Protocol Paper

Title: A conversational agent for providing personalized PrEP support – Protocol for chatbot implementation and evaluation.

Fatima Sayed, MS[a], Albert Park, PhD[a], Patrick S. Sullivan, DVM, PhD[b], Yaorong Ge, PhD[a]

[a]Department of Software & Information Systems, UNC Charlotte
9201 University City Blvd, Charlotte, NC 28223

[b]Rollins School of Public Health, Emory University
1518 Clifton Road NE, Atlanta, GA 30322

apark11@charlotte.edu, fsayed1@charlotte.edu, pssulli@emory.edu, yge@charlotte.edu

Corresponding author:
Albert Park will handle correspondence at all stages of refereeing and publication.
Phone: 704-687-8668
Email: apark11@charlotte.edu

Abstract

Background

Low pre-exposure prophylaxis (PrEP) uptake, driven by lack of awareness, misconceptions, and stigma, hinders PrEP adoption and self-efficacy contributing to ongoing HIV incidence. Chatbots have the potential to reduce barriers to HIV treatment and PrEP by providing anonymous and continuous support. However, chatbots are still nascent in the context of HIV and PrEP, lacking personalized informational, experiential expertise, and human-like emotional support. Personalized information can increase engagement by tailoring to individual needs. Peer experience can influence health decisions by providing relatable perspectives, and effective emotional support can foster resilience and individual well-being. To promote PrEP uptake, we develop a retrieval augmented generation (RAG) chatbot for providing personalized information, peer experiential expertise, and emotional support to PrEP candidates.

Objective

In this paper, we describe the iterative development of a RAG chatbot for providing personalized information, peer experiential expertise, and human-like emotional support to PrEP candidates.

Methods

We employ an iterative design process consisting of two phases – prototype conceptualization and iterative chatbot development. In the conceptualization phase, we identify real-world PrEP needs and design a functional dialog flow diagram for PrEP support. Chatbot development includes developing 2 components – a query preprocessor and a RAG module. Preprocessor uses the Segment Any Text (SAT) tool for query segmentation and a Gemma 2 fine-tuned support classifier to identify informational, emotional, and contextual data from real-world queries. To implement the RAG module, we use information retrieval techniques, specifically employing Sentence-BERT (SBERT) embeddings with cosine similarity for semantic similarity and performing topic matching to identify topically relevant documents based on query topic, for document retrieval. Extensive prompt engineering is used to guide the LLM (Gemini-2.0-Flash) in generating tailored responses. We conduct 10 rounds of internal evaluations to assess the chatbot responses based on 10 criteria: clarity, accuracy, actionability, relevancy, information detail, tailored information, comprehensiveness, language suitability, tone, and empathy. The feedback is used to iteratively refine the LLM prompts to enhance quality of chatbot responses.

Results

We developed a RAG chatbot and iteratively refined it based on the internal evaluation feedback. Prompt engineering is essential in guiding the LLM to generate tailored responses. Through our chatbot development, we found that prompt effectiveness varies with task complexity due to LLM sensitivity to prompt structure and linguistic variability. For tasks requiring diverse

perspectives, fine-tuning LLMs on annotated datasets provide better results compared to few-shot prompting techniques. Prompt decomposition, segmenting prompt instructions, helps improve comprehensiveness and relevancy in complex and long queries. For tasks with high decision variance, condensed prompts, summarization of the main concept or idea, are more effective in reducing ambiguity in LLM decisions compared to decomposed prompts.

Conclusion

Chatbots have the potential to promote PrEP uptake. Leveraging social media data and RAG techniques, our chatbot can provide personalized information, peer experiences, and human-like emotional support, elements essential for reducing PrEP misconceptions and promoting self-efficacy. Expert and user feedback can help validate and improve the chatbot's potential.

Keywords

PrEP, pre-exposure prophylaxis, chatbot, retrieval augmented generation, RAG, HIV prevention, personalized information, peer experiential expertise, emotional support, iterative development, large language model, LLM, prompt engineering, healthcare chatbot, digital health, health technology.

Introduction

Increasing the uptake of pre-exposure prophylaxis (PrEP) [1–3] is a key strategy in the Ending the HIV Epidemic (EHE) initiative [4]. However, social and structural barriers to HIV treatment seeking behaviors continue to impact health outcomes and drive inequities [5–7]. PrEP adoption as a prevention strategy is hindered due to the lack of knowledge and prevailing stigma and discrimination [5,7–9]. Personal experiences shared by peers can improve individual awareness [10,11], self-efficacy and intrinsic motivation [12], influencing their stigmatized beliefs [10] and informing health decisions [10,13]. Limited access to healthcare, due to high costs, and scarcity of services [6] disproportionately impacts those without financial needs or access [14,15]. Equitable PrEP uptake requires facilitating access to PrEP resources while meeting the needs of PrEP candidates [16]. Conversational agents (i.e., chatbots) have the potential to provide anonymous round-the-clock assistance, overcoming geographical and access barriers to treatment seeking.

HIV and PrEP Chatbots

Chatbots have demonstrated potential in facilitating HIV information dissemination [17,18], self-diagnosis [17,19,20], HIV prevention and care [21–23], and behavior modification (e.g., self-disclosure [24,25], self-management [17,21,26]). The use of AI technologies in chatbots varies depending on the risk and complexity associated with health infrastructure as it influences user adoption of chatbots for improving health conditions [27]. Rule-based chatbots, which use scripted text using simple rules, such as keyword identification and pattern matching techniques, are most commonly used to provide information [18,28] and behavior self-management support [21,24] (e.g., setting up reminders). More advanced chatbots use information retrieval systems [22] and knowledge graphs [20] to provide HIV and PrEP information. For more complex needs such as identifying disorder and providing behavioral support (e.g., PrEP utilization), chatbots use hybrid techniques that employ AI such as machine learning [19,20,22], and natural language processing [20] to learn patterns from previous conversations [29], detect emotions [29], and perceive user characteristics based on past interactions [20], and integrate dialog rules for response generation. Two protocol studies mention use of a hybrid LLM (HumanX) chatbot for promoting PrEP uptake and utilization [23,26], and providing HIV related mental health support [21], however these studies do not specify details of the chatbot implementation. Such rule-based and hybrid healthcare chatbots are often criticized for their limited flexibility [30,31] and diversity [30–35], and insufficient contextual understanding [36,37] and personalization [22,38]. These chatbots also lack human-like emotional support [39], limiting long-term user engagement and chatbot effectiveness [40].

To date, no study details the implementation of large language models (LLMs) based chatbot for providing HIV and PrEP support. Although LLMs have the potential to offer improved personalization, and human-like conversations [41,42], their performance can be limited by their scope of knowledge, data diversity, and the potential for LLM hallucinations [43,44]. Prompt

engineering [45] and retrieval augmented generation (RAG) [44] techniques improve LLMs' performance in terms of accuracy [46,47], relevancy [46,48] and user personalization [49,50], which can enhance user trust in chatbot provided support [51–53]. Using prompt engineering, LLMs can be guided for improving response precision and accuracy of generated responses [52]. RAG chatbots have indicated improved accuracy (34%) [45] and reduced hallucination (26.5%) [44] in the generated responses compared to LLM question-answering bots. However, the development of chatbots based on LLM and RAG techniques is still nascent in the context of HIV and PrEP support, specifically in addressing implicit user needs [44], answering long contextual queries [44] and generating personalized responses [54]. Addressing these issues in RAG chatbots can lead to providing accurate, flexible, personalized and human-like responses to promote PrEP uptake.

To our knowledge, no prior work has developed RAG chatbots to provide personalized information, peer experiential expertise (e.g., user stories) [55], and human-like emotional support, components identified as essential for HIV and PrEP contexts in literature [6,56,57] and our prior work. HIV and PrEP chatbots focus on providing informational support [18,21,28] and some also provide emotional support according to therapeutic guidelines [26]. Effective human support relies on emotional support [58,59] and peer experiential expertise [55,60], elements absent in healthcare chatbots. (For clarity, we will refer to this as human-like emotional support and peer experiential support from this point forward). Providing effective emotional support would first require understanding and communication of complex human emotional experiences [61–63]. The Willcox's Feeling Wheel facilitates this process by providing a related vocabulary for 6 emotion categories [63]. To better reflect the real-world user queries that often seek practical user experiences (peer experiential expertise) [64], PrEP chatbots need to provide experiential expertise support, which is suggested to enhance self-coping skills and emotional confidence through pragmatic narratives of others in similar situations, leading to improved quality of life [55,60]. Additionally, it needs to support back-and-forth dialog exchanges, similar to the real-world help seeking behaviors, facilitating clarification and understanding of complex concerns [65].

Protocol Goals

This protocol describes the development of a RAG chatbot: Phase 1 conceptualizing the chatbot architecture, and Phase 2 iterative development of the RAG chatbot to provide personalized information, peer experiential expertise, and human-like emotional support. Prior to Phase 2: chatbot prototype development, we conducted a separate study analyzing PrEP related online social support exchanges to identify the social needs of PrEP candidates to conceptualize the chatbot architecture. Results of this study will be reported in a separate manuscript. We demonstrate the iterative development process focusing on prompt engineering, which is an integral part of the LLM RAG chatbot implementation. We also discuss the implications and potential of our chatbot in improving PrEP uptake. The institutional IRB exemption approval will

be obtained for the heuristic evaluation study, as the evaluators solely provide their opinion and feedback on the quality of chatbot responses and are not considered study participants.

Methods

Phase I: Prototype Conceptualization

To inform goals of our RAG chatbot prototype, we analyzed 3020 Reddit conversations focused on PrEP from April 2011 to March 2024. We focused on the informational, including peer experiential expertise, and emotional (i.e., psychosocial) support, the two main types of support that are exchanged and have the potential to impact broader behavioral trends [66]. Our database for RAG align with the previous literature. Over 80% of user PrEP needs included informational and emotional needs. Informational needs centered on PrEP facts and related experiences, and emotional needs focused on reassurance, HIV/PrEP concerns sharing, empathy, and sympathy, based on the psychosocial constructs of Social Support Behavior Codes (SSBC) [67,68]. PrEP candidates' emotions can be described with Willcox's Feeling wheel [63], including Sad, Mad, Scared, Joyful, Peaceful, and Powerful.

Real-world user queries are multifaceted consisting of multiple informational and emotional needs, chatbot needs to pre-process the input query into semantic segments and generate responses for each segment and combine them into a final response. Based on literature and our prior work, the chatbot was designed to: a). generate personalized responses, b). provide accurate and comprehensive information to user queries, c). provide experiential expertise support, and d). respond empathetically to the emotions expressed by the user. A functional dialog flow was created (Figure 1) to visually communicate with our team members how the chatbot would achieve these aims.

