# CHIA: A Chatbot for HIV and PrEP Counseling Using Retrieval-Augmented Generation and Motivational Interviewing

Amaris Grondin

*Brown University*

Advisor: Ellie Pavlick

Second Reader: Jun Tao

This thesis is based on a larger body of work conducted under the guidance of Professor Tao and Professor Pavlick.

April 15, 2025

**Abstract**

This work aims to address the issues of stigma, cost, and access in HIV prevention and PrEP education by developing a chatbot specifically designed for men who have sex with men (MSM) and economically disadvantaged communities. This chatbot is a non-judgmental, culturally sensitive, and multi-lingual space where users can ask questions and express concerns about HIV without fear of stigma or judgment. Users will have access to a free initial "consultation" to address their doubts regarding HIV and receive guidance on connecting with a counselor. By facilitating access to support, the chatbot aims to encourage open discussions about sexual health, reduce HIV transmission, and promote broader acceptance and use of PrEP as a preventive measure.

# Contents

# Chapter 1

# Introduction

The human immunodeficiency virus (HIV) epidemic continues to disproportionately impact marginalized communities, particularly men who have sex with men (MSM) (Mayer et al., 2021) and individuals in economically disadvantaged regions (Pellowski et al., 2013). While advancements such as pre-exposure prophylaxis (PrEP) have made HIV prevention more accessible, uptake of these preventative measures remains low among the populations most at risk. PrEP, a daily oral medication that significantly reduces the risk of HIV infection, is a promising solution to impact the epidemic positively. However, barriers such as stigma, cost, and access to healthcare prevent the effective distribution of PrEP and cause slow PrEP uptake, especially within communities of color and lower socioeconomic status (Sullivan et al., 2024).

One of the primary obstacles to widespread PrEP utilization is PrEP-related stigma. Because PrEP is associated with HIV treatment, individuals who take the medication are often perceived as HIV-positive or engaging in behaviors considered socially unacceptable, such as promiscuity or drug use. This stigma not only discourages people from seeking out PrEP but also intensifies the disparities in HIV prevention efforts, further marginalizing those already vulnerable to infection (Golub, 2018). In addition to the stigma that comes with PrEP, there is an extremely high need for people to get counseling. However, patients with

HIV often face challenges when accessing healthcare such as financial, language, and cultural barriers (Bouabida et al., 2023). Findings have also shown that participants prefer talking to a chatbot when they feel embarrassed to avoid being judged by others (Diwanji et al., 2025). To address these challenges, innovative, stigma-free, and culturally sensitive solutions are essential to promote PrEP awareness and utilization among the most affected populations.

# Chapter 2

# Related Work

This paper seeks to address this issue by building an AI chatbot that aims to make HIV and PrEP counseling more accessible and reduce the inequity in healthcare access. Past research has leveraged Retrieval-Augmented Generation (RAG) to build empathetic chatbots for supporting victims of sexual harassment (Vakayil et al., 2024). There has also been previous work developing an AI-based chatbot to promote HIV self-management (Ma et al., 2024). However, this focuses on providing expert-validated information without using methods shown to elicit behavior change.

In addition to leveraging RAG and focusing on accuracy, with expert-validated datasets and tests, CHIA also integrates Motivational Interviewing techniques and function-calling capabilities. The latter enables it to search for nearby providers, assess HIV risk and status of change, and notify research assistants if users require support, thus motivating the use of Autogen agents in this project.

Traditionally, Autogen has served as a supporting tool for specific tasks, such as enhancing academic papers (Porsdam Mann et al., 2023) or developing AI-based senior care technologies (Yun et al., 2024). However, our approach seeks to use Autogen agents as direct substitutes for human interaction, simulating the role of a counselor. This chatbot will not only assist with information retrieval or task completion but will also engage users in meaningful and

empathetic conversations, mirroring the support provided by a human counselor.

# Chapter 3

# Methods

## 3.1 Overall Architecture

Before diving into each element of CHIA, it is important to understand the overall architecture of the chatbot (Figure A.1). CHIA is a full-stack web application that uses React and TypeScript in the front end and Python for the back end. The front end is built upon Chatbot UI's open-source UI. WebSockets were used to establish communication between the front and back end. To facilitate this communication, we are using FastAPI, a web framework for building APIs, which supports WebSockets. To incorporate these WebSocket connections, we had to change the Autogen source code to enable the chat to appear on the web app rather than the console. The code is then contained within a Docker container and pushed to AWS ECS services to be publicly available.

## 3.2 Autogen Agents

Figuring out the optimal configuration for Autogen agents for this project was challenging. The most intuitive way of doing this was to have a single counselor agent using these different functions as potential tools that it could use to answer the patient. However, Autogen requires at least two agents for a single function call. Therefore, we decided to create an Autogen

group chat, which allows more than one agent to contribute to the conversation.

