

Protocol

A Conversational Agent for Providing Personalized Preexposure Prophylaxis (PrEP) Support: Protocol for Chatbot Implementation and Evaluation

Fatima Sayed¹, MS; Albert Park², PhD; Patrick Sean Sullivan³, DVM, PhD; Yaorong Ge¹, PhD

¹Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, United States

²School of Health Information Science, University of Victoria, Victoria, BC, Canada

³Department of Epidemiology, Emory University, Atlanta, GA, United States

Corresponding Author:

Albert Park, PhD
School of Health Information Science
University of Victoria
PO Box 1700 STN CSC
Victoria, BC V8W 2Y2
Canada
Phone: +1-250-721-8575
Email: alpark@uvic.ca

Abstract

Background: Chatbots have the potential to reduce barriers to preexposure prophylaxis (PrEP), including lack of awareness, misconceptions, and stigma, by providing anonymous and continuous support. However, in the context of PrEP, chatbots are still nascent; they lack personalized informational expertise, peer experiential expertise, and human-like emotional support to facilitate future PrEP uptake. These personalized and relatable forms of support are crucial for increasing engagement, influencing health decisions, and fostering resilience and well-being.

Objective: In this paper, we describe the iterative development and evaluation plans of a retrieval-augmented generation (RAG) chatbot for providing personalized information, peer experiential expertise, and human-like emotional support to PrEP candidates.

Methods: We used an iterative design process consisting of 2 phases: prototype conceptualization and iterative chatbot development. In the conceptualization phase, we identified real-world PrEP needs and designed a functional dialogue flow diagram for PrEP support. Chatbot development included developing 2 components: a query preprocessor and a RAG module. The preprocessor uses the Segment Any Text (SAT) tool (developed by Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl) 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 used Sentence-Bidirectional Encoder Representations from Transformers (SBERT) embeddings with cosine similarity, and performed topic matching to identify topically relevant documents based on the query topic to support document retrieval. Extensive prompt engineering was used to guide the large language model (LLM), Gemini-2.0-flash, in generating tailored responses. We conducted 10 rounds of internal evaluations to assess and improve the chatbot responses based on 10 criteria: clarity, accuracy, actionability, relevancy, information detail, tailored information, comprehensiveness, language suitability, tone, and empathy. Finally, we conducted a blinded comparative study and used a linear mixed model to validate the chatbot against a general LLM and real-world user 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 responses tailored to information, experiential, and emotional user needs. We found that prompt effectiveness varied with task complexity; this was likely due to LLM sensitivity to the structure of prompts and to linguistic variability. Prompt decomposition and segmenting prompt instructions helped improve comprehensiveness and relevancy for complex and long queries. The linear mixed model analysis revealed that while both the general LLM and RAG were preferred over user responses, the general LLM maintained a superior edge over the RAG chatbot across most criteria, with the exceptions of empathy and tone.

Conclusions: Our RAG chatbot leverages social media data to provide personalized information, peer experiences, and human-like emotional support; these elements are essential for addressing PrEP misconceptions and promoting self-efficacy. Further analysis incorporating expert and user feedback will be conducted to help validate and improve the chatbot's potential.

International Registered Report Identifier (IRRID): PRR1-10.2196/79966

JMIR Res Protoc 2026;15:e79966; doi: [10.2196/79966](https://doi.org/10.2196/79966)

Keywords: HIV prevention; chatbots; conversational agents; large language models; retrieval-augmented generation; generative AI; PrEP; preexposure prophylaxis

Introduction

Background

Increasing the uptake of preexposure 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 has been hindered in the United States by lack of PrEP knowledge and prevailing anticipated and experienced stigma and discrimination [5,7-9]. Personal experiences shared by peers can improve the awareness [10,11], self-efficacy, and intrinsic motivation [12] of people considering PrEP use, mitigating their stigmatized beliefs [10], and informing health decisions [10,13]. People without financial means or ready access to providers experience limited access to health care [14,15]. Service deserts result in a lack of physically proximate service providers [6]; scarcity of services also disproportionately impacts those without financial means or access [16]. Equitable PrEP uptake requires facilitating access to PrEP resources while meeting the diverse needs of PrEP candidates [16]. Conversational agents (ie, chatbots) have the potential to provide anonymous, round-the-clock assistance in finding a PrEP provider, answering questions about PrEP, and providing emotional support. These services can help to overcome geographical and access barriers to seeking PrEP care.