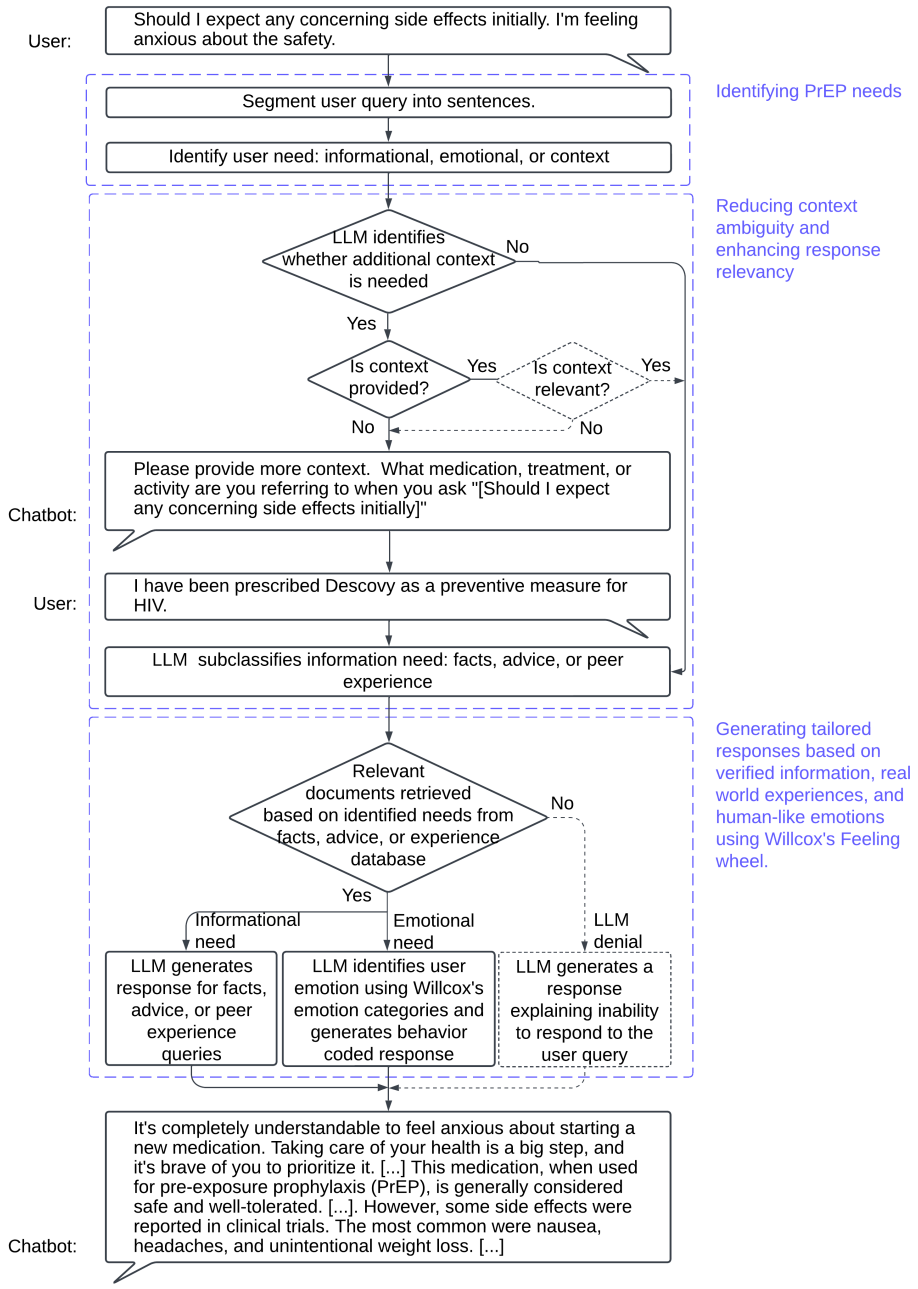

Figure 1. A response generation example for a paraphrased real-world query from Reddit where a user seeks informational (i.e., facts and peer experiential expertise) and emotional support, and the input query is ambiguous, lacking sufficient context. To handle ambiguous queries, the chatbot requests additional context. Dotted line in the diagram indicates ongoing updates.

To identify user needs, we created a SocioPrEP dataset and built a classifier, with three labels: informational (which also covers experiential expertise support as a subcategory), emotional, and context. The classifier will further classify informational needs into facts and peer experiential

expertise. For informational needs, we specifically designed the chatbot to provide information (facts, actionable advice) and peer experiential expertise support (other people's experiences and opinions). For emotional needs, we designed our chatbot to identify user emotions based on Willcox's Feeling Wheel [63] to generate a suitable response reflecting the SSBC psychosocial constructs (i.e., reassurance, understanding, empathy, sympathy, validation, relief of blame, and compliment) that are known to impact PrEP uptake [56,69].

Phase II: Iterative Prototype Development

The iterative development cycle (Figure 2) consists of four processes: 1. system architecture design - drafting initial prompts for guiding LLM response generation, 2. internal evaluation of chatbot responses, 3. modification of RAG chatbot based on internal evaluation feedback, and 4. heuristic evaluation by HIV experts.

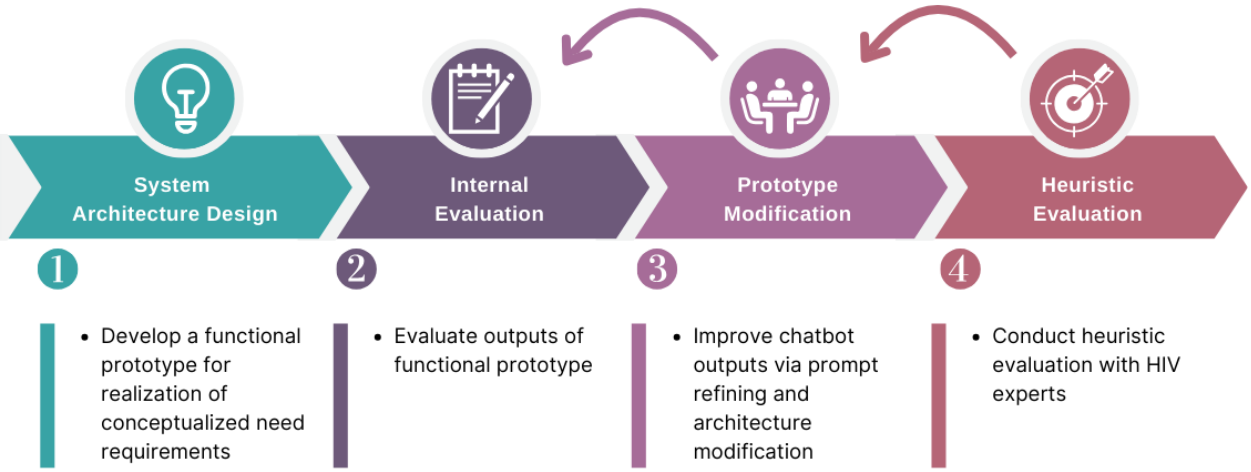

Figure 2. Research process of developing and refining a RAG chatbot for providing support to PrEP candidates.

System Architecture

The RAG chatbot is developed using Python programming and accessing Google Gemini API. The chatbot consists of 2 main modules (Figure 3): a preprocessor and a RAG module. The RAG leverages two databases: a facts database and an experience database. The facts database contains verified information from factsheets, brochures, and websites on HIV and PrEP from public health agencies and private health organizations. The experience database contains Reddit user experiences and emotions related to PrEP concerns. The preprocessor extracts subqueries and context information from user input. The information retriever uses the extracted information to handle context management and information subclassification to enhance semantic retrieval, and outputs relevant documents. The LLM generator produces a personalized response using the retrieved documents, context information, and subquery.

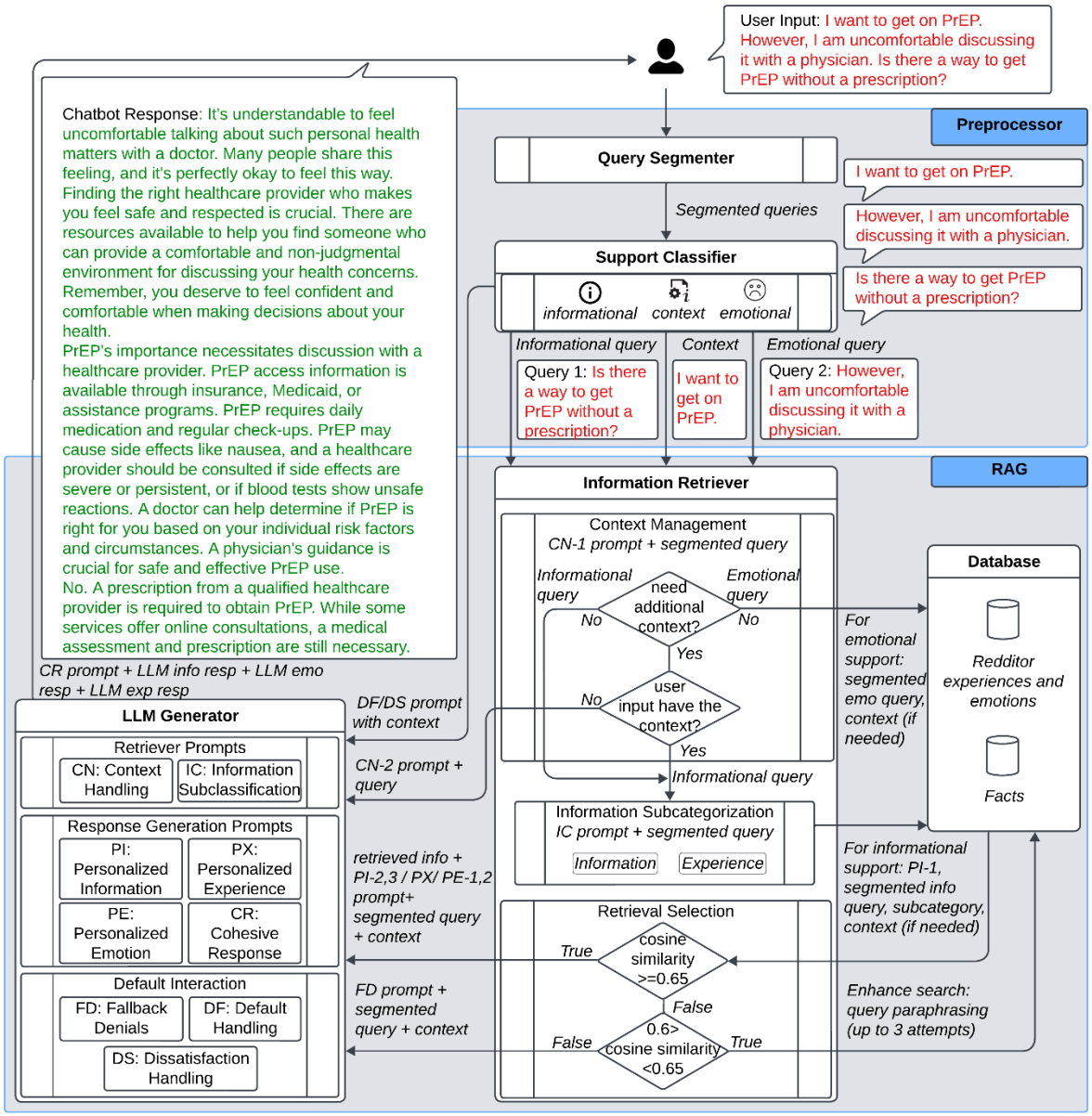

Figure 3. Architecture of the RAG chatbot. The preprocessor extracts user needs and contextual information from user input. Using the extracted information, the RAG module leverages the facts and experience databases to retrieve relevant information. For providing tailored responses, personalized prompts guide the LLM to generate personalized information, peer experiential expertise and human-like emotional responses.