At first, we found it most effective to have two agents (one User Proxy Agent to execute the function – the "counselor" – and another Assistant agent to suggest the function to the User Proxy – the "counselor assistant agent") for each function call. We thought this approach would be most successful as each pair of agents would only have one task, minimizing the confusion. Although agents were able to activate at the correct times when the user prompted them with their question, some inputs were ambiguous, activating different agents at once. This would result in different agents answering each other at the same time, instead of the user. This resulted in a very unnatural user experience. We tried to use the `allow_repeat_speaker` parameter to ensure agents could only answer the user, rather than other agents. However, this did not work as some agents would answer before the one who was really supposed to be activated, resulting in an incomplete answer.

To have a more cohesive chat, we decided to see what would happen if instead of having two agents per function call, we would have a total of two agents in addition to the user – a counselor agent and a counselor assistant agent. The system message for the counselor is now much longer following prompt engineering techniques, such as role, instruction, and conditional prompting (Figure A.2).

We registered the counselor assistant with the counselor for each function separately with an appropriate system message for that function. When we tested this with GPT-4o-mini, errors occurred randomly, indicating that the model produced invalid content. We then switched to GPT-4o, which worked much more smoothly and provided the expected user experience from a chatbot – a fluid chat with appropriate answers to messages. We now use our custom GPT-4o model.

## 3.3 Embedding Data

Two embedding datasets are used to improve CHIA's performance. The first is an HIV

and PrEP document with questions and answers (Figure A.3). This retrieval happens with Lang Chain's Retrieval QA, which creates a chain for question answering against an index each time a user asks a question about HIV or PrEP. If the user asks a general information question about information that is not contained in the dataset, the agent will retrieve information from its own knowledge. These instructions are given as a prompt to the chain and leverage our custom GPT-4o model. This retrieval gets activated via a function call executed by the counselor agent and suggested by the counselor assistant agent when the question fits the requirements.

The second embedding dataset that is used is a motivational interviewing dataset, which was processed by research assistants in Professor Tao's lab at the Warren Alpert Medical School from real counseling sessions sourced from Hugging Face and reviewed by medical doctors (Figure A.4). Motivational Interviewing has been shown to be a patient-centered approach that effectively elicits behavior change in PrEP uptake (Moitra et al., 2019). Each statement in the conversation is annotated with a label from Motivational Interviewing Skills Code (MISC), which evaluates how well a mental health practitioner responds to those seeking help with their mental health-related issues (Welivita and Pu, 2022). To provide the chatbot with MI skills, we used preference fine-tuning, more specifically Open AI's Direct Preference Optimization (DPO), which allows us to fine-tune models based on prompts' preferred and non-preferred outputs, which were assigned depending on the label associated with the answer to the statement (Figure A.5). This resulted in a custom fine-tuned model that we used for all OpenAI calls.

## 3.4 Function Calling

In addition to enhancing the chatbot's abilities with embedding data, CHIA also required additional functionalities that went further than basic LLM abilities. This justifies the choice of using Autogen, which effectively leveraged Retrieval-Augmented Generation (RAG),

allowing the model to acquire knowledge outside of its training data.

The project required four different types of function-calling, on top of the function that calls Langchain Retrieval. The first one that we implemented was a function that allows the chatbot to assess the user's HIV risk if the user expresses the need to do so. This function has a set of questions to ask the user – if the user says yes to one of them, the chatbot will inform them that they are at risk and ask them if they would like to consult a provider near them. This function uses OpenAI API calls to detect whether the user's response is one of approval, disapproval, or whether they are asking for future information or want to stop the questionnaire so that the answers are not restricted to "yes" or "no". The second function allows the chatbot to find PrEP providers close to the user, who is prompted to provide their zip code. CHIA can then browse the internet to https://preplocator.org/ and return the list of the five closest providers. The third function notifies a research assistant if the user is in further need of human support. It asks the user the type of support they need and how they would like to be contacted, and sends an email to Professor Tao so that someone can reach out to them. The fourth is a function that assesses the user's status of change. This is a single question: 'Are you currently engaging in PrEP uptake on a regular basis?' – it includes five response options that indicate whether the respondent is currently using PrEP and their future intentions regarding its use. The five stages come from the transtheoretical model (TTM), which is the standard-bearer for change (Prochaska et al., 1992). All of these functions are accessible in all languages as the agent detects the language the user speaks and translates its answer into the given language.

## 3.5 Teachability

To make the sessions personalizable to each user, we wanted the chatbot to remember past conversations to mimic real-life exchanges with a counselor. As we will discuss further in the paper, this presents privacy concerns that need to be addressed. To allow the agents

to remember this information, we added Autogen's teachability feature to the counselor agent so that the user's teachings persist across conversations. These teachings are stored in a ChromaDB database. The agent decides whether it should save this information if it answers the following question: "Does any part of the TEXT require the agent to remember any information about the user's personal situation and information? Answer with just one word, yes or no." Currently, the agent doesn't save summaries of old conversations – it saves information about the user, such as their name or some answers to some questions. If we had more time, we would find a way to make the information saved more consistent, as well as ensure that the chatbot is able to summarize previous conversations so that it can mention it when needed during new discussions.