Integrating chatbots into existing public health programs can significantly mitigate treatment barriers while enhancing user trust and adoption. However, to ensure scalability, deployment must account for infrastructure constraints and linguistic diversity. In low-resource settings where internet access, device affordability, and digital literacy may be limited [17], chatbots should be designed for local deployment to ensure data privacy and accessibility. This includes supporting offline functionality for mobile devices and using lightweight architectures to ensure accessibility on older hardware. Furthermore, to cater to cultural and behavioral preferences, it is essential to incorporate multilingual and multiplatform support. Providing such user-centric personalization is a vital step in reducing health disparities and ensuring that digital interventions are both accessible and equitable.

HIV and PrEP Chatbots

Chatbots have demonstrated potential in disseminating HIV information [18,19], self-testing [18,20,21], access to and uptake of HIV prevention and care [22-24], and making

positive behavioral changes relevant to HIV prevention and care (eg, self-disclosure of HIV status [25,26] and self-management of PrEP adherence [18,22,27]). AI technologies can be useful in developing and operating chatbot services, but their contribution is variable depending on the risk and complexity associated with health infrastructure. This is because risk and complexity influence the adoption of chatbots by users for improving health conditions [28]. Rule-based chatbots, which use scripted text based on simple rules (ie, keyword identification and pattern-matching techniques), are most commonly used to provide information [19,29] and support self-management of health behaviors (eg, setting up reminders) [22,25]. More advanced chatbots use information retrieval systems [23] and knowledge graphs [21] to help users identify HIV risk and navigate relevant information to mitigate barriers to PrEP uptake. For more complex needs such as identifying disorders and providing behavioral support (eg, for PrEP retention), chatbots use hybrid techniques that use AI tools (eg, machine learning [20,21,23] and natural language processing [21]) to learn patterns from previous conversations [30]. These approaches can also detect emotions [30] and perceive user characteristics based on past interactions [21], and can integrate dialogue rules for response generation. Two protocols for studies mention the use of a hybrid large language model (LLM; HumanX) chatbot for promoting PrEP uptake and use [24,27] and for providing HIV-related mental health support [22]. However, these studies do not specify details of the chatbot implementation. Such rule-based and hybrid health care chatbots are often criticized for their limited flexibility [31,32] and diversity of content [31-36], and insufficient contextual understanding [37,38] and personalization [23,39]. Such chatbots have also been reported to lack human-like emotional support [40], and this limits long-term user engagement and chatbot effectiveness [41]. LLMs, leveraging vast knowledge bases, can generate responses that are varied and context-relevant compared to rule-based and hybrid systems [42]. Beyond information retrieval, LLMs have demonstrated a significant capacity for cognitive empathy, the intellectual ability to identify and label a user's emotional state and respond with supportive language [43,44]. However, researchers have noted that these responses often lack affective empathy [44], as the models lack human-like emotional experience and intrinsic motivation [43-45], leading to impersonal responses in sensitive and stigma-associated contexts [45]. This empathy gap suggests that while LLMs can extend general warmth, they struggle to provide the deep, personalized emotional connection

found in human-to-human emotional support (ie, human-like emotional support).