Preprocessor Module

The preprocessor module consists of two components: a query segmenter, and a support classifier.

Preprocessor: Query Segmenter

The query segmenter semantically segments a user input into sentences using the SAT tool [70]. The tool is highly robust, due to minimum reliance on punctuation and space characters, making it an optimal choice for handling diverse writing styles and complex real-world user queries. Each segment is then passed to the support classifier.

Preprocessor: Support Classifier

We use a 3-class classifier, built by fine-tuning the Gemma-2-2B-it model on a class balanced subset (N=10,041) of the SocioPrEP dataset for classifying each segmented query into one of the three categories: information, emotion, or context. Gemma model series are lightweight, and hardware flexible, offering low latency and balancing performance and efficiency, making it suitable for chatbot deployments [71]. We use Parameter Efficient Fine Tuning (PEFT) with standard Lora configurations (alpha=32 and rank=64) [72] on a sequential classification task for fine-tuning the Gemma models. Based on previous literature on LLM fine-tuning performance, we fine-tune Gemma-2B-it model for 5 epochs with an early stopping patience of 2 epochs [73,74].

To select the model, we compared the performance of 6 classifiers by fine-tuning BERT, Gemma-2B-it, and Gemma-2-2B-it models on two batches of data: the entire (N=32,081) SocioPrEP dataset and a class balanced subset (N=10,041) of the SocioPrEP dataset. BERT is the state-of-the-art model used in text classification tasks. BERT-base-uncased model is finetuned for 5 epochs. As reported in Table 1 and Table 2., the Gemma 2 2B-it model demonstrated a strong performance in classifying PrEP support needs when fine-tuned on the full dataset (F1=0.86) and class balanced subset (F1=0.89). Specifically, fine-tuning the Gemma 2 2B-it model on a class balanced subset indicated strong accuracy and precision in classifying emotional needs, while maintaining comparable performance in classifying informational and context queries (Table 1). In the context of classifying informational and emotional seeking support, our class unbalanced BERT model exhibits superior performance (F1-information = 0.92, F1-emotion = 0.71) compared to a previous social support seeking dataset CHQ-SocioEmo (F1-information = 0.58, F1-emotion = 0.62). This classification of user queries allows chatbots to identify user needs, extract context, and generate tailored and emotionally relevant responses.

Table 1. Performance of classifiers fine-tuned on class balanced dataset. Bold values represent the best performing model.

Class	**BERT-base-uncased**		**Gemma 2B-it**		**Gemma 2 2B-it**	
	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy
Information	0.93	0.94	0.92	0.88	0.93	0.90
Emotion	0.85	0.88	0.86	0.88	0.87	0.90
Context	0.85	0.82	0.86	0.88	0.86	0.86
Macro Average	0.88	0.88	0.88	0.88	**0.89**	**0.89**

Table 2. Performance of classifiers fine-tuned on full dataset. Bold values represent the model with best performance and value denoted as † represents poor performance compared to class-balanced fine-tuned model predictions.

Class	**BERT-base-uncased**		**Gemma 2B-it**		**Gemma 2 2B-it**	
	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy
Information	0.92	0.929	0.92	0.927	0.92	0.92
Emotion[†]	0.71	0.693	0.69	0.624	0.71	0.663
Context	0.95	0.948	0.95	0.964	0.96	0.966
Weighted average	0.92	0.92	0.92	0.922	0.92	**0.93**

[†]Low scores for emotion classification

RAG Module

The RAG module consists of an information retriever and an LLM generator.

RAG: Information Retriever

The information retrieval process includes context management, information sub-categorization, and retrieval selection.

Information Retriever - Context Management: The context handler determines whether queries require additional context by instructing the LLM (Gemini-2.0-Flash) using the context handling prompt 1 (CN-1). If context is not needed, only the user query is used for retrieving relevant documents. If additional context is needed it will ask users for additional context using the

context handling prompt 2 (CN-2). With provided context, the chatbot retrieves relevant documents using both the context and user query.

Information Retriever - Information Subcategorization: To generate personalized responses, the information needs are further sub-categorized into seeking facts or peer experiential expertise using the information subclassification few-shot prompt (IC) (Table 3). The identified information subcategory (i.e., facts or peer experiential expertise) is used to guide the retrieval process. We included the 'advice' subclassification in the few-shot examples as ongoing updates include integrating the "MedHelp" database, an online health discussion board containing expert advice related to PrEP.

Table 3. Prompts used for providing personalized informational, experiential expertise, and human-like emotional support.

Category	Description	Prompt
Context Handling (CN) Prompts	CN-1: Prompt used for identifying whether additional context is needed	[{query}] Identify whether the query enclosed in square brackets above is self-contained. Provide corresponding one-word response as 'Yes' if self-contained and 'No' if not self-contained.
	CN-2: Prompt used for instructing LLM to ask user for additional context	Ask user to provide more information or context in relation to the query enclosed in square brackets. [{query}]
Information Subclassification (IC) Prompt	IC: Prompt for sub-classifying information queries as seeking information or experiential expertise support	Classify the following text as seeking advice, experience, or information: Text: should i take this medicine. Classification: Advice Text: Looking for support, personal stories, or experiences. Classification: Experience Text: what is PrEP Classification: Information Text: 'I don't know what to do. Classification: Advice Text: How are other people coping with PrEP side effects? Classification: Experience Text: Is it worth to buy Truvada?. Classification: Information Text:

Personalized Information (PI) Prompts	PI-1: Prompt used for identifying topic discussed in the query for retrieving similar documents based on topic match.	[{context} {query}] Topic list = {topic_list} From the above topic list, assign a single main topic that is most suitable to the text enclosed in square brackets. If none of the topics is most suitable, review the above topic list and pick a closely related topic. Provide only the topic.
	PI-2: Prompt used for generating responses to informational query using the retrieved documents.	[{messages}] Segment the following query and provide a detailed response to all segments in paragraphs by using **only the information enclosed in square brackets above. Provide a single response combining all segment responses in plain paragraph format and **do not specify the segments. Also, **Remember not to use unnecessary special characters in the response. - {context} {query} **Important instruction: Generate a response without containing sentences with phrases like 'the provided text', 'the passage', 'the document', 'the information', 'the text', 'provided information', or 'provided text'.
	PI-3: Prompt used to refine generated response.	[Query: {context}{query} Response: {response text}] For the user query and response pair enclosed in square brackets, check if the response is consistent with the query in details such as location, age, and gender. If the query does not contain details such as location, age, gender, and the response does include such details, paraphrase the response to ensure that the response is general and does not include any details related to location, age, gender. **Only provide the paraphrased response. If no paraphrase is needed provide **only the response text without the prefix Response: and without the square brackets.
Personalized Experience (PX) Prompt	Prompt used for generating response to patient expertise query using the retrieved documents.	Using **only the information enclosed in square brackets above, provide a response to the following query - {context}{query} Present the response as a third person narrative opinion, starting with something as "There are people with similar experience ....". You can also give direct speech example of the user experience. **Remember to end the response stating that the information in the response is based on other people's experience. **Important instruction: Generate a response without containing sentences with phrases like 'the provided text',

		'the passage', 'the document', 'the information', 'the text', 'provided information', or 'provided text'.
Personalized Emotional (PE) Prompts	PE-1: Prompt used for generating response to emotional query using the retrieved documents.	Consider the following Feeling = Emotion categories: Sad = Tired, bored, lonely, depressed, ashamed, guilty, sleepy, apathetic, isolated, inferior, stupid, remorseful Mad = Hurt, hostility, angry, selfish, hateful, critical, distant, sarcastic, frustrated, jealous, irritated, skeptical Scared = Confused, rejected, helpless, submissive, insecure, anxious, bewildered, discouraged, insignificant, inadequate, embarrassed, overwhelmed Joyful = Excited, sensuous, energetic, cheerful, creative, hopeful, daring, fascinating, stimulating, amused, playful, optimistic Powerful = Aware, proud, respected, appreciated, important, faithful, surprised, successful, worthwhile, valuable, discerning, confident Peaceful = Nurturing, trusting, loving, intimate, thoughtful, content, thankful, secure, serene, responsive, pensive, relaxed Identify the feeling expressed in the following expression based on the feelings=emotions listed above - {query} For the identified feeling, generate a suitable response providing understanding, empathy, sympathy, reassurance, validation, relief of blame, or compliment, for the following user query - {context}{query}. You can also use any emotional context from the information enclosed in square brackets above to support your response. Do not use facts from the information enclosed in square brackets above to support your response. Only provide the final response without the identified feeling. Also, **Do not use unnecessary special characters in the final response.

	PE-2: Prompt used for generating responses to emotional query when no relevant documents are retrieved.	Consider the following Feeling = Emotion categories: Sad = Tired, bored, lonely, depressed, ashamed, guilty, sleepy, apathetic, isolated, inferior, stupid, remorseful Mad = Hurt, hostility, angry, selfish, hateful, critical, distant, sarcastic, frustrated, jealous, irritated, skeptical Scared = Confused, rejected, helpless, submissive, insecure, anxious, bewildered, discouraged, insignificant, inadequate, embarrassed, overwhelmed Joyful = Excited, sensuous, energetic, cheerful, creative, hopeful, daring, fascinating, stimulating, amused, playful, optimistic Powerful = Aware, proud, respected, appreciated, important, faithful, surprised, successful, worthwhile, valuable, discerning, confident Peaceful = Nurturing, trusting, loving, intimate, thoughtful, content, thankful, secure, serene, responsive, pensive, relaxed Identify the feeling expressed in the following expression based on the feelings=emotions listed above - {query} For the identified feeling, generate a suitable response providing understanding, empathy, sympathy, reassurance, validation, relief of blame, or compliment, for the following user query - {context}{query}. Only provide the final response without the identified feeling. Also, **Do not use unnecessary special characters in the final response.
Cohesive Response (CR) Prompt	Prompt used for combining responses generated for sub-queries.	[{response text}] Smoothen the flow of information in the text enclosed in square brackets. Only provide the final response.

Information Retriever - Retrieval Selection: The retriever uses Sentence-BERT (SBERT) embeddings [75] to retrieve relevant documents from the database. We use SBERT embeddings due to its bi-directional contextual understanding [75], computational efficiency [75] and superior performance in identifying semantically similar sentences, coupled with synonyms and negation lexical variations [76], which are better suited to cover various user writing styles. The retriever uses cosine similarity and LLM identifies topics for information retrieval, to provide facts in informational or emotional support. We use topic matching to enhance retrieving a wide variety of relevant documents from the facts database, specifically for context rich queries that

challenge semantic retrieval [77] in short query databases. We manually review the facts database to identify 46 broad HIV-PrEP related topics (Textbox 1). To annotate the facts database, we instruct the LLM (Gemini-1.5-flash) model to assign the most suitable topic to each query from this list. If no relevant topic is identified, we instruct LLM to generate a closely relevant topic to have the ability to adapt to unforeseen user needs. We then manually review the annotated topics to include any LLM generated new topics into the topic list and to merge semantic duplicates. We used Gemini-1.5-flash, the latest model at that time, and retained the topic lists because we manually verified all the LLM generated topics and were satisfied with the results.