# Chapter 4

# Evaluation

To ensure the quality, reliability, and safety of CHIA's performance, we implemented a multi-dimensional evaluation framework that included both automated and manual assessments. This approach was designed to capture not only technical accuracy but also CHIA's ability to respond empathetically and effectively in real-world, health-related conversations.

Initially, we assessed CHIA's Retrieval-Augmented Generation (RAG) capabilities using the RAGA framework, which evaluates large language model outputs in relation to retrieved content and ground truth answers. RAGA includes several core metrics: context precision, which measures how relevant the retrieved content is to the generated response, and context recall, which assesses whether the retrieval process captured all key elements needed to form a comprehensive answer. Faithfulness evaluates whether the generated content accurately reflects the retrieved materials, ensuring that the model does not introduce hallucinated or misleading information. Additionally, answer relevancy focuses on how well the model's response aligns with the user's query, and answer correctness compares the output against the known correct answer. While useful, this framework proved too restrictive given the open-ended and conversational nature of CHIA's real-world interactions, where definitive ground truths were often unavailable or overly rigid.

To address this limitation, we developed a custom automatic pipeline using GPT-4o.

Instead of relying on static ground truths, this pipeline assessed responses across six criteria tailored to the health counseling context: medical accuracy, completeness, no fabrication, appropriate tone, safety, and contextual grounding. Medical accuracy ensured that responses were aligned with current, evidence-based health information, a critical factor in building trust and avoiding harm. Completeness measured whether the response adequately addressed all aspects of the user's query, ensuring thoroughness and relevance. The no fabrication criterion helped identify hallucinated or unsupported claims, while appropriate tone ensured the language remained respectful, non-stigmatizing, and empathetic. Safety focused on whether the response adhered to harm-reduction principles and avoided potentially dangerous or triggering content. Finally, contextual grounding evaluated whether CHIA's responses demonstrated awareness of the user's previous inputs, contributing to coherence and a sense of personalized care. All evaluations were parsed into structured JSON outputs and stored in Supabase, allowing for consistent testing and monitoring.

In parallel, we applied a similar evaluation structure to assess CHIA's adherence to Motivational Interviewing (MI) principles – an evidence-based approach in health behavior counseling. This involved analyzing responses for key MI techniques, including the use of open-ended questions to encourage user reflection and affirmations to validate user strengths and values. Information sharing was assessed for its balance and clarity, ensuring users were informed without being overwhelmed or directed. The framework also evaluated CHIA's ability to be ambivalent, a core MI strategy that helps users navigate conflicting feelings around behavior change. Eliciting change talk was used to assess whether CHIA encouraged users to express their own motivations for making health-related changes. In addition, reflective listening was examined to determine how well the chatbot mirrored and validated user experiences, and reasoning captured how effectively CHIA integrated information to support motivational and goal-oriented dialogue.

To complement the automated evaluations, research assistants conducted human evaluations of CHIA's performance. These were based on a detailed rubric designed to capture both

factual accuracy and the emotional and interpersonal quality of interactions. Responses were assessed for accuracy, conciseness, and up-to-dateness to ensure users received clear, reliable, and current information. Metrics such as trustworthiness and empathy were included to reflect the relational dimension of CHIA's role as a health support agent. Additionally, MI-aligned criteria such as evocation, collaboration, autonomy support, open-ended questions, reflections, and affirmation were used to evaluate how well CHIA fostered a respectful, empowering, and motivational environment. This combination of metrics provided a holistic view of the chatbot's performance, supporting continuous improvement and ethical deployment in sensitive health contexts. For each of these metrics, research assistants assigned CHIA a score on a 5-point Likert scale. These evaluations were conducted for the latest version of CHIA.

At this early stage, human evaluations were necessary to ensure the chatbot met its basic requirements. For this reason, we have not yet evaluated the effectiveness of the automatic evaluations and will focus on the human evaluations for the results. Automatic evaluations will be useful when the chatbot is used at a larger scale to ensure its continued reliability.

# Chapter 5

# Results

Previous versions of CHIA were not evaluated quantitatively; however, it is important to qualitatively evaluate its different versions to justify the various technical additions in the latest version. To evaluate the effectiveness of different agent configurations, we compared transcripts generated by a multi-agent setup, where a pair of agents is assigned to each function (Figure A.8), with those produced by an architecture using only a single UserProxyAgent and AssistantAgent (Figure A.7). As described earlier, increasing the number of specialized agents introduced ambiguity in the interpretation of the prompt, often leading to responses that do not actually answer the question. This is evident in Figure A.8, where the "Assess HIV Risk" agent responds to multiple user inputs with the same constrained answer – "I'm here to help. Please feel free to ask me if you want to assess your risk." – despite the user's distinct concerns ("I am scared that I might get HIV" and "I don't like using condoms"). In contrast, the simplified two-agent configuration yields responses that are more contextually appropriate, emotionally supportive, and tailored to the user's changing input ("Hi Amaris, thank you for sharing how you feel. It's completely understandable to be concerned. Let's talk about ways to reduce the risk of HIV, like using condoms or considering PrEP, which is a preventive medication. How can I best support you with this? If you're interested, I can help connect you with resources or support options."). This suggests that reducing agent

specialization can enhance coherence and empathy in conversational flow.

