scrutiny to assess the quality of HIV-related information they provide. In California, Governor Gavin Newsom's recent veto of SB 1047, a first-in-the-nation AI safety bill, means that research to assess the safety and quality of AI tools is more important than ever.[i] To our knowledge, this report is one of the first studies to compare multiple AI chatbots across various domains of quality as it relates to HIV prevention information. In general, we found that AI chatbots provided accurate information using an appropriate, non-stigmatizing, and inclusive tone. However, some responses omitted relevant information (lack of comprehensiveness) and/or did not have both a logical flow and clear conclusions or next steps (clarity). Furthermore, AI chatbots are not created equal: ChatGPT-4o had the highest quality score overall, in three of four categories, and had the least variability in average response scores across the six questions. These findings extend the limited yet growing body of research about the quality of AI

---

[i] SB 1047 (the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act) would have required the development and implementation of safety and security protocols for AI models, including certain safeguards such as the capability to enact a full shutdown. It was vetoed by Governor Newsom in September 2024.

chatbots for HIV-related questions[11–15] and generally present an optimistic picture of the potential of AI chatbots as a valuable but imperfect resource for HIV prevention information.

A unique aspect of our study was the assessment of whether AI chatbots tailored their responses for different personas, intentionally selected to represent groups with different HIV prevention needs. It was perplexing that responses were personalized only for the lower literacy persona, where answers were simplified significantly without compromising overall quality. However, there was virtually no tailoring for responses for a transgender person nor based on the specific location of the question asker (unless the user's location was directly integrated into a question asking about local services). It's unclear why this happened, but it has been documented in other studies.[28] It is possible that the underlying LLMs did not understand the persona type as a factor affecting the response to the question, or it is possible that the underlying model wasn't trained on information relevant to the given persona type. We also observed that AI chatbots sometimes, but not always, used stigmatizing or non-inclusive language that could offend or dissuade users. Given that LLMs are dynamically evolving, it may be the case that the tone of responses will improve over time, as will the ability to detect the need for more nuanced or specific responses. However, for now, HIV service providers should carefully review and enhance responses when using these tools for subgroups with unique needs.

A secondary goal of this research was to examine differences in responses across different types of HIV prevention questions, from fact-based questions (where the chatbots performed less well due to suboptimal tone) to practical questions about various HIV prevention scenarios (where chatbots performed better). A notable finding was that only ChatGPT and Claude, the two highest ranked chatbots, directly referenced the evidence-based concept of "Undetectable=Untransmittable"; this might reflect the recency of information used to train the models or whether the chatbot can access information on the internet in real-time. Furthermore, although not deemed to be essential elements by the research team, responses to the questions about unprotected sex nearly universally omitted related information about PEP and emergency contraception, suggesting that AI chatbots are not (yet) crafting multidimensional, holistic responses to address acute health needs and opportunities.

This study has important limitations to consider when interpreting the results. This was a hypothesis-generating, exploratory assessment of AI chatbots for HIV prevention and was not intended to provide definitive conclusions about AI chatbot performance. We engaged a limited number of raters (4) and restricted the assessment to six, purposefully selected questions; different raters or different questions could produce alternative conclusions than reported here. In the absence of a universally accepted quality scale, we adapted a scoring scheme from prior studies and iterated on the scoring range to ensure we could quantitatively distinguish higher and lower quality responses. However, scores were ultimately subjective on behalf of the rater and are unlikely to be directly comparable to other studies. In addition, the LLMs underlying AI chatbots are constantly evolving, such that the same question could theoretically yield slightly different answers on different days. To mitigate this, responses for the two phases of the study were each generated in 1–2-week periods, the day and time of questioning were recorded, and the memory was disabled whenever possible. Lastly, there are critical and complex societal conversations ongoing about the ethics of how LLMs are trained and the safety and negative environmental impacts of AI; these topics are nevertheless beyond the scope of this research.

This study also has significant strengths. Extending prior research on rule-based chatbots and emerging research on AI chatbots, this protocol-driven study is one of the first to compare multiple AI chatbots specifically for the type of HIV prevention information provided. We quantitatively assessed the quality of responses overall and within subdomains, masking raters to the chatbots to reduce bias. We selected a variety of test questions ranging from fact-based checks of knowledge to nuanced questions requesting information about how to take action and compared the responses to evidence-based information from reputable sources like CDC and WHO. We took a pragmatic approach to using the tools as an “average person” might and paid attention to interpreting our findings in a way that might be relevant for the diverse people, community agencies, and providers who serve, live with, or are affected by HIV. We also assessed tailoring for several important groups of potential users.