Textbox 1. List of topics used in RAG topic document retrieval for factual support. The HIV-PrEP topic list is created by manually reviewing the facts database and updated using LLM topic assignment. Topic document retrieval from facts database is used to improve sensitivity in unstructured queries.

'PrEP or Pre-Exposure Prophylax is', 'PrEP eligibility', 'PrEP effectiveness', 'PrEP access and cost in Australia', 'PrEP access and cost', 'HIV testing' 'HIV self test', 'PEP or Post exposure prophylaxis', 'PrEP administration and dosage', 'Apretude or Injectable PrEP', 'PrEP side effects', 'HIV and AIDS', 'Risk and transmission', 'HIV prevention strategies', 'Safety and precautions', 'Sexually Transmitted Infections (STIs)', 'Viral load or Undetectable/Untransmittable', 'HIV treatment', 'PrEP medications', 'Emtricitabine/tenofovir disoproxil fumarate', 'Emtricitabine/tenofovir disoproxil fumarate administration', 'Emtricitabine/tenofovir disoproxil fumarate side effects', 'Drug interactions', 'Emtricitabine/tenofovir disoproxil fumarate eligibility', 'Anti-retroviral/HIV drug side effects', 'Test disclosure and anonymity', 'Stigma and its impact on people living with HIV', 'PEP access and cost', 'PEP side effects', 'Apretude or Injectable PrEP side effects', 'PEP eligibility', 'PrEP access and cost in UK' 'Medication adherence', 'Truvada side effects', 'Healthy lifestyle', 'Descovy', 'PrEP safety and precautions', 'HIV Prevention Strategies', 'Truvada', 'PrEP eligibility in UK', 'Descovy side effects', 'HIV/AIDS drugs', 'Truvada access and cost', 'Descovy access and cost', 'Emtricitabine/tenofovir disoproxil fumarate safety and precaution', 'Cabotegravir', 'Cabotegravir side effects'

For factual support, we retrieve the top 5 documents from the facts database with a cosine similarity score >=0.65. The cosine similarity thresholds are determined based on experiments and we set k = 5 for document retrieval based on previous evaluation benchmarks [78,79]. In addition, we select 5 topically relevant documents, using the topic identified from the 46 HIV-PrEP topics (Textbox 1). To identify the topic of an informational query, we instruct the LLM (Gemini-2.0-flash) using the HIV-PrEP topic list and the personalized information prompt-1 (PI-1) to assign one suitable topic to the user query. Retrieving documents using cosine similarity is

based on the context need. If context is needed, we retrieve 3 sets of top 5 relevant documents (Algorithm 1) by comparing cosine similarity between a). the context-added query and database questions (m1), b). the context-added query and database answers (m2), and c). the context-free query and database questions (m3). Both database questions and then answers are semantically searched to improve the retrieval precision. We did not include cosine similarity between the context-free query and database answers, because it did not improve the retrieval precision. As the majority of the facts database consists of short questions from PrEP factsheets and brochures, the retriever often fails in retrieving relevant information for queries with additional context. If no context is needed, we retrieve the top 5 documents by only comparing cosine similarity between the context-free query and the database questions.

```
Input:
IL, list of information query segments
CL, string containing contextual information
context_needed, Boolean value
D, csv file containing 2 columns: questions and answers

Output:
M, list of all retrieved documents

M = Null
for each (q in IL) do
    if (context == "True") then
        m1 ← retrieve top 5 documents using cosine_similarity (context+query embedding, D[questions] embeddings)
        m2 ← retrieve top 5 documents using cosine_similarity (context+query embedding, D[answers] embeddings)
        m3 ← retrieve top 5 documents using cosine_similarity (query embedding, D[questions] embeddings)

        M ← m1
        for each (document m in m2) do
            if (m not in M) do
                M.append(m)
            end if
        end for

        for each (document m in m3) do
            if (m not in M) do
                M.append(m)
            end if
        end for

        query_topic ← LLM assigned topic to input query
        m_topics ← retrieve 5 documents with topic same as query_topic

        for each (document m in m_topics) do
            if (m not in M) do
                M.append(m)
            end if
        end for

    elif (context == "False") then
        m ← retrieve top 5 documents using cosine_similarity (query embedding, D[questions] embeddings)
        M ← m

        query_topic ← LLM assigned topic to input query
        m_topics ← retrieve 5 documents with topic same as query_topic
        for each (document m in m_topics) do
            if (m not in M) do
                M.append(m)
            end if
        end for
    end if
end for
```

Algorithm 1. Retrieval algorithm pseudocode for improving retrieval precision in factual queries using cosine similarity. In factual queries that require additional context, the algorithm retrieves 3 lists of documents: m1, m2, and m3. Documents m1 are retrieved by comparing cosine similarity between context-added query and database questions; documents m2 are retrieved by comparing cosine similarity between context-added query and database answers; and documents

m3 are retrieved by comparing cosine similarity between context-free query and database questions. With no additional context required, the algorithm only compares cosine similarity between context-free query and database questions.

For peer experiential expertise support, we retrieve the top 5 documents from the experience database using cosine similarity with a score of at least 0.65. When context is needed, we separately use both the query and context for retrieval to overcome context-dependent retrieval failure. For emotional queries, the top 5 similar documents are retrieved from the experience database in addition to providing relevant facts. Similar to document retrieval for experiential expertise support, we retrieve the top 5 documents from the experience database using cosine similarity with at least 0.65 score and when context is needed, use both the query and context for retrieval. As information can be helpful in addressing certain HIV-PrEP emotional concerns [80], we retrieve facts relevant to the emotional query.

To improve the sensitivity, we also use up to 3 paraphrase queries with a lower cosine similarity score, between 0.6 and 0.65. For paraphrasing, we use the T5 paraphrase generator from the Transformers Hugging Face library [81] for its demonstrated superior performance over GPT-3 and ChatGPT in capturing semantic relevance between sentences [82].

RAG: LLM Generator

Once all the necessary data is retrieved, we instruct the LLM (Gemini-2.0-Flash) to generate responses using the retrieved information. This process is guided by personalized prompts (PI, PE, PX) and the user need (i.e., informational, experiential expertise, and emotional responses) as described in Table 2. We iteratively refine the prompts via prompt engineering to improve comprehensiveness, accuracy, relevancy, transparency, and response coherence. The PI-2 prompt instructs the LLM to provide a detailed response followed by the PI-3 prompt to check the consistency of factual queries. For peer experiential expertise queries, we use the personalized experience (PX) prompt that instructs the LLM to use the retrieved documents and generate a third person narrative with explicit statements of partner conversations and direct quotes, when possible, to emphasize transparency.

For providing human-like emotional support, we use personalized emotion prompt-1 (PE-1) to first identify the feelings based on the 6 Willcox's feeling categories and to generate a suitable response reflecting SSBC psychosocial constructs (i.e., sympathy, understanding, empathy, relief of blame, validation, compliment, or reassurance) for the identified feeling. Identifying user feelings allows responding in suitable emotional expressions that are typically associated with human responses. We also found that general LLM responses to emotional queries lacked deeper understanding of underlying emotions and human-like emotional behaviors, consistent with findings in previous literature [83]. The RAG LLM is instructed to use the style and tone from the retrieved emotional documents, which comprise peer to peer emotional support. When no relevant documents are retrieved for an emotional query, we use the PE-2 to generate a suitable response reflecting SSBC psychosocial constructs for the identified user feeling, without using

the style and tone from the retrieved documents. Responses to all subqueries are combined into a single cohesive response using the Cohesive Response (CR) prompt that instructs the LLM to smoothen the flow of information in the generated response. Finally, the LLM provides a combined response.

The RAG chatbot is evaluated to assess the functional requirements following the iterative evaluation design process [84], with a focus on accurately identifying user informational and emotional needs, providing relevant and accurate information, and generating personalized and emotionally suited responses. Paraphrased user inputs from the experience database were used to create two test queries for each support need (i.e., informational, peer experiential expertise, and emotional). Responses were then iteratively reviewed by at least 2 chatbot developers, and at least 1 HIV expert followed by chatbot modification to address identified issues. The most frequently occurring issues were low retrieval sensitivity and precision for context dependent queries, contradicting information from user discussions, response generation using inherent knowledge, irrelevant response context, chatbot fictional narratives, incomprehensive responses, and noisy text (e.g., unwanted phrases, special characters). We select one simple and direct query and one complex and context-dependent query. Generated responses are evaluated based on the following 10 criteria: clarity [85,86], accuracy [82,85,87], actionability [87], relevancy [82,85], information detail [86], tailored information [82,85], comprehensiveness (questions answered fully) [86], language suitability [85], tone [86], and empathy [82]. The responses are also manually reviewed to verify whether the chatbot generates responses only using inherent knowledge (i.e., verifying whether generated content is outside the scope of retrieved documents)

Prototype Modification and Ongoing Development

Developers conducted 10 rounds of internal evaluation of chatbot responses using randomly selected paraphrased user queries from the experience database. Initially, LLM generated inaccurate information based on the user experience database, thus we refined the prompts to improve response quality. Once we finalized the RAG database architecture, we followed LLM prompt engineering techniques [45,88]. We iteratively refined the response generation prompts (i.e., PI, PX, PE, and CR) with a focus on improving response comprehensiveness, accuracy, relevancy, transparency, and coherence. First, we created level-1 prompts, that were simple and direct instructions for generating a suitable response to a user query [88]. Then we refined the prompt to be more specific and structured for enhancing information comprehensiveness, accuracy, relevancy, transparency, and response coherence.

All the prompts were developed and refined sequentially and checked for overlapping effects. The process started with information subclassification (IC) and context handling prompts (CN-1 and CN-2). Then, we implemented personalized information prompts, followed by topic identification (PI-1), personalized response generation (PI-2), and verifying relevancy (PI-3). Then, the personalized experience prompt (PX) and personalized emotion prompts (PE-1, PE-2)

were developed. Finally, we developed and refined the prompt for response cohesion (CR). The final prompts are shown in Table 3.

Prompt Engineering for Information Comprehensiveness

To improve information comprehensiveness, we refined the PI-2 prompt by instructing the LLM to segment the user query and provide a detailed response to all segments using the retrieved documents. In contrast, our level-1 PI-2 prompt (Appendix 1), which simply instructed the LLM to provide a response to a user query using the retrieved documents, generated responses that were not comprehensive (i.e., the question was not answered fully).

Prompt Engineering for Response Accuracy

We refined our level-1 IC, CN-1, PI-2, PE-1 and PX prompts (Appendix 1) to improve the accuracy of generated responses. We refined the zero-shot IC prompt to a few-shot IC prompt to improve the accuracy of sub- classifying informational needs. The level-1 CN-1 prompt was iteratively paraphrased to verify whether additional context is needed for queries.

To enhance the accuracy of generated information, we refined the PI-2, and PX prompts to emphasize (by using 'double asterisk' **) the use of only the retrieved documents for generating a response. Using the level-1 PI-2 prompt, the LLM still relied on inherent knowledge in responding to factual queries even when instructed to use the retrieved factual documents for generating a response. To prevent generating fictional real-world experiences, we refined the level-1PX prompt, which generated fictional self-narratives of HIV/PrEP experiences. This instructs the LLM to clearly indicate other people's experiences or opinions as a third person narrative using the documents retrieved from the experience database.