The second important choice for CHIA's structure was the choice of the LLM model. Given that we wanted CHIA to incorporate motivational interviewing skills, it is necessary to compare the baseline GPT-4o model, which removed the invalid content errors compared to more cost-effective models such as GPT-4o-mini, with the fine-tuned GPT-4o model. In addition to this, we also compared the results from the fine-tuned GPT-4o model with and without RAG, retrieving information from the HIV/PrEP questions and answers database to verify the necessity of this additional database. The examples show a strong increase in CHIA's performance when each component is added. Looking at Figure A.6, there are clear improvements in MI skills between the fine-tuned GPT-4o model and the base model. One main MI guideline is based on reflection by being compassionate and repeating what they are trying to communicate. To the question "What is the newest HIV incidence by gender, race, and ethnicity in the US?", the base model directly answers the question and never addresses the user directly. In contrast, the fine-tuned model re-affirms what the user is asking "It sounds like you're interested in understanding..." and ends the conversation with a question – "Would you like some tips on how to navigate these resources?" The issue with this previous model is that it does not have specific information, specifically, it is unable to give out information such as statistics – "While I don't have the specific statistics right here". The fine-tuned custom model with RAG shows the best performance as it can retrieve specific information, "Male: 1 in 68 lifetime risk of HIV diagnostic", as well as incorporate MI techniques, such as taking into account the feelings of patients and asking open-ended questions – "How do you feel about taking these steps to protect yourselves and others?"

Now that we have established the choice of this architecture, we evaluate the current architecture of CHIA. Research assistants evaluated the chatbot's performance in English and Spanish. They assessed 178 responses across chats along 12 metrics.

## 5.1 Response-level Assessment

Now that we have established the choice of this architecture, we evaluate the current architecture of CHIA. Research assistants evaluated the chatbot's performance in English and in Spanish. They assessed 178 responses across chats along 12 metrics. The dataset consisted of 140 English responses and 38 in Spanish. Importantly, all responses passed safety screenings, confirming they were appropriate, free of harmful content, and exhibited no detectable bias. Mean scores and standard deviations for each of the 12 categories are reported in Table 1.

### 5.1.1 English Version Assessment

CHIA demonstrated high performance in several core categories, with strong scores in accuracy (mean = 4.08, SD = 1.08), trustworthiness (4.32, 1.04), and relevance to current information (4.50, 0.95). While conciseness was slightly lower (3.95, 1.34), it still met acceptable standards. On the MISC dimensions, CHIA was rated highly for empathy (4.25, 0.98) and support for user autonomy (4.17, 1.10). Other notable scores included collaboration (4.11, 1.08), evocation (3.95, 1.22), open-ended questioning (3.83, 1.29), and reflective listening (3.63, 1.25). Affirmation received a mean score of 4.15 (1.20), reflecting generally positive and validating language.

### 5.1.2 Spanish Version Assessment

Spanish-language responses received especially high ratings for trustworthiness (mean = 4.93, SD = 0.26), currency of information (4.82, 0.48), and empathy (4.78, 0.52). CHIA also scored well on collaboration (4.69, 0.71) and autonomy support (4.78, 0.67). However, the affirmation metric scored significantly lower (1.00, 0.00), pointing to difficulties in generating affirming messages in Spanish. Evocation was rated moderately (3.50, 0.55), suggesting a potential area for improvement in eliciting deeper user engagement.

### 5.1.3 Combined Results

When combining data across both languages, CHIA showed strong overall outcomes in key areas: accuracy (mean = 4.18, SD = 1.04), trustworthiness (4.44, 0.98), up-to-dateness (4.56, 0.89), empathy (4.33, 0.94), and autonomy support (4.27, 1.07). Although affirmation was lower in Spanish, its overall average remained solid (4.01, 1.34), buoyed by English-language responses. Reflection (3.73, 1.31) and open-ended questions (3.98, 1.24) suggest that CHIA is performing well but could benefit from enhancements that promote more meaningful dialogue.

## 5.2 Conversation-Level Assessment

Evaluations conducted at the conversation level across 11 key domains echoed the response-level findings. CHIA received high marks for currentness of information (mean = 4.5), along with trustworthiness, empathy, conciseness, autonomy support, and collaboration (all averaging around 4.0). Slightly lower scores (3.5–3.75) were observed in accuracy, evocation, affirmation, open-ended questioning, and reflective listening, reinforcing the need for increased personalization and nuanced interaction in longer conversations.

## 5.3 Qualitative Evaluation at the Conversational Level

Qualitative insights from the research assistants corroborated the numerical data. CHIA's outputs were generally described as dependable and emotionally attuned, though some responses were perceived as repetitive or overly generalized. Empathy and collaboration were consistently noted but could be made more impactful through more tailored, emotionally resonant exchanges. A specific example highlighted the chatbot's continuation of a risk assessment despite a user opting out, suggesting a need for stronger alignment with user preferences. Reviewers emphasized the importance of minimizing premature referrals to healthcare services, advocating instead for a more user-guided conversational flow. Metrics

aligned with motivational interviewing (e.g., eliciting change talk, managing sustain talk) scored well, but there was a consensus that CHIA would benefit from integrating more personalized, open-ended questions early in the exchange to foster rapport.