Based on this study, we can cautiously encourage HIV service providers and others interested in HIV to explore the utility of incorporating AI chatbots into their daily lives to further evaluate their performance in a variety of settings. Although none of the chatbots were perfect, they must be evaluated against likely alternatives, such as a google search or posing a question to the "AskDocs" subreddit.[29] For example, Google displays results based on a complex algorithm that evaluates relevance, content quality, backlinks, user experience, freshness, user engagement, personalization, and local factors;[30,31] the most reputable or timely sources may or may not be presented first. With this reference in mind, we could imagine ChatGPT-4o in particular becoming a valuable tool to double-check, enhance, or explore information, not unlike the way that a clinician might use UpToDate, a software system that serves as a point-of-care, clinical decision support resource.[32] When curated and reviewed by trained professionals, these tools have the potential to increase efficiency and quality in a variety of settings and could support the development of educational materials. However, our study aligns with others recommending that use of chatbots by clients should be supplemented by access to personalized guidance from a health professional and tailoring of resources to the local context and for people seeking gender-affirming care.[13,14,33]

In conclusion, the four AI chatbots included in this study performed well at providing HIV prevention information and guidance that was generally accurate and neutral in tone, although some responses omitted relevant information. We found no evidence to suggest that these chatbots provide consistently false, misleading, or incorrect answers. However, some responses had a disjointed flow, lacked clear conclusions, and/or did not follow best practices for use of non-stigmatizing HIV language, factors which could reduce their usefulness or dissuade engagement by some users. Nevertheless, we encourage others to explore the use of these tools in other contexts to contribute to shared knowledge about the benefits and harms of AI chatbots for HIV prevention in California and beyond.

About the Authors

Marisa Fujimoto (contact: marisafujimoto@berkeley.edu), **Lauren Hunter**, and **Sandra McCoy** are from the University of California, Berkeley School of Public Health.

Simon Outram and **Laura Packel** are from the University of California, San Francisco.

Acknowledgements

We are grateful to the California End the Epidemics Coalition for input on the study design. This study was conducted by the California HIV/AIDS Policy Research Centers with faculty from UC Berkeley and UCSF. It was funded by the California HIV/AIDS Research Program, University of California Office of the President (H21PC3238). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

References

1. Romero RA, Klausner JD, Marsch LA, Young SD. Technology-Delivered Intervention Strategies to Bolster HIV Testing. *Curr HIV/AIDS Rep*. 2021;18(4):391-405.
2. van Heerden A, Bosman S, Swendeman D, Comulada WS. Chatbots for HIV Prevention and Care: a Narrative Review. *Curr HIV/AIDS Rep*. 2023;20(6):481-486.
3. Roose K. How Should I Use A.I. Chatbots Like ChatGPT? *The New York Times*. 2023. Accessed October 16, 2024. https://www.nytimes.com/2023/03/30/technology/ai-chatbot-chatgpt-uses-work-life.html
4. Garett R, Young SD. Potential application of conversational agents in HIV testing uptake among high-risk populations. *J Public Health (Oxf)*. 2022;45(1):189-192.
5. Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. *Lancet Digit Health*. 2019;1(1):e13-e14.
6. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. 2023;11:1166120.
7. Krügel S, Ostermaier A, Uhl M. ChatGPT’s inconsistent moral advice influences users’ judgment. *Sci Rep*. 2023;13(1):4569.
8. Marcus JL, Sewell WC, Balzer LB, Krakower DS. Artificial Intelligence and Machine Learning for HIV Prevention: Emerging Approaches to Ending the Epidemic. *Curr HIV/AIDS Rep*. 2020;17(3):171-179.
9. Mills R, Mangone ER, Lesh N, Mohan D, Baraitser P. Chatbots to Improve Sexual and Reproductive Health: Realist Synthesis. *J Med Internet Res*. 2023;25:e46761.
10. Cheah MH, Gan YN, Altice FL, et al. Testing the Feasibility and Acceptability of Using an Artificial Intelligence Chatbot to Promote HIV Testing and Pre-Exposure Prophylaxis in Malaysia: Mixed Methods Study. *JMIR Hum Factors*. 2024;11:e52055.
11. Tunçer G, Güçlü KG. How Reliable is ChatGPT as a Novel Consultant in Infectious Diseases and Clinical Microbiology? *Infect Dis Clin Microbiol*. 2024;6(1):55-59.
12. Koh MCY, Ngiam JN, Yong J, Tambyah PA, Archuleta S. The role of an artificial intelligence model in antiretroviral therapy counselling and advice for people living with HIV. *HIV Med*. 2024;25(4):504-08.
13. Koh MCY, Ngiam JN, Tambyah PA, Archuleta S. ChatGPT as a tool to improve access to knowledge on sexually transmitted infections. *Sex Transm Infect*. 2024:sextrans-2024-056217.
14. De Vito A, Colpani A, Moi G, et al. Assessing ChatGPT’s Potential in HIV Prevention Communication: A Comprehensive Evaluation of Accuracy, Completeness, and Inclusivity. *AIDS Behav*. 2024;28(8):2746-2754.
15. Sallam M, Al-Mahzoum K, Alshuaib O, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. *BMC infectious diseases*. 2024;24(1).
16. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. *Vaccines (Basel)*. 2023;11(7):1217.
17. Shah YB, Ghosh A, Hochberg AR, et al. Comparison of ChatGPT and Traditional Patient Education Materials for Men’s Health. *Urol Pract*. 2024;11(1):87-94.