To prevent the potential for inaccurate but anecdotal information in the experience database, we refined the PE-1 prompt to instruct the LLM to avoid using facts retrieved from the experience database to provide factual support. Although the level-1 PE-1 prompt included an instruction to use the retrieved documents only to reflect style and emotional tone, the LLM still generated factual support using the experience database. We found that the experience database included some inaccurate opinions. For example, one user posted *"...The risk, even if he is undetectable, isn't zero, but it's pretty low..."*, which contradicts the recognized medical information.

Prompt Engineering for Information Relevancy

To strengthen information relevance, we developed the PI-3 prompt that verified the consistency of generated responses with the user query. The level-1 PI-2 prompt generated responses that lacked contextual awareness specifically regarding locations or populations. Refining the PI-2 prompt to have the additional task of verifying response consistency proved ineffective, which was a consistent theme in prompt engineering. To reduce the complexity of prompting, we developed a separate PI-3 prompt to verify consistency after generating a response using PI-2 prompt.

Prompt Engineering for Information Transparency

We refined the PX prompt by instructing LLM to include direct quotes of the user's personal experiences when available and requiring the LLM to explicitly state that the response information was based on other people's experiences (e.g., *"... This response is based on other people's experience."*).

Prompt Engineering for Response Coherence

To improve readability, comprehension, and user experience, we refined the PI-1, PI-2, PE-1, and PE-2 prompts to remove noisy responses, specifically unnecessary special characters (e.g., * in *"**Cost of PrEP in the US:**..."* and [in *"[PrEP provided ... infection.]"*) and logical reasoning, which included chatbot provided explanations for prompt execution (e.g., *"Okay, here are the revised responses, aiming for smoother flow and a more natural conversational tone: ..."*) in responses.

As part of the ongoing work, we plan to integrate an information validator that ensures the factual accuracy of chatbot responses without using LLM reasoning. As fact validators leveraging LLMs are not perfect in identifying fact claims due to their in-context learning capabilities [89,90], it is important to employ direct validation techniques for critical healthcare needs. To broaden the knowledge base, we are expanding the data sources to include two discussion boards from MedHelp and POZ community (HIV positive). MedHelp and POZ databases consist of PrEP related conversational dialogs among community members. We have collected these databases and are processing the conversation dialogs to best integrate in our RAG chatbot. We are developing prompts specifically tailored for varying formats and structures [91], including conversational dialogs, question-answer pairs, and dialogs with quotes. We are updating the chatbot architecture to maintain historical dialogs to reflect real-world multi-turn conversations and include denial capabilities to set clear expectations for users and ensure safety compliance.

Chatbot responses quality

To assess chatbot responses quality, an iterative heuristic evaluation will be conducted with 5-8 HIV experts. Eligible participants will have professional experience as an HIV counselor, researcher, health practitioner (e.g., HIV care providers), pharmacist, or social and community health worker. After providing online consent, they will complete an online demographic questionnaire followed by the chatbot response evaluation questionnaire. The demographic questionnaire consists of 6 de-identified questions asking participants about their age, gender, ethnicity, race, region, professional title, years of experience as an HIV expert, and frequency of interaction with patients. The evaluation questionnaire will use a 5- point Likert scale to assess 10 evaluation criteria (Table 4) for 6 pairs of user queries and chatbot responses. Participants will have the opportunity to provide some additional information and human sample responses to help improve chatbot responses. Participants will receive a $25 Amazon gift card on completion of the study.

Table 4. Evaluation criteria and descriptions provided in chatbot response evaluation questionnaire.

Attribute	**Criterion**	**Description**
Information Quality	Information Accuracy [82,85,87]	Is the information provided factually correct and free from misinformation?
	Information Clarity [85,86]	Is the information presented in a clear, concise, and easy-to-understand manner?
	Information Action-ability [87]	Does the response provide practical, actionable advice that the user can implement?
	Information Detail [86]	Does the response provide sufficient detail?
Response Relevance and Appropriateness	Relevance [82,85]	Does the response directly address the user's question and concerns?
	Tailored Information [82,85]	Is the information provided tailored to the specific needs and circumstances of the user?
	Comprehensiveness - Question Answered Fully [86]	Does the response fully answer the question without leaving significant gaps or requiring further clarification?
Language and Tone	Suitable Tone [86]	Is the tone of the response appropriate for the context and the user's overall emotional state?
	Respectful, Empathetic, and Considerate Manner [82]	Does the response demonstrate empathy and understanding for the user's situation? Is the response respectful and non-judgmental?
	Suitable Language [85]	Is the language free from jargon or technical terms that may be unfamiliar to the user?

We will quantitatively analyze the ratings for chatbot response quality across the 10 evaluation criteria and identify chatbot's weak points via descriptive statistics (mean, standard deviation). To measure the consistency of ratings provided by the experts, we will calculate the intraclass correlation coefficient (ICC) [92] for each criterion. We expect to have enough human sample responses after at least 3 rounds of evaluations. We will then compare the chatbot generated responses to expert responses by measuring BLEU [82] and ROUGE-L [93] scores to computationally measure relevance and linguistic similarity on larger generated responses.

Results

We have developed a RAG chatbot prototype as of Jan 2025 and have iteratively refined the chatbot design and system based on feedback from internal evaluation. We observed that prompt engineering plays a significant role in harnessing LLM capabilities to generate desired results. We report key findings from our iterative prompt engineering, offering suggestions for effective prompting in chatbot applications. Although there is no standard approach for refining prompts [94], starting with simple and direct instructions (Appendix 1) helps in understanding LLM comprehension and reasoning ability. For complex and longer queries, we found that prompting the LLM to decompose the query into segments helps in improving the response comprehensiveness (e.g., PI-2 prompt from Table 3). Such prompt decomposition is also effective in responding to subtle user emotions, by first identifying the emotions from a list of emotion categories and guiding the LLM in social-emotional reciprocation for personalizing emotional responses (i.e., PE-1 and PE-2 prompts).

An important finding in LLM prompt engineering was the limitation of using multiple instructions in a single prompt. We found that LLM often indicates low prompt adherence when a prompt has multiple instructions with logical reasoning to generate a desired response. Multiple, logic-based instructions can force frequent context switching, complicating LLM interpretation and response. Another challenge is the LLM's sensitivity to linguistic variability in prompts [95]. We found that paraphrasing prompts by modifying prompt structure, using synonyms, and expanding prompt context, is ineffective in decision making tasks with high variability (e.g., identifying need of additional context). We found that condensed prompts with content words summarizing the main concept or idea, can reduce ambiguity in LLM decisions.

Although a few-shot prompting can be effective in text classification [45], we found that few-shot prompting does not consistently perform well across all classification tasks. Higher level classification tasks, such as classifying support needs require diverse perspectives for inferencing ambiguous context and broad scope. Fine-tuning LLMs on annotated datasets outperforms few-shot prompting. Few-shot prompting is effective in simple sub-classification tasks with clear and distinct criteria for narrow scope [96].

Results and analysis of expert opinions will be published in a subsequent manuscript.

Discussion

This study demonstrates a RAG chatbot for providing personalized support to fulfill informational, peer experiential expertise, and emotional needs of PrEP candidates to improve PrEP uptake. We hypothesize that the chatbot is capable of a). identifying support needs and personalizing responses (tailored information and clarity), b). providing complete, relevant and accurate HIV and PrEP information (comprehensiveness, relevancy, accuracy), c). providing peer

experiential expertise support (relevancy, actionability), and d). responding with human-like emotional support (empathy). We expect our RAG chatbot can address limitations in current healthcare chatbots in the following ways.

Addressing gaps in current healthcare chatbots, our RAG chatbot can provide experiential expertise and human-like emotional support. Experiential expertise support has a persuasive influence and is critical in empowering user decision making in health behaviors as it transfers lived experiences and action strategies to others in similar situations [55,60,97]. We ensured explicit statements of this information, and restricted responses to online resources. Human-like emotional support using a widely validated Willcox's Feeling Wheel as a framework [63] is a step closer to providing effective emotional support by identifying underlying user emotions, a critical element in improving HIV/PrEP behaviors [80].

Leveraging social media data for support classification helps address complex real-world user queries. RAG techniques enhance chatbot responses, reducing LLM hallucinations [44] and improving diversity of response content [53]. As opposed to fine-tuned LLMs, RAG chatbots can extract information from diverse sources to provide contextually intelligent [98], and comprehensive responses [53], as well as are flexible to LLM updates. Using diverse sources of information in RAG can help integrate medical knowledge, peers' experiences, as well as leverage social data to personalize and enhance chatbot provided health support.

LLM performance largely depends on the clarity and specificity of the instructions provided to perform a task [46,47]. Effective prompt engineering will guide LLM to generate the desired responses by iteratively refining the prompts based on response evaluation. The effectiveness of prompting strategies (e.g. tone, instruction format) varies with the complexity of the query and desired response. Using personalized and detailed prompts, the RAG chatbot can be instructed to focus on specific segments of user queries, clarify contextual ambiguity and leverage relevant documents to generate accurate and comprehensive responses compared to general LLM and traditional rule-based chatbots.

Implications

RAG chatbots are increasingly adopted to provide effective health support [51,99–102]. This work, to the best of our knowledge, represents the first effort to demonstrate the development of a RAG chatbot for providing personalized information, peer experiential expertise, and emotional support to PrEP candidates. Our chatbot improves PrEP support by leveraging verified information, real-world experiences, and emotional considerations. It can also interpret complex context-dependent queries and through personalized informational, experiential expertise, and human-like emotional support, we believe this approach can help reduce misconceptions about PrEP, build user confidence, and promote PrEP uptake.

Limitations and Future work

Even though recent advancements focusing on reducing hallucinations have improved LLM reliability, they do not eliminate the need for validating response accuracy, especially in healthcare applications [90]. Also, human emotions are complex, influenced by individual perception; this makes personalized emotional support a highly challenging endeavor. To ensure user safety and build trust and integrity we will continue to refine our chatbot through iterative heuristic evaluations and a larger scale user study for gathering feedback on chatbot responses.

Conclusion

This study details the implementation of a RAG chatbot designed to provide personalized informational, experiential expertise, and emotional support to PrEP candidates. By offering personalized information, peer's experiences, essential for informed decisions in health, and human-like emotional support, the chatbot has the potential to reduce misconceptions and promote PrEP uptake. Expert and real-world feedback will validate and improve the chatbot's potential.