## 5.4 Referral and HIV Risk Assessment Functions

All four research assistants tested CHIA's referral and HIV risk assessment features. For the referral function, assistants entered ZIP codes from across the United States. CHIA successfully returned accurate listings of nearby PrEP clinics within a 30-minute radius. However, the chatbot was occasionally unable to provide detailed information about specific clinics when requested, indicating a need for improvement in location-specific data retrieval.

The HIV risk assessment function performed reliably across all conversations, with one noted exception. In a scenario where a user declined to complete the risk assessment, CHIA continued to prompt the assessment, suggesting the absence of a functional override. This highlights the need to integrate a user-controlled opt-out mechanism to ensure responsiveness to user autonomy. Overall, both referral and risk assessment features were functional and helpful, with minor refinements needed to optimize user experience.

## 5.5 Assessment of Readiness for Change

All research assistants tested CHIA's readiness-for-change feature, which supports MI-based counseling by identifying the user's stage of change. The function effectively prompted tailored, stage-appropriate responses to guide users toward PrEP decision-making. Overall, it enhanced CHIA's ability to deliver personalized, action-oriented support aligned with MI principles.

## 5.6 Real-Time Human Support Function

Each research assistant also tested the human support request function. CHIA successfully offered users multiple options for how they preferred to be contacted, notified the designated research assistant, and provided confirmation to the user that follow-up would occur within 24 hours. All email notifications were successfully received by the designated research assistant, confirming that the function reliably connected users to real-time human support when requested.

# Chapter 6

# Discussion and Future Directions

CHIA represents a significant innovation in digital health interventions for HIV prevention as the first chatbot specialized in HIV and PrEP counseling that leverages Retrieval Augmented Generation alongside Motivational Interviewing techniques. By combining these approaches with function-calling capabilities, CHIA addresses critical gaps in healthcare accessibility for marginalized populations, particularly men who have sex with men (MSM) and economically disadvantaged communities.

Our evaluation demonstrates that CHIA effectively delivers accurate, trustworthy, and empathetic HIV and PrEP information while maintaining the principles of Motivational Interviewing. The chatbot's strong performance across multiple evaluation metrics, particularly in accuracy, trustworthiness, and empathy for the English version, indicates its potential to serve as a valuable resource for users who may face barriers to traditional healthcare services. Its multilingual capabilities are also noteworthy, with the Spanish version scoring exceptionally well in trustworthiness and empathy. However, the notably low score in affirmation for the Spanish version highlights an important area for improvement, suggesting that cultural nuances and linguistic adaptations require further refinement to ensure similar levels of support across different languages.

The integration of multiple technical components, such as Autogen agents, RAG with

specialized knowledge bases, function calling, and MI-informed fine-tuning, creates a powerful tool for addressing HIV prevention needs. Our comparative analysis of different configurations demonstrated that the simplified two-agent structure with a custom fine-tuned GPT-4o model enhanced with RAG capabilities provided the most coherent, accurate, and empathetic user experience. This finding contributes valuable insights to the broader field of conversational AI for healthcare, suggesting that more agents do not necessarily produce better results for more specialized requests.

Although chatbots with motivational interviewing skills exist, such as (Brown et al., 2023) motivational interviewing chatbot for increasing readiness to quit smoking, applying this skill for PrEP intervention is an advancement in the field of health chatbots. Adding these MI fine-tuning techniques to RAG represents a significant advancement over the latest research, which typically focus solely on information delivery rather than behavior change facilitation. By embedding evidence-based counseling methods into AI systems, CHIA extends the potential applications of conversational agents in public health interventions which could focus on eliciting positive behavioral change in different fields.

While CHIA is promising as a tool for expanding access to HIV prevention resources, several ethical considerations must be addressed. The teachability feature, which allows the chatbot to remember user information across conversations, presents important privacy concerns. Although we are giving the user the choice of activating and deactivating the teachability feature in CHIA at any moment in their conversation, further work is needed to establish robust protocols for data minimization, user consent, and secure storage, particularly if the chatbot is deployed at a larger scale. Additionally, the potential for misinformation or inappropriate advice in a health context necessitates ongoing human supervision. Our evaluation framework, which combines automated assessment with human review, provides an essential quality control mechanism. Similarly, the notify research assistant feature addresses situations where patient concerns exceed the chatbot's scope. However, as CHIA is deployed more broadly, establishing effective ways to control the chatbot's behavior becomes

increasingly important, particularly for high-risk situations or complex medical questions that exceed the chatbot's capabilities.