To date, no study details the implementation of an LLM-based chatbot for providing various HIV and PrEP support, including emotional, informational, and peer experiential support to facilitate access to PrEP information and promote PrEP uptake and adherence. Although LLMs can offer improved personalization and human-like conversations [42,46], their performance can also be limited by their scope of specialized knowledge, diversity of available data, and the potential for LLM hallucinations [47, 48]. Fine-tuned LLMs have been effective in performing domain-specific tasks; however, the small size of the training corpus limits their ability to generalize model capabilities across diverse user needs [31]. Prompt engineering [49] and retrieval-augmented generation (RAG) [48] techniques improve LLMs' performance in terms of accuracy [50,51], relevancy [50,52], data privacy [53], and user personalization [54,55], which can enhance user trust in chatbot-provided support [56-58]. Using prompt engineering, LLMs can be guided to improve the precision and accuracy of generated responses [57]. RAG chatbots have indicated improved accuracy (34%) [49] and reduced hallucination (26.5%) [48] in generated responses compared to LLM question-answering bots. Compared to fine-tuned LLMs, RAG provides improved quality of responses in specialized domains by leveraging diverse external knowledge [27]. Using effective prompts to guide LLMs [32], RAG chatbots can offer greater personalization and relevance compared to rule-based and hybrid systems [27].

However, the development of chatbots based on LLM and RAG techniques is still nascent in the context of HIV and PrEP support and has not been implemented to address implicit user needs [48], answer long contextual queries [48], and generate personalized responses [59]. Addressing these issues in RAG chatbots can lead to providing accurate, flexible, personalized, and human-like responses to promote PrEP uptake.

Available literature [60-62] and our prior work have identified human-like emotional support components as essential elements to support behavior change for HIV medication taking and PrEP uptake. However, no prior work has developed RAG chatbots to provide personalized information and peer experiential expertise (eg, user stories) [63]. Existing HIV and PrEP chatbots focus on providing informational support [19,22,29] and emotional support according to therapeutic guidelines [27]. Effective support provided by humans relies on emotional support (ie, human-like emotional support) [64,65] and peer experiential expertise (ie, peer experiential support) [63,66], elements which are absent in existing health care chatbots. Providing effective human-like emotional support first requires understanding and communicating complex human emotional experiences [67-69]. The Willcox Feeling Wheel facilitates this process by providing a related vocabulary for 6 emotion categories [69]. To better reflect the real-world user queries that often seek practical user experiences (peer experiential expertise) [70], PrEP chatbots should

provide peer experiential expertise support through pragmatic narratives of others in similar situations, which might enhance self-coping skills and emotional confidence, leading to improved quality of life [63,66]. Additionally, chatbots need to support back-and-forth dialogue exchanges, similar to real-world help-seeking behaviors, to facilitate clarification and understanding of complex concerns [71].

We develop a RAG chatbot that addresses the knowledge gaps in providing personalized PrEP support. We develop a support classifier to identify implicit PrEP user needs from long context queries by fine-tuning Google's Gemma model. The RAG chatbot uses personalized prompts and an external knowledge base (facts and real-world PrEP experiences) to generate responses tailored to users' informational, experiential, and emotional needs. To provide human-like emotional support, we use the Willcox Feeling Wheel to guide the chatbot in responding empathetically to the emotions expressed by the user. The chatbot also provides personalized peer experiential expertise support through pragmatic narratives of users in similar situations. We use extensive prompt engineering and fact validation techniques to improve accuracy, personalization, and transparency in generated responses.

Protocol Goals

This protocol describes the development of a RAG chatbot in 2 phases. Phase 1 describes conceptualizing the chatbot architecture, and Phase 2 includes iterative development of the RAG chatbot to provide personalized information, peer experiential expertise, and human-like emotional support. Before 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 paper. We demonstrate the iterative development process focusing on prompt engineering, which is an integral part of the LLM RAG chatbot implementation. An exploratory user study was conducted to assess the quality of chatbot responses compared to real-world user responses and general LLM responses. We also discuss the implications of these findings, highlighting the feasibility of the chatbot prototype as a tool to facilitate future PrEP support.