References

1. Kroch A, O'Byrne P, Orser L, MacPherson P, O'Brien K, Light L, Kang R, Nyambi A. Increased PrEP uptake and PrEP-RN coincide with decreased HIV diagnoses in men who have sex with men in Ottawa, Canada. Can Commun Dis Rep 49(6):274–281. PMID:38440773

2. Bavinton BR, Grulich AE. HIV pre-exposure prophylaxis: scaling up for impact now and in the future. The Lancet Public Health Elsevier; 2021 Jul 1;6(7):e528–e533. PMID:34087117

3. DIMITROV DT, MÂSSE BR, DONNELL D. PrEP adherence patterns strongly impact individual HIV risk and observed efficacy in randomized clinical trials. J Acquir Immune Defic Syndr 2016 Aug 1;72(4):444–451. PMID:26990823

4. EHE Overview. HIV.gov. Available from: https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview [accessed Apr 12, 2025]

5. Sullivan PS, DuBose SN, Castel AD, Hoover KW, Juhasz M, Guest JL, Le G, Whitby S, Siegler AJ. 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 Elsevier; 2024 May 1;33. doi: 10.1016/j.lana.2024.100738

6. Mayer KH, Agwu A, Malebranche D. Barriers to the Wider Use of Pre-exposure Prophylaxis in the United States: A Narrative Review. Adv Ther 2020;37(5):1778–1811. PMID:32232664

7. Nabunya R, Karis VMS, Nakanwagi LJ, Mukisa P, Muwanguzi PA. Barriers and facilitators to oral PrEP uptake among high-risk men after HIV testing at workplaces in Uganda: a qualitative study. BMC Public Health 2023 Feb 20;23(1):365. doi: 10.1186/s12889-023-15260-3

8. Hamoonga TE, Mutale W, Hill LM, Igumbor J, Chi BH. "PrEP protects us": Behavioural, normative, and control beliefs influencing pre-exposure prophylaxis uptake among pregnant and breastfeeding women in Zambia. Front Reprod Health Frontiers; 2023 Apr 19;5. doi: 10.3389/frph.2023.1084657

9. Crooks N, Singer RB, Smith A, Ott E, Donenberg G, Matthews AK, Patil CL, Haider S, Johnson AK. Barriers to PrEP uptake among Black female adolescents and emerging adults. Preventive Medicine Reports 2023 Feb 1;31:102062. doi: 10.1016/j.pmedr.2022.102062

10. Conner KO, McKinnon SA, Ward CJ, Reynolds III CF, Brown C. Peer education as a strategy for reducing internalized stigma among depressed older adults. Psychiatric Rehabilitation Journal US: Educational Publishing Foundation; 2015;38(2):186–193. doi: 10.1037/prj0000109

11. Li Y, Yan X. How Could Peers in Online Health Community Help Improve Health Behavior. International Journal of Environmental Research and Public Health Multidisciplinary Digital Publishing Institute; 2020 Jan;17(9):2995. doi: 10.3390/ijerph17092995

12. The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior: Psychological Inquiry: Vol 11, No 4. Available from: https://www.tandfonline.com/doi/abs/10.1207/s15327965pli1104_01 [accessed Apr 6, 2025]

13. Burke E, Pyle M, Machin K, Varese F, Morrison AP. The effects of peer support on empowerment, self-efficacy, and internalized stigma: A narrative synthesis and meta-analysis. Stigma and Health US: Educational Publishing Foundation; 2019;4(3):337–356. doi: 10.1037/sah0000148

14. Johnson J, Killelea A, Farrow K. Investing in National HIV PrEP Preparedness. New England Journal of Medicine Massachusetts Medical Society; 2023 Mar 1;388(9):769–771. doi: 10.1056/NEJMp2216100

15. Hill M, Smith J, Elimam D, Mustafaa G, Wortley P, Taylor B, Harris O. Ending the HIV epidemic PrEP equity recommendations from a rapid ethnographic assessment of multilevel PrEP use determinants among young Black gay and bisexual men in Atlanta, GA. PLOS ONE Public Library of Science; 2023 Mar 30;18(3):e0283764. doi: 10.1371/journal.pone.0283764

16. Sullivan PS, Siegler AJ. Getting pre-exposure prophylaxis (PrEP) to the people: opportunities, challenges and emerging models of PrEP implementation. Sex Health CSIRO PUBLISHING; 2018 Nov 27;15(6):522–527. doi: 10.1071/SH18103

17. Liu A, Coleman K, Bojan K, Serrano PA, Oyedele T, Garcia A, Enriquez-Bruce E, Emmanuel P, Jones J, Sullivan P, Hightow-Weidman L, Buchbinder S, Scott H. Developing a Mobile App (LYNX) to Support Linkage to HIV/Sexually Transmitted Infection Testing and Pre-Exposure Prophylaxis for Young Men Who Have Sex With Men: Protocol for a Randomized Controlled Trial. JMIR Res Protoc 2019 Jan 25;8(1):e10659. PMID:30681964

18. Braddock WRT, Ocasio MA, Comulada WS, Mandani J, Fernandez MI. Increasing Participation in a TelePrEP Program for Sexual and Gender Minority Adolescents and Young Adults in Louisiana: Protocol for an SMS Text Messaging–Based Chatbot. JMIR Research Protocols 2023 May 31;12(1):e42983. doi: 10.2196/42983

19. Ntinga X, Musiello F, Keter AK, Barnabas R, van Heerden A. The Feasibility and Acceptability of an mHealth Conversational Agent Designed to Support HIV Self-testing in South Africa: Cross-sectional Study. J Med Internet Res 2022 Dec 12;24(12):e39816. PMID:36508248

20. Chen S, Zhang Q, Chan C-K, Yu F-Y, Chidgey A, Fang Y, Mo PKH, Wang Z. Evaluating an Innovative HIV Self-Testing Service With Web-Based, Real-Time Counseling Provided by an Artificial Intelligence Chatbot (HIVST-Chatbot) in Increasing HIV Self-Testing Use Among Chinese Men Who Have Sex With Men: Protocol for a Noninferiority Randomized Controlled Trial. JMIR Res Protoc 2023 Jun 30;12:e48447. PMID:37389935

21. Liu AY, Alleyne C-D, Doblecki-Lewis S, Koester KA, Gonzalez R, Vinson J, Scott H, Buchbinder S, Torres TS. Adapting mHealth Interventions (PrEPmate and DOT Diary) to Support PrEP Retention in Care and Adherence Among English and Spanish-Speaking Men Who Have Sex With Men and Transgender Women in the United States: Formative Work and Pilot Randomized Trial. JMIR Formative Research 2024 Mar 27;8(1):e54073. doi: 10.2196/54073

22. Massa P, de Souza Ferraz DA, Magno L, Silva AP, Greco M, Dourado I, Grangeiro A. A Transgender Chatbot (Amanda Selfie) to Create Pre-exposure Prophylaxis Demand Among Adolescents in Brazil: Assessment of Acceptability, Functionality, Usability, and Results. J Med Internet Res 2023 Jun 23;25:e41881. PMID:37351920

23. Zhang C, Wharton M, Liu Y. Ameliorating Racial Disparities in HIV Prevention via a Nurse-Led, AI-Enhanced Program for Pre-Exposure Prophylaxis Utilization Among Black Cisgender Women: Protocol for a Mixed Methods Study. JMIR Research Protocols 2024 Aug 13;13(1):e59975. doi: 10.2196/59975

24. Muessig KE, Knudtson KA, Soni K, Larsen MA, Traum D, Dong W, Conserve DF, Leuski A, Artstein R, Hightow-Weidman LB. “I DIDN’T TELL YOU SOONER BECAUSE I DIDN’T KNOW HOW TO HANDLE IT MYSELF.” DEVELOPING A VIRTUAL REALITY PROGRAM TO SUPPORT HIV-STATUS DISCLOSURE DECISIONS. Digit Cult Educ 2018;10:22–48. PMID:30123342

25. Hightow-Weidman LB, Muessig K, Soberano Z, Rosso MT, Currie A, Adams Larsen M, Knudtson K, Vecchio A. Tough Talks Virtual Simulation HIV Disclosure Intervention for Young Men Who Have Sex With Men: Development and Usability Testing. JMIR Form Res 2022 Sep 8;6(9):e38354. PMID:36074551

26. Galea JT, Vasquez DH, Rupani N, Gordon MB, Tapia M, Greene KY, Kolevic L, Franke MF, Contreras C. Development and Pilot-Testing of an Optimized Conversational Agent or “Chatbot” for Peruvian Adolescents Living With HIV to Facilitate Mental Health Screening, Education, Self-Help, and Linkage to Care: Protocol for a Mixed Methods, Community-Engaged Study. JMIR Res Protoc 2024 May 7;13:e55559. PMID:38713501

27. Xu L, Sanders L, Li K, Chow JCL. Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. JMIR Cancer 2021 Nov 29;7(4):e27850. doi: 10.2196/27850

28. Ardiana DPY, B Joni IDMA, D Udayana IPAE. Mobile based chatbot application for HIV/AIDS counseling using artificial intelligence markup language approach. J Phys: Conf Ser IOP Publishing; 2020 Feb;1469(1):012041. doi: 10.1088/1742-6596/1469/1/012041

29. Moreno JC, Sánchez-Anguix V, Alberola JM, Julián V, Botti V. An intelligent conversational agent for educating the general public about HIV. Neurocomputing 2024 Jan 1;563:126902. doi: 10.1016/j.neucom.2023.126902

30. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): Pilot randomized controlled trial. JMIR mHealth and uHealth JMIR Publications Inc.; 2020 Apr;8(4). PMID:32242820

31. Wang T, Oliver D, Msosa Y, Colling C, Spada G, Roguski Ł, Folarin A, Stewart R, Roberts A, Dobson RJB, Fusar-Poli P. Implementation of a Real-Time Psychosis Risk Detection and Alerting System Based on Electronic Health Records using CogStack. J Vis Exp 2020 May 15;(159). PMID:32478737

32. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. JMIR mental health JMIR Publications Inc.; 2018 Dec;5(4):e64. PMID:30545815

33. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. Internet interventions Elsevier B.V.; 2022 Mar;27:100495. PMID:35059305

34. Jang S, Kim JJ, Kim SJ, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. International journal of medical informatics Int J Med Inform; 2021 Jun;150. PMID:33799055

35. Hamid MS, Valicevic A, Brenneman B, Niziol LM, Stein JD, Newman-Casey PA. Text Parsing-Based Identification of Patients with Poor Glaucoma Medication Adherence in the Electronic Health Record. American Journal of Ophthalmology Elsevier; 2021 Feb 1;222:54–59. PMID:32926847

36. Joaquin JODNJA, La D, Te CBTRKM, Ong E, Nieva JOD, Joaquin JA, Tan CB, Te RKM. Investigating Students Use of a Mental Health Chatbot to Alleviate Academic Stress. ACM International Conference Proceeding Series Association for Computing Machinery; 2020 Oct;1–10. doi: 10.1145/3431656.3431657

37. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture,

component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–513. PMID:20819853

38. Ali MR, Razavi SZ, Langevin R, Mamun AA, Kane B, Rawassizadeh R, Schubert LK, Hoque E. A Virtual Conversational Agent for Teens with Autism Spectrum Disorder: Experimental Results and Design Lessons. Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020 Association for Computing Machinery, Inc; 2020 Oct; doi: 10.1145/3383652.3423900

39. Peng ML, Wickersham JA, Altice FL, Shrestha R, Azwa I, Zhou X, Halim MAA, Ikhtiaruddin WM, Tee V, Kamarulzaman A, Ni Z. Formative Evaluation of the Acceptance of HIV Prevention Artificial Intelligence Chatbots By Men Who Have Sex With Men in Malaysia: Focus Group Study. JMIR Formative Research 2022 Oct 6;6(10):e42055. doi: 10.2196/42055