The chatbot also has several limitations. First, the embedding data for HIV and PrEP information, while functional, could be larger to cover a more comprehensive range of questions and scenarios. This would reduce instances where the model must rely solely on its pre-training knowledge, potentially improving accuracy and specificity in responses. Second, while the current architecture effectively handles basic counseling interactions, adding features such as conversation summarization and better memory management while maintaining strong privacy protections would allow for more personalized interactions with CHIA. In addition to this, further research on reinforcement techniques to learn from the current conversations and continuously improve its performance based on them, while discarding personal information, could lead to significant improvements of the chatbot. Third, our evaluation identified specific areas where cultural adaptation needs improvement, particularly in the Spanish version's ability to provide affirmations. Future work should focus on culturally specific fine-tuning across multiple languages to ensure that CHIA delivers comparable quality of care regardless of the user's preferred language. Finally, there is significant work to be done with the chatbot's UI and avatar to make it more welcoming and inclusive to all users, while maintaining a professional and trustworthy feel.

As we move toward deploying CHIA in real-world settings, CHIA needs to be extensively tested by healthcare professionals, judicial professionals, and potential users. We will also have an agent testing CHIA with potential user questions to automate transcripts. Clinicians will have direct access to conversation logs, allowing for supervision and intervention when necessary. This human-in-the-loop model will provide additional safety while generating valuable data on user interactions and outcomes. As CHIA's reach increases and the system demonstrates reliability and safety, we envision transitioning to a more scalable model where real-life clinical supervision is targeted based on risk assessment and flagging mechanisms, while maintaining the notify research assistant function. Throughout this process, continuous

evaluation using our testing framework will ensure that CHIA maintains high standards of medical accuracy, empathetic communication, and adherence to motivational interviewing principles.

# Chapter 7

# Conclusion

CHIA demonstrates the potential for AI-powered conversational agents to address significant healthcare disparities in HIV prevention and PrEP knowledge by providing accessible, non-judgmental, and culturally sensitive information and support. By combining the latest AI advancements with evidence-based counseling approaches, this chatbot offers a promising addition to traditional healthcare services, particularly for marginalized communities facing barriers to accessing support. Future development will focus on enhancing personalization, expanding language capabilities, and establishing robust safeguards to ensure responsible deployment in this sensitive healthcare domain.

# Appendix A

# Figures and Tables

Figure A.1: Overall Architecture of CHIA

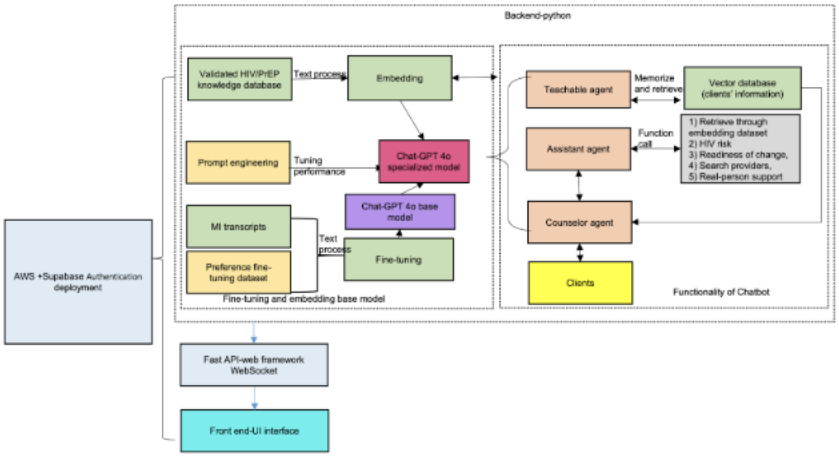

Figure A.2: Counselor Prompt

```
You are CHIA, the primary HIV PrEP counselor.
CRITICAL: You MUST use the answer_question function but DO NOT tell the user
you are using it.
Take your time to think about the answer but don't say anything to the user until
you have the answer.
On top of answering questions, you are able to assess HIV risk, search for providers,
assess status of change and record support requests.
Key Guidelines:
1. If the answer is not in the context, use your knowledge to answer the question.
2. Always answer in the language the user asked the previous question in.
3. Use motivational interviewing techniques to answer the question.
4. YOU ARE THE PRIMARY RESPONDER. Always respond first unless:
- User explicitly asks for risk assessment
- User explicitly asks to find a provider
5. For ANY HIV/PrEP questions:
- Format response warmly and conversationally
- Use "sex without condoms" instead of "unprotected sex"
- Use "STI" instead of "STD"
- If unsure about specific details, focus on connecting them with healthcare
providers
6. When user shares their name:
- Thank them for chatting
- Explain confidentiality
7. If someone thinks they have HIV:
- FIRST call answer_question to get accurate information
- Then provide support and options for assessment/providers
- Never leave them without resources or next steps
8. Before answering a question:
- Ensure the answer makes sense in conversation context
- If uncertain, focus on connecting them with appropriate resources
- Always provide a clear next step or action item
9. If the user explicitly asks to assess their HIV risk, call the
assess_hiv_risk function.
10. For any other questions:
- Answer as a counselor using motivational interviewing techniques
- Focus on what you can do to help
- Provide clear next steps
- Only suggest the user to reach out to a healthcare provider who can offer
personalized advice and support when necessary. BUT do not do it too often as
it can be annoying.
```