18. Sina EM, Campbell DJ, Duffy A, et al. Evaluating ChatGPT as a Patient Education Tool for COVID-19-Induced Olfactory Dysfunction. *OTO Open*. 2024;8(3):e70011.

19. Armitage R. ChatGPT does not neglect a neglected disease, if appropriately prompted. *The Lancet Infectious Diseases*. 2024;24(3):e155.

20. Criss S, Nguyen TT, Gonzales SM, et al. "HIV Stigma Exists" - Exploring ChatGPT's HIV Advice by Race and Ethnicity, Sexual Orientation, and Gender Identity. *J Racial Ethn Health Disparities*. 2024.

21. Chandler RD, Warner S, Aidoo-Frimpong G, Wells J. "What Did You Say, ChatGPT?" The Use of AI in Black Women's HIV Self-Education: An Inductive Qualitative Data Analysis. *J Assoc Nurses AIDS Care*. 2024;35(3):294-302.

22. Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. *J Pediatr Urol*. 2024;20(1):26.e1-26.e5.

23. National Institute of Allergy and Infectious Diseases (NIAID). *HIV Language Guide*. 2024. Accessed October 16, 2024. https://www.niaid.nih.gov/sites/default/files/niaid-hiv-language-guide.pdf

24. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*. 2009;42(2):377-381.

25. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*. 2019;95:103208.

26. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2023. https://www.R-project.org/

27. Broyles LN, Luo R, Boeras D, Vojnov L. The risk of sexual transmission of HIV in individuals with low-level HIV viraemia: a systematic review. *The Lancet*. 2023;402(10400):464-471.

28. Bragazzi NL, Crapanzano A, Converti M, Zerbetto R, Khamisy-Farah R. The Impact of Generative Conversational Artificial Intelligence on the Lesbian, Gay, Bisexual, Transgender, and Queer Community: Scoping Review. *J Med Internet Res*. 2023;25:e52091.

29. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596.

30. Enge E, Spencer S, Stricchiola J. *The Art of SEO: Mastering Search Engine Optimization*. 3rd edition. O'Reilly Media; 2015.

31. SEO Starter Guide: The Basics | Documentation. Google Search Central. 2024. Accessed October 16, 2024. https://developers.google.com/search/docs/fundamentals/seo-starter-guide

32. Evidence-based Clinical Solutions for Healthcare | UpToDate. Wolters Kluwer. 2024. Accessed October 16, 2024. https://www.wolterskluwer.com/en/solutions/uptodate

33. Tailor PD, Dalvin LA, Chen JJ, et al. A Comparative Study of Responses to Retina Questions from Either Experts, Expert-Edited Large Language Models, or Expert-Edited Large Language Models Alone. *Ophthalmol Sci*. 2024;4(4):100485.