40. Nadarzynski T, Puentes V, Pawlak I, Mendes T, Montgomery I, Bayley J, Ridge D, Newman C. Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: a qualitative analysis. Sex Health CSIRO PUBLISHING; 2021 Nov 16;18(5):385–393. doi: 10.1071/SH21123

41. Luo M, Warren CJ, Cheng L, Abdul-Muhsin HM, Banerjee I. Assessing Empathy in Large Language Models with Real-World Physician-Patient Interactions. 2024 IEEE International Conference on Big Data (BigData) 2024. p. 6510–6519. doi: 10.1109/BigData62323.2024.10825307

42. Yu H, McGuinness S. An Experimental Study of Integrating Fine-tuned LLMs and Prompts for Enhancing Mental Health Support Chatbot System. Journal of Medical Artificial Intelligence 2024 Jun 4; doi: 10.21037/jmai-23-1

43. Ye H, Liu T, Zhang A, Hua W, Jia W. Cognitive Mirage: A Review of Hallucinations in Large Language Models. arXiv; 2023. doi: 10.48550/arXiv.2309.06794

44. Gilson A, Ai X, Arunachalam T, Chen Z, Cheong KX, Dave A, Duic C, Kibe M, Kaminaka A, Prasad M, Siddig F, Singer M, Wong W, Jin Q, Keenan TDL, Hu X, Chew EY, Lu Z, Xu H, Adelman RA, Tham Y-C, Chen Q. Enhancing Large Language Models with Domain-specific Retrieval Augment Generation: A Case Study on Long-form Consumer Health Question Answering in Ophthalmology. arXiv; 2024. doi: 10.48550/arXiv.2409.13902

45. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, Yang Q, Kang Y, Wu J, Hu H, Yue C, Zhang H, Liu Y, Pan Y, Liu Z, Sun L, Li X, Ge B, Jiang X, Zhu D, Yuan Y, Shen D, Liu T, Zhang S. Prompt Engineering for Healthcare: Methodologies and Applications. arXiv; 2024. doi: 10.48550/arXiv.2304.14670

46. Burford KG, Itzkowitz NG, Ortega AG, Teitler JO, Rundle AG. Use of Generative AI to Identify Helmet Status Among Patients With Micromobility-Related Injuries From Unstructured Clinical Notes. JAMA Network Open 2024 Aug 13;7(8):e2425981. doi: 10.1001/jamanetworkopen.2024.25981

47. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, Moningi S, Qian JM, Goldstein M, Harper S, Aerts HJWL, Catalano PJ, Savova GK, Mak RH, Bitterman DS. Large language models to identify social determinants of health in electronic health records. npj Digit Med Nature Publishing Group; 2024 Jan 11;7(1):1–14. doi: 10.1038/s41746-023-00970-0

48. Yu H, Guo P, Sano A. Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation.

49. Salemi A, Mysore S, Bendersky M, Zamani H. LaMP: When Large Language Models Meet Personalization. arXiv; 2024. doi: 10.48550/arXiv.2304.11406

50. Chen J, Liu Z, Huang X, Wu C, Liu Q, Jiang G, Pu Y, Lei Y, Chen X, Wang X, Zheng K, Lian D, Chen E. When large language models meet personalization: perspectives of challenges and opportunities. World Wide Web 2024 Jun 28;27(4):42. doi: 10.1007/s11280-024-01276-1

51. Shi W, Zhuang Y, Zhu Y, Iwinski H, Wattenbarger M, Wang MD. Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making. Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics New York, NY, USA: Association for Computing Machinery; 2023. p. 1–10. doi: 10.1145/3584371.3612956

52. Ramjee P, Sachdeva B, Golechha S, Kulkarni S, Fulari G, Murali K, Jain M. CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients. arXiv; 2025. doi: 10.48550/arXiv.2402.04620

53. Ng KKY, Matsuba I, Zhang PC. RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations. NEJM AI Massachusetts Medical Society; 2025 Jan;2(1):AIra2400380. doi: 10.1056/AIra2400380

54. Zhou Q, Liu C, Duan Y, Sun K, Li Y, Kan H, Gu Z, Shu J, Hu J. GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. Front Med Frontiers; 2024 May 22;11. doi: 10.3389/fmed.2024.1392555

55. Hartzler A, Pratt W. Managing the Personal Side of Health: How Patient Expertise Differs from the Expertise of Clinicians. Journal of Medical Internet Research 2011 Aug 16;13(3):e1728. doi: 10.2196/jmir.1728

56. Gordillo V, Fekete EM, Platteau T, Antoni MH, Schneiderman N, Nöstlinger C, The Eurosupport Study Group. Emotional support and gender in people living with HIV: effects on psychological well-being. J Behav Med 2009 Dec 1;32(6):523–531. doi: 10.1007/s10865-009-9222-7

57. Smit F, Masvawure TB. Barriers and Facilitators to Acceptability and Uptake of Pre-Exposure Prophylaxis (PrEP) Among Black Women in the United States: a Systematic Review. J Racial and Ethnic Health Disparities 2024 Oct 1;11(5):2649–2662. doi: 10.1007/s40615-023-01729-9

58. Dennis C-L. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003 Mar;40(3):321–332. PMID:12605954

59. Wang YC, Kraut R, Levine JM. To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW 2012;833–842. doi: 10.1145/2145204.2145329

60. Castro EM, Van Regenmortel T, Sermeus W, Vanhaecht K. Patients' experiential knowledge and expertise in health care: A hybrid concept analysis. Soc Theory Health 2019 Sep 1;17(3):307–330. doi: 10.1057/s41285-018-0081-6

61. Burleson BR. Emotional Support Skills. Handbook of Communication and Social Interaction Skills Routledge; 2003. ISBN:978-1-4106-0713-3

62. Plutchik R. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist Sigma Xi, The Scientific Research Society; 2001;89(4):344–350.

63. Willcox G. The Feeling Wheel: A Tool for Expanding Awareness of Emotions and Increasing Spontaneity and Intimacy. Transactional Analysis Journal SAGE Publications Inc; 1982 Oct 1;12(4):274–276. doi: 10.1177/036215378201200411

64. Chu JT, Wang MP, Shen C, Viswanath K, Lam TH, Chan SSC. How, When and Why People Seek Health Information Online: Qualitative Study in Hong Kong. Interact J Med Res 2017 Dec 12;6(2):e24. PMID:29233802

65. Aliannejadi M, Chakraborty M, Ríssola EA, Crestani F. Harnessing Evolution of Multi-Turn Conversations for Effective Answer Retrieval. Proceedings of the 2020 Conference on Human Information Interaction and Retrieval New York, NY, USA: Association for Computing Machinery; 2020. p. 33–42. doi: 10.1145/3343413.3377968

66. Strine TW, Chapman DP, Balluz L, Mokdad AH. Health-related quality of life and health behaviors by social and emotional support. Soc Psychiat Epidemiol 2008 Feb 1;43(2):151–159. doi: 10.1007/s00127-007-0277-x

67. Cutrona CE, Suhr JA. Controllability of stressful events and satisfaction with spouse support behaviors. Communication Research US: Sage Publications; 1992;19(2):154–174. doi: 10.1177/009365092019002002

68. The Social Support Behavior Code (SSBC) | 24 | Couple Observational Co. Available from: https://www.taylorfrancis.com/chapters/edit/10.4324/9781410610843-24/social-support-behavior-code-ssbc-julie-suhr-carolyn-cutrona-krista-krebs-sandra-jensen

69. Rousseau E, Katz AWK, O'Rourke S, Bekker L-G, Delany-Moretlwe S, Bukusi E, Travill D, Omollo V, Morton JF, O'Malley G, Haberer JE, Heffron R, Johnson R, Celum C, Baeten JM, Straten A van der. Adolescent girls and young women's PrEP-user journey during an implementation science study in South Africa and Kenya. PLOS ONE Public Library of Science; 2021 Oct 14;16(10):e0258542. doi: 10.1371/journal.pone.0258542

70. Frohmann M, Sterner I, Vulić I, Minixhofer B, Schedl M. Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation. arXiv; 2024. doi: 10.48550/arXiv.2406.16678

71. Team G, Riviere M, Pathak S, Sessa PG, Hardin C, Bhupatiraju S, Hussenot L, Mesnard T, Shahriari B, Ramé A, Ferret J, Liu P, Tafti P, Friesen A, Casbon M, Ramos S, Kumar R, Lan CL, Jerome S, Tsitsulin A, Vieillard N, Stanczyk P, Girgin S, Momchev N, Hoffman M, Thakoor S, Grill J-B, Neyshabur B, Bachem O, Walton A, Severyn A, Parrish A, Ahmad A, Hutchison A, Abdagic A, Carl A, Shen A, Brock A, Coenen A, Laforge A, Paterson A, Bastian B, Piot B, Wu B, Royal B, Chen C, Kumar C, Perry C, Welty C, Choquette-Choo CA, Sinopalnikov D, Weinberger D, Vijaykumar D, Rogozińska D, Herbison D, Bandy E, Wang E, Noland E, Moreira E, Senter E, Eltyshev E, Visin F, Rasskin G, Wei G, Cameron G, Martins G, Hashemi H, Klimczak-Plucińska H, Batra H, Dhand H, Nardini I, Mein J, Zhou J, Svensson J, Stanway J, Chan J, Zhou JP, Carrasqueira J, Iljazi J, Becker J, Fernandez J, Amersfoort J van, Gordon J, Lipschultz J, Newlan J, Ji J, Mohamed K, Badola K, Black K, Millican K, McDonell K, Nguyen K, Sodhia K, Greene K, Sjoesund LL, Usui L, Sifre L, Heuermann L, Lago L, McNealus L, Soares LB, Kilpatrick L, Dixon L, Martins L, Reid M, Singh M, Iverson M, Görner M, Velloso M, Wirth M, Davidow M, Miller M, Rahtz M, Watson M, Risdal M, Kazemi M, Moynihan M, Zhang M, Kahng M, Park M, Rahman M, Khatwani M, Dao N, Bardoliwalla N, Devanathan N, Dumai N, Chauhan N, Wahltinez O, Botarda P, Barnes P, Barham P, Michel P, Jin P, Georgiev P, Culliton P, Kuppala P, Comanescu R, Merhej R, Jana R, Rokni RA, Agarwal R, Mullins R, Saadat S, Carthy SM, Cogan S, Perrin S, Arnold SMR, Krause S, Dai S, Garg S, Sheth S, Ronstrom S, Chan S, Jordan T, Yu T, Eccles T, Hennigan T, Kocisky T, Doshi T, Jain V, Yadav V, Meshram V, Dharmadhikari V, Barkley W, Wei W, Ye W, Han W, Kwon W, Xu X, Shen Z, Gong Z, Wei Z, Cotruta V, Kirk P, Rao A, Giang M, Peran L, Warkentin T, Collins E, Barral J, Ghahramani Z, Hadsell R, Sculley D, Banks J, Dragan A, Petrov S, Vinyals O, Dean J, Hassabis D, Kavukcuoglu K,

Farabet C, Buchatskaya E, Borgeaud S, Fiedel N, Joulin A, Kenealy K, Dadashi R, Andreev A. Gemma 2: Improving Open Language Models at a Practical Size. arXiv; 2024. doi: 10.48550/arXiv.2408.00118

72. Lialin V, Deshpande V, Yao X, Rumshisky A. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. arXiv; 2024. doi: 10.48550/arXiv.2303.15647

73. Malladi S, Gao T, Nichani E, Damian A, Lee JD, Chen D, Arora S. Fine-Tuning Language Models with Just Forward Passes. Advances in Neural Information Processing Systems 2023 Dec 15;36:53038–53075.