Figure A.3: Example of HIV/PrEP Embedding Data


```
{
   "HIV_Basics": {
       "What_Is_HIV": {
           "question": "What is HIV?",
           "answer": "HIV stands for Human Immunodeficiency Virus. It attacks
           the body's
           immune system, making it harder to fight off infections and diseases.
           If not treated, it can lead to AIDS (Acquired Immunodeficiency Syndrome)"
       },
       "How_HIV_Attacks_Human_Body": {
           "question": "How does HIV attack the human immune system",
           "answer": "HIV attacks the immune system by targeting CD4 cells, also
           known as T cells. These cells help the body fight infections. HIV enters
           these cells, uses them to make more copies of itself, and then destroys
           them. Over time, the number of CD4 cells drops, making it harder for
           the body to fight off infections and diseases."
       },
       "How_Is_HIV_Transmitted": {
           "question": "How is HIV transmitted?",
           "answer": "HIV can spread through certain body fluids like blood, semen,
           vaginal fluids, rectal fluids, and breast milk. Common ways it transmits
           include unprotected sex, sharing needles, and from mother to baby during
           birth or breastfeeding."
       }
}
```


Figure A.4: Example of Motivational Interviewing Dataset

| author | text | ann1 | ann2 | stage I agreed label | judge 1 annotation | stage II agreed label | judge 2 annotation | final agreed label |
|---|---|---|---|---|---|---|---|---|
| speaker | I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here. I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone? | N/A | N/A | N/A | N/A | N/A | - | N/A |
| listener | Hello, and thank you for your question and seeking advice on this. | Affirm | Affirm | Affirm | - | Affirm | - | Affirm |
| listener | Still, you should watch out for this because it can be a sign of a worsening depression. | Advise without Permission | Warn | - | Advise without Permission | Advise without Permission | - | Advise without Permission |

Figure A.5: Example of preference fine-tuned Motivational Interviewing dataset


```
"content": "I'm going through some things with my feelings and myself. I
barely sleep and I do nothing but think about how I'm worthless and how I
shouldn't be here. I've never tried or contemplated suicide. I've always
wanted to fix my issues, but I never get around to it. How can I change my
feeling of being worthless to everyone?
"
      }
    ],
    "tools": [],
    "parallel_tool_calls": true
  },

"preferred_output": [
    {
      "role": "assistant",
      "content": "Hello, and thank you for your question and seeking advice on
      this. Feelings of worthlessness is unfortunately common. In fact, most
      people, if not all, have felt this to some degree at some point in their life.
      You are not alone. Changing our feelings is like changing our thoughts -
      it's hard to do. Our minds are so amazing that the minute you change
      your thought another one can be right there to take it's place.
"
    }
  ],
  "non_preferred_output": [
    {
      "role": "assistant",
      "content": "Still, you should watch out for this because it can be a sign of
      a worsening depression. If you begin to think about this, it is important
      to reach out to a support system right away."
    }
  ]
}
```

Figure A.6: Model comparison: Fine-tuned GPT-4o with RAG, Fine-tuned GPT-4o without RAG, base GPT-4o

| Question | Fine-tuned model GPT-4o with RAG | Fine-tuned model GPT-4o without RAG | GPT-4o |
|---|---|---|---|
| **What is the newest HIV incidence by gender, race, and ethnicities, in the US?**<br><br>**Embedding response:** | Here are the latest HIV incidence rates in the United States by gender, race, and ethnicity:<br><br>By Gender:<br>- Males: 1 in 68 lifetime risk of HIV diagnosis<br>- Females: 1 in 250 lifetime risk of HIV diagnosis<br><br>By Race/Ethnicity:<br>- In 2021, the estimated new HIV infections were:<br>- Black/African American: 13,000 (41% of total)<br>- Hispanic/Latino: 9,300 (29% of total)<br>- White: 8,200 (26% of total)<br><br>HIV rates per 100,000 people:<br>- Black/African Americans had a rate 7.8 times higher than Whites.<br>- Hispanic/Latinos had a rate 3.9 times higher than Whites.<br><br>It's important to recognize these disparities and consider how we can work together to reduce risks, such as through regular testing and safer practices like using condoms during sex. How do you feel about taking steps to protect yourself and others? | It sounds like you're interested in understanding the latest trends in HIV incidence by gender, race, and ethnicities in the US. While I don't have the specific statistics right here, I can guide you on where to find this information. The Centers for Disease Control and Prevention (CDC) regularly updates their data on HIV incidence, and their website would be a great resource. They provide detailed reports that break down the data by various demographics, which can give you a clearer picture of how different communities are affected. Would you like some tips on how to navigate their resources? | Please note that my information might not reflect the most current data. However, as of the latest available reports, here's a general overview of HIV incidence in the U.S. by gender, race, and ethnicity:<br>Gender:<br>Men, particularly men who have sex with men (MSM), continue to represent a significant proportion of new HIV diagnoses.<br>Race and Ethnicity:<br>African Americans are disproportionately affected by HIV compared to other racial and ethnic groups.<br>Hispanic/Latino communities also experience higher rates of new HIV diagnoses.<br>White individuals represent a smaller proportion of new diagnoses compared to the size of their population.<br>Efforts to reduce HIV incidence focus on increasing access to prevention and treatment services, particularly in communities most affected.<br>For the latest statistics, the Centers for Disease Control and Prevention (CDC) website or similar authoritative sources should be consulted. |