Methods

RAG Architecture Overview

To improve accuracy and empathy, we designed an RAG architecture that explicitly separates informational, experiential expertise, and emotional support. Unlike a traditional LLM, which processes a user query based solely on its internal training data, this RAG framework uses a support classifier to identify specific user needs. The RAG chatbot then queries specialized knowledge bases to deliver tailored, evidence-based responses to the identified social support needs. This modular approach helps in controlling hallucination and improving the accuracy of the support provided. The mechanism involves segmenting user queries into specific needs and retrieving information from separate factual and

experiential databases; the system ensures that responses are both evidence-based and empathetically aligned. The key features of the RAG chatbot include the following:

1. **Modular retrieval:** The system first segments and classifies user queries into 3 distinct categories—informational, experiential expertise, or emotional support. Using the enriched query segment, the RAG chatbot retrieves documents from specialized databases. It pulls from a clinical facts database for informational needs and an anonymized peer-experience database for experiential and emotional needs. This ensures that the chatbot only references verified data for facts and draws from real-world human narratives to provide empathetic and experiential support.
2. **Contextual enrichment:** A dedicated context manager handles query enrichment by ensuring sufficient information is present to respond to the query. If the query is underspecified, the manager prompts the user for additional context to refine the retrieval process.
3. **Response generation:** The retrieved context is integrated with specialized system prompts to generate the final response. During this stage, the system leverages a human emotional framework to map the user's expressed feelings to a matching emotional vocabulary. This approach facilitates nuanced emotional support that aligns with the user's specific emotional state and is grounded in real-world human empathy. To maintain cultural authenticity and relatability, the system incorporates direct quotes and real-world narratives from the experience database, providing responses grounded in genuine human experience rather than synthetic advice.

Phase I: Prototype Conceptualization

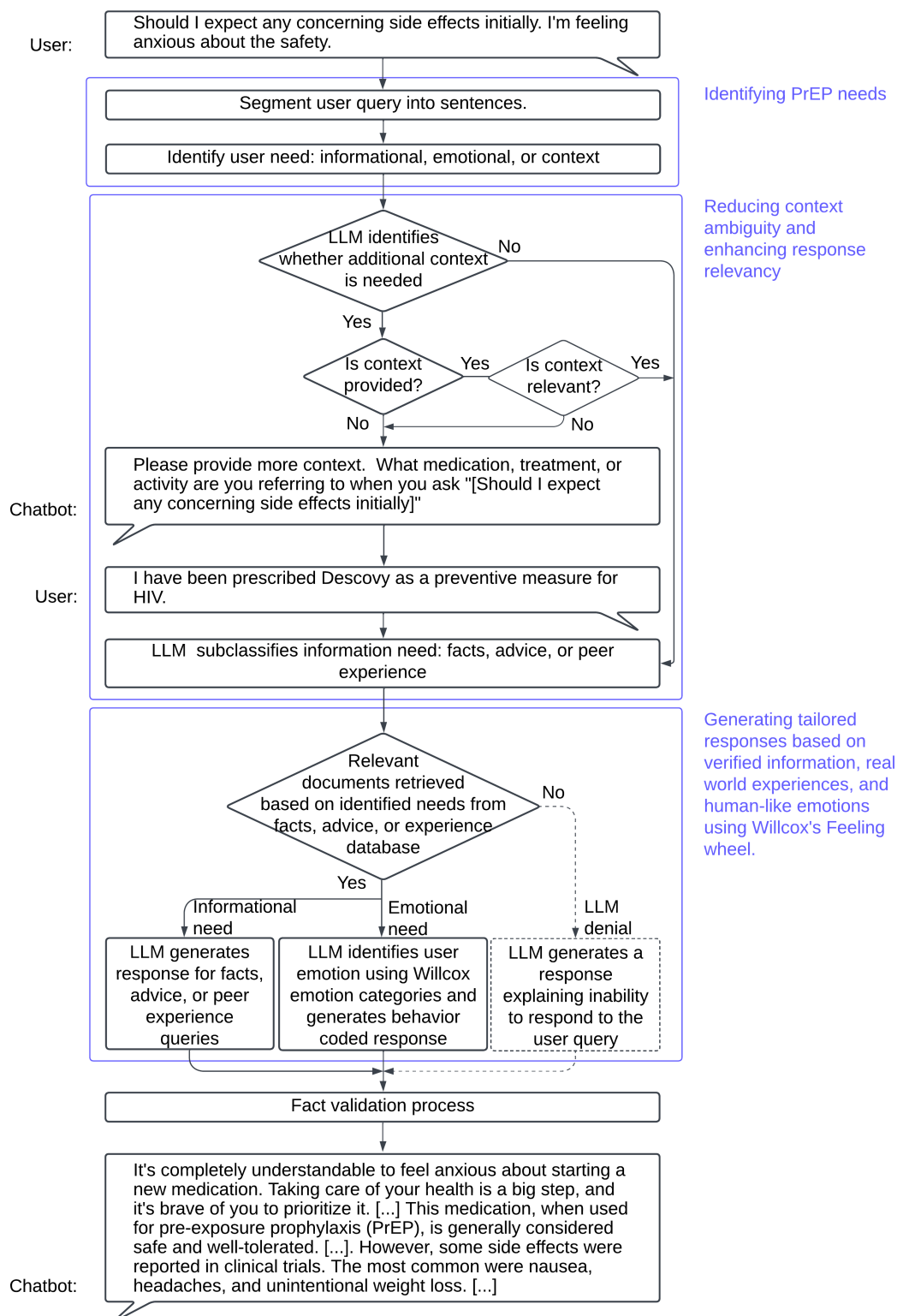
To inform the goals of our RAG chatbot prototype, we analyzed 3020 publicly available Reddit conversations focused on PrEP from April 2011 to March 2024. We focused on the informational (including peer experiential expertise) and emotional (ie, psychosocial) support because these are the 2 main types of support that are exchanged in this forum and because these have the potential to impact broader

behavioral trends related to PrEP use [72]. Our database for RAG aligns with the previous literature [73]. We found that more than 88% (2658/3020) of user PrEP needs included informational and emotional need items. Informational needs centered on PrEP facts and related experiences, and emotional needs focused on reassurance, HIV and PrEP concerns sharing, empathy, and sympathy, based on the psychosocial constructs of Social Support Behavior Code (SSBC) [74, 75]. PrEP candidates' emotions can be described with the Willcox Feeling Wheel [69], including sad, mad, scared, joyful, peaceful, and powerful.

Real-world user queries were multifaceted, consisting of multiple informational and emotional needs. Chatbots must (1) preprocess each input query into semantic segments, (2) generate responses for each segment, and (3) combine them into a final response. Based on literature [73] and our prior work, our chatbot was designed to (1) generate personalized responses, (2) provide accurate and comprehensive information to user queries, (3) provide peer experiential expertise support, and (4) respond empathetically to the emotions expressed by the user. Functional dialogue flow diagrams were created (Figure 1) to visually communicate with our team members and developers how the chatbot would achieve these aims.

To identify user needs, we created a PrEP-related dataset. This dataset was annotated with 3 categories: informational (including facts and peer experiential expertise), emotional, and context. Subsequently, we built a classifier to parse user inputs and categorize them into the 3 categories. The classifier further classified informational needs into facts and peer experiential expertise. For informational needs, we specifically designed the chatbot to provide information (eg, facts and actionable advice) and peer experiential expertise support (ie, other people's experiences and opinions). For emotional needs, we designed the chatbot to identify user emotions based on the Willcox Feeling Wheel [69] and to generate a suitable response using relevant SSBC psychosocial constructs known to impact PrEP uptake (eg, reassurance, understanding, empathy, sympathy, validation, relief of blame, and compliment) [60,76].

Figure 1. An example of a generated response using a paraphrased real-world query from Reddit. The retrieval-augmented generation process includes (1) semantically segmenting the query into sentences and classifying each segment into 1 of the 3 categories: “information,” “context,” or “emotion”; (2) checking for additional context; (3) retrieving the top 5 relevant documents from specific databases; (4) identifying user emotions from emotional segments and retrieving documents via topic matching; and (5) generating a response using personalized prompts. The dotted line in the diagram indicates ongoing updates. LLM: large language model; PrEP: preexposure prophylaxis.



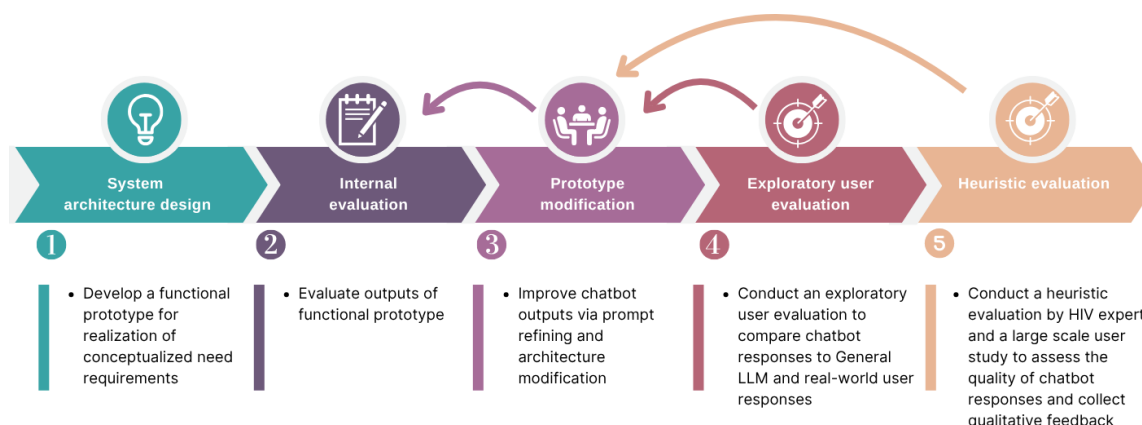
Phase II: Iterative Prototype Development

Overview

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 the RAG chatbot based on internal evaluation feedback, (4) exploratory user evaluation to compare the RAG chatbot responses against general LLM and real-world user responses, and (5) a heuristic evaluation by HIV experts and a large-scale user study to validate accuracy, safety, ethical alignment, and user trust of chatbot responses prior to real-world deployment.

Figure 2. Iterative development and refinement process of the retrieval-augmented generation chatbot for providing support to preexposure prophylaxis candidates. The process initiated with (1) system architecture design, followed by (2) internal evaluation involving developers and an HIV expert to assess functional requirements. This resulted in a recursive cycle of (3) prototype modification, in which the retrieval-augmented generation chatbot is modified based on internal evaluation and findings from (4) exploratory user study and future (5) heuristic evaluation and large-scale user study. LLM: large language model.



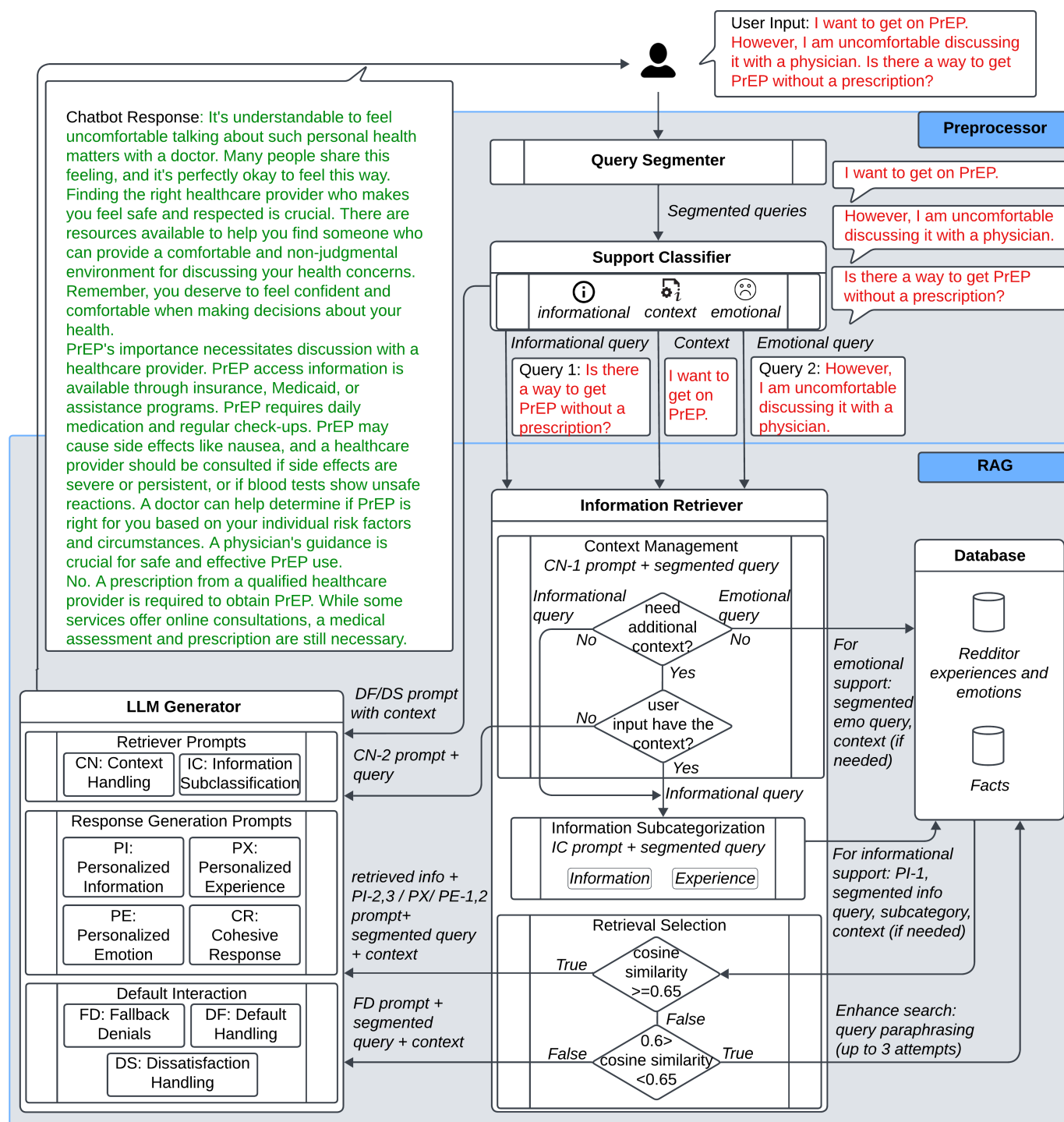
System Architecture

Overview

The RAG chatbot was developed using Python programming and by accessing the Google Gemini API. The chatbot consists of 2 main modules (Figure 3): a preprocessor and a RAG module. The RAG, in turn, leverages 2 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. To enhance model determinism and minimize irrelevant generation, the information retriever uses the extracted information to handle context management and information subclassification to enhance semantic retrieval and output relevant documents. Finally, the LLM generator synthesizes the retrieved documents and user context through personalized prompt engineering, providing a robust framework for aligning model behavior with user-specific needs while maintaining high factual fidelity.

Figure 3. Architecture of the retrieval-augmented generation chatbot used to generate personalized information, peer experiential expertise, and human-like emotional responses for providing tailored responses. LLM: large language model; PrEP: preexposure prophylaxis; RAG: retrieval-augmented generation.



Preprocessor Module

The preprocessor module consists of 2 components: a query segmenter and a support classifier.

Query Segmenter

The query segmenter semantically segments a user input into sentences using the SAT tool [77]. The tool is highly robust because it is minimally reliant on punctuation and space characters. This makes it an optimal choice for handling

diverse writing styles and complex real-world user queries. Each segment is then passed to the support classifier.

Support Classifier

Overview

The support classifier was used to classify the intent of the input query segments to guide the retrieval process. For segments identified as informational, a few-shot information subclassifier served as a routing layer: it directed experience-based queries to the experiential database and fact-based

queries to the factual database. This classification-driven routing ensures that the retrieval pipeline is grounded in the appropriate knowledge database. A detailed description of this routing logic and how classifier labels trigger specific database retrieval is provided in the “Information Subcategorization” section.

Dataset

We created the PrEP dataset, which consists of semantic sentences extracted from online user PrEP discussions and manually labeled them into 1 of the 3 categories—informational, emotional, and context. For fine-tuning, we used 2 batches of data: the entire ($n=32,081$) PrEP dataset (class-unbalanced) and a class-balanced subset ($n=10,041$) of the PrEP dataset.

Classifier Fine-Tuning

We developed 6 support classifiers by fine-tuning each batch of data using 3 models: Bidirectional Encoder Representations from Transformers (BERT)-base-uncased model, Gemma 2B-it [78], and Gemma 2 2B-it [78]. Gemma model series is lightweight and hardware-flexible, offering low latency and balancing performance and efficiency, making it suitable for chatbot deployments [78]. BERT is the state-of-the-art model used in text classification tasks.

We used Parameter-Efficient Fine-Tuning (PEFT) with standard Low-Rank Adaptation (LoRA) configurations ($\alpha=32$ and $\text{rank}=64$) [79] on a sequential classification task for fine-tuning the Gemma models. Based on previous literature on LLM fine-tuning performance, we fine-tuned the Gemma 2B-it model for 5 epochs with an early stopping patience of 2 epochs [80,81]. Similarly, the BERT model was fine-tuned for 5 epochs.

Evaluation Metrics

We used F_1 -score and accuracy to compare the performance of the 6 classifiers on the class-balanced (macro average) and class-unbalanced (weighted average) datasets.

RAG Module

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

Information Retriever

The information retrieval process includes context management, information subcategorization, and retrieval selection.

Context Management

The context handler instructs the LLM (Gemini-2.0-flash) using the context handling prompt 1 (CN-1) to determine

whether queries require additional context. If no additional context is needed, only the user query is used for retrieving relevant documents. If additional context is needed, our chatbot will ask users for additional context using the context handling prompt 2 (CN-2). With additional context provided by the user, the chatbot then retrieves relevant documents using both the context and user query.

Information Subcategorization

To generate personalized responses, the information needs were further subcategorized into (1) seeking facts or (2) requests for peer experiential expertise, using the information subclassification few-shot prompt (IC; Table S1 in [Multimedia Appendix 1](#)). The identified information subcategory (ie, facts or peer experiential expertise) was used to guide the retrieval process. For queries seeking facts, documents were retrieved from the facts database, and for queries seeking peer experiential expertise, documents were retrieved from the experience database. 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.

Retrieval Selection

The retriever uses Sentence-Bidirectional Encoder Representations from Transformers (SBERT) embeddings [82] to retrieve relevant documents from the database. We used SBERT embeddings due to their bidirectional contextual understanding [82], computational efficiency [82], and superior performance in identifying semantically similar sentences, coupled with synonyms and negation lexical variations [83], which are better suited to cover various user