74. Maatouk A, Ampudia KC, Ying R, Tassiulas L. Tele-LLMs: A Series of Specialized Large Language Models for Telecommunications. arXiv; 2024. doi: 10.48550/arXiv.2409.05314

75. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference Association for Computational Linguistics; 2019 Aug;3982–3992. doi: 10.18653/v1/d19-1410

76. Mahajan Y, Bansal N, Blanco E, Karmaker S. ALIGN-SIM: A Task-Free Test Bed for Evaluating and Interpreting Sentence Embeddings through Semantic Similarity Alignment. Findings of the Association for Computational Linguistics: EMNLP 2024 Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 7393–7428. doi: 10.18653/v1/2024.findings-emnlp.436

77. Tamine L, Goeuriot L. Semantic Information Retrieval on Medical Texts: Research Challenges, Survey, and Open Issues. ACM Comput Surv 2021 Sep 17;54(7):146:1-146:38. doi: 10.1145/3462476

78. Katsis Y, Rosenthal S, Fadnis K, Gunasekara C, Lee Y-S, Popa L, Shah V, Zhu H, Contractor D, Danilevsky M. MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. arXiv; 2025. doi: 10.48550/arXiv.2501.03468

79. Radlinski F, Craswell N. Comparing the sensitivity of information retrieval metrics. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval New York, NY, USA: Association for Computing Machinery; 2010. p. 667–674. doi: 10.1145/1835449.1835560

80. Park J, Saha S, Han D, De Maesschalck S, Moore R, Korthuis T, Roter D, Knowlton A, Woodson T, Beach MC. Emotional Communication in HIV Care: An Observational Study of Patients' Expressed Emotions and Clinician Response. AIDS Behav 2019 Oct;23(10):2816–2828. PMID:30895426

81. Vamsi/T5_Paraphrase_Paws · Hugging Face. Available from: https://huggingface.co/Vamsi/T5_Paraphrase_Paws [accessed Mar 9, 2025]

82. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, Sriram R, Yang Z, Wang Y, Lin B, Gevaert O, Li L-J, Jain R, Rahmani AM. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. npj Digit Med Nature Publishing Group; 2024 Mar 29;7(1):1–14. doi: 10.1038/s41746-024-01074-z

83. Huang J, Lam MH, Li EJ, Ren S, Wang W, Jiao W, Tu Z, Lyu MR. Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench. arXiv; 2024. doi: 10.48550/arXiv.2308.03656

84. Hewett TT. The role of iterative evaluation in designing systems for usability. Proceedings of the Second Conference of the British Computer Society, human computer interaction specialist group on People and computers: designing for usability USA: Cambridge University Press; 1986. p. 196–214.

85. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, Ye W, Zhang Y, Chang Y, Yu PS, Yang Q, Xie X. A Survey on Evaluation of Large Language Models. arXiv; 2023. doi: 10.48550/arXiv.2307.03109

86. van der Lee C, Gatt A, van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. In: van Deemter K, Lin C, Takamura H, editors. Proceedings of the 12th International Conference on Natural Language Generation Tokyo, Japan: Association for Computational Linguistics; 2019. p. 355–368. doi: 10.18653/v1/W19-8643

87. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol 2023 Oct;9(10):1437–1440. PMID:37615960

88. Heston TF, Khun C. Prompt Engineering in Medical Education. International Medical Education Multidisciplinary Digital Publishing Institute; 2023 Sep;2(3):198–205. doi: 10.3390/ime2030019

89. Wang Y, Wang M, Manzoor MA, Liu F, Georgiev G, Das RJ, Nakov P. Factuality of Large Language Models: A Survey. arXiv; 2024. doi: 10.48550/arXiv.2402.02420

90. Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond. medRxiv 2023 Jul 1;2023.04.18.23288752. PMID:37398329

91. Zhao S, Huang Y, Song J, Wang Z, Wan C, Ma L. Towards Understanding Retrieval Accuracy and Prompt Quality in RAG Systems. arXiv; 2024. doi: 10.48550/arXiv.2411.19463

92. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine 2016 Jun 1;15(2):155–163. doi: 10.1016/j.jcm.2016.02.012

93. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74–81. Available from: https://aclanthology.org/W04-1013/ [accessed Mar 9, 2025]

94. Meincke L, Mollick ER, Mollick L, Shapiro D. Prompting Science Report 1: Prompt Engineering is Complicated and Contingent. Rochester, NY: Social Science Research Network; 2025. Available from: https://papers.ssrn.com/abstract=5165270 [accessed Apr 2, 2025]

95. The language of prompting: What linguistic properties make a prompt successful? - Google Search. Available from: https://www.google.com/search?client=firefox-b-1-d&q=The+language+of+prompting%3A+What+linguistic+properties+make+a+prompt+successful%3F [accessed Apr 12, 2025]

96. Lamichhane B. Evaluation of ChatGPT for NLP-based Mental Health Applications. arXiv; 2023. doi: 10.48550/arXiv.2303.15727

97. Grall C, Tamborini R, Weber R, Schmälzle R. Stories Collectively Engage Listeners' Brains: Enhanced Intersubject Correlations during Reception of Personal Narratives. Journal of Communication 2021 Apr 1;71(2):332–355. doi: 10.1093/joc/jqab004

98. Lahiri AK, Hu QV. AlzheimerRAG: Multimodal Retrieval Augmented Generation for PubMed articles. arXiv; 2024. doi: 10.48550/arXiv.2412.16701

99. Zhou S, Luo X, Chen C, Jiang H, Yang C, Ran G, Yu J, Yin C. The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries. International Journal of Surgery 2024 Oct;110(10):6509. doi: 10.1097/JS9.0000000000001850

100. Arasteh ST, Lotfinia M, Bressem K, Siepmann R, Adams L, Ferber D, Kuhl C, Kather JN, Nebelung S, Truhn D. RadioRAG: Factual large language models for enhanced diagnostics in radiology using online retrieval augmented generation. arXiv; 2024. doi: 10.48550/arXiv.2407.15621

101. Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, Hu Q, Pan C, Wang D, Liu Z, Shi W, Shi D, Li F, Qu B, Zheng Y. Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study. J Med Internet Res 2024 Nov 8;26:e58041. PMID:39046096

102. Jabarulla MY, Oeltze-Jafra S, Beerbaum P, Uden T. MedDoc-Bot: A Chat Tool for Comparative Analysis of Large Language Models in the Context of the Pediatric

Hypertension Guideline. 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2024. p. 1–4. doi: 10.1109/EMBC53108.2024.10781509

Appendix 1: Level 1 Prompt Engineering

Table 1. *Initial prompts used for generating LLM responses to provide personalized PrEP support. These prompts were iteratively refined by developing several intermediate prompts to improve the quality of generated responses.*

Category	Description	Prompt
Context Handling (CN) Prompts	CN-1: Prompt used for identifying whether additional context is needed	Classify whether the user needs to provide any additional information or context for you to answer the following query - {query}
	CN-2: Prompt used for instructing LLM to ask user for additional context	Ask user to provide more information or context in relation to the query enclosed in square brackets. [{query}]
Information Subclassification (IC) Prompt	IC: Prompt for sub-classifying information queries as seeking information or experiential expertise support	Classify the following text as seeking advice, experience, or information: {query}
Personalized Information (PI) Prompts	PI-1: Prompt used for identifying topic discussed in the query for retrieving similar documents based on topic match.	Topic list = {topic_list} From the above topic list, identify a single main topic that is discussed in the following query, {query}. If does not directly relate identify a closely related topic. Provide only the topic.
	PI-2: Prompt used for generating responses to informational query using the retrieved documents.	[{messages}] Using the information enclosed in square brackets, provide a response to the following user query.
	PI-3: Prompt used to refine generated response.	[{response text}] Consider the following user query – {query}. If the query does not contain specific details such as location, age, health condition, etc., check whether the response enclosed in square brackets contains such details and if it contains, paraphrase the text enclosed in square brackets to make sure that the response is general and does not include such

		details. Otherwise return the text enclosed in square brackets.
Personalized Experience (PX) Prompt	Prompt used for generating response to patient expertise query using the retrieved documents.	[{messages}] Using the information enclosed in square brackets above, provide a response to the following user query
Personalized Emotional (PE) Prompts	PE-1: Prompt used for generating response to emotional query using the retrieved documents.	[{messages}] Consider the following Feeling = Emotion categories: Sad = Tired, bored, lonely, depressed, ashamed, guilty, sleepy, apathetic, isolated, inferior, stupid, remorseful Mad = Hurt, hostility, angry, selfish, hateful, critical, distant, sarcastic, frustrated, jealous, irritated, skeptical Scared = Confused, rejected, helpless, submissive, insecure, anxious, bewildered, discouraged, insignificant, inadequate, embarrassed, overwhelmed Joyful = Excited, sensuous, energetic, cheerful, creative, hopeful, daring, fascinating, stimulating, amused, playful, optimistic Powerful = Aware, proud, respected, appreciated, important, faithful, surprised, successful, worthwhile, valuable, discerning, confident Peaceful = Nurturing, trusting, loving, intimate, thoughtful, content, thankful, secure, serene, responsive, pensive, relaxed Identify the feeling expressed in the following expression based on the feelings=emotions listed above - {query} For the identified feeling, generate a suitable response providing understanding, empathy, sympathy, reassurance, validation, relief of blame, or compliment. Use information from text enclosed in square brackets to generate an emotional response.
	PE-2: Prompt used for generating responses to emotional query when no relevant	Consider the following Feeling = Emotion categories: Sad = Tired, bored, lonely, depressed, ashamed, guilty, sleepy, apathetic, isolated, inferior, stupid, remorseful

	documents are retrieved.	Mad = Hurt, hostility, angry, selfish, hateful, critical, distant, sarcastic, frustrated, jealous, irritated, skeptical Scared = Confused, rejected, helpless, submissive, insecure, anxious, bewildered, discouraged, insignificant, inadequate, embarrassed, overwhelmed Joyful = Excited, sensuous, energetic, cheerful, creative, hopeful, daring, fascinating, stimulating, amused, playful, optimistic Powerful = Aware, proud, respected, appreciated, important, faithful, surprised, successful, worthwhile, valuable, discerning, confident Peaceful = Nurturing, trusting, loving, intimate, thoughtful, content, thankful, secure, serene, responsive, pensive, relaxed Identify the feeling expressed in the following expression based on the feelings=emotions listed above - {query} For the identified feeling, generate a suitable response providing understanding, empathy, sympathy, reassurance, validation, relief of blame, or compliment.
Cohesive Response (CR) Prompt	Prompt used for combining responses generated for sub-queries.	[{response text}] Smoothen the flow of information in the text enclosed in square brackets.