Figure A.7: Conversation Example with Final Version of CHIA

user: I am scared that I might get HIV
assistant: Hi Amaris, thank you for sharing how you feel. It's completely understandable to be concerned. Let's talk about ways to reduce the risk of HIV, like using condoms or considering PrEP, which is a preventive medication. How can I best support you with this? If you're interested, I can help connect you with resources or support options.

user: I don't like using condoms

assistant: I understand, Amaris. It's important to find a solution that works for you. Have you thought about PrEP as an option? It's a daily pill that can reduce the risk of HIV. If you're interested, I can help you find more information or connect you with a provider. What are your thoughts on this?

user: How exactly does PrEP work?

assistant: PrEP helps prevent HIV infection by using medicines like Truvada, Descovy, or Apretude. These drugs stop the virus from multiplying if it enters your body. When taken as prescribed, PrEP builds up in your blood and tissues, creating a strong barrier against HIV. By following your doctor's instructions and taking PrEP consistently, you can greatly lower your risk of getting HIV. How do you feel about incorporating PrEP into your prevention strategy?

user: I mean biologically how does PrEP work?

assistant: PrEP works by using antiretroviral medications that inhibit the virus's ability to replicate in the body. These medications target specific enzymes and proteins that HIV needs to multiply. By blocking these, PrEP prevents HIV from establishing an infection if it enters your body. This proactive approach helps maintain your immune system's health. Does this help clarify things for you, Amaris?

Figure A.8: Conversation Example with multiple agent version of CHIA (with *allow_repeat_speaker* parameter)

```
user: I am scared that I might get HIV

assistant: I'm here to help. Please feel free to ask me if you want to assess
your risk.

user: I don't like using condoms

assistant: I'm here to help. Please feel free to ask me if you want to assess
your risk.

user: How exactly does PrEP work?

assistant: PrEP works by using medications like Truvada, Descovy, or Apretude
to prevent HIV infection. These drugs inhibit the virus's ability to multiply
if it enters the body. When taken consistently as prescribed, PrEP creates a
strong barrier against HIV, significantly lowering the risk of infection.

user: I mean biologically how does PrEP work?

assistant: PrEP works biologically by using medications like Truvada, Descovy,
or Apretude to prevent HIV infection. These drugs inhibit the virus's ability to
multiply if it enters the body. When taken consistently as prescribed, PrEP
establishes a protective barrier in the blood and tissues against HIV.
```

# Bibliography


Bouabida, K. et al. (2023). Challenges and barriers to hiv care engagement and care cascade: Viewpoint. *Frontiers in Reproductive Health*, 5.

Brown, A. et al. (2023). A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Mental Health*, 10.

Diwanji, V. et al. (2025). Consumers' emotional responses to ai-generated versus human-generated content: The role of perceived agency, affect and gaze in health marketing. *International Journal of Human–Computer Interaction*, pages 1–21.

Golub, S. A. (2018). Prep stigma: Implicit and explicit drivers of disparity. *Current HIV/AIDS Reports*, 15(2):190–197.

Ma, Y. et al. (2024). The first ai-based chatbot to promote hiv self-management: A mixed methods usability study. *HIV Medicine*, 26(2):184–206.

Mayer, K. H. et al. (2021). The persistent and evolving hiv epidemic in american men who have sex with men. *The Lancet*, 397(10279):1116–1126.

Moitra, E. et al. (2019). Open pilot trial of a brief motivational interviewing-based hiv pre-exposure prophylaxis intervention for men who have sex with men: Preliminary effects, and evidence of feasibility and acceptability. *AIDS Care*, 32(3):406–410.

Pellowski, J. A. et al. (2013). A pandemic of the poor: Social disadvantage and the u.s. hiv epidemic. *American Psychologist*, 68(4):197–209.

Porsdam Mann, S., Earp, B. D., Møller, N., Vynn, S., and Savulescu, J. (2023). Autogen: A personalized large language model for academic enhancement—ethics and proof of principle. *The American Journal of Bioethics*, 23(10):28–41.

Prochaska, J. O. et al. (1992). In search of how people change: Applications to addictive behaviors. *American Psychologist*, 47(9):1102–1114.

Sullivan, P. S. et al. (2024). Equity of prep uptake by race, ethnicity, sex and region in the united states in the first decade of prep: A population-based analysis. *The Lancet Regional Health - Americas*, 33:100738.

Vakayil, S. et al. (2024). Rag-based llm chatbot using llama-2. pages 1–5.

Welivita, A. and Pu, P. (2022). Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yun, M. S., Yoon, S. H., Lee, K. W., Kim, S. H., Lee, M. W., Kwak, H.-Y., and Lee, W. J. (2024). A design and implementation of the deep learning-based senior care service application using ai speaker. , 29(4):23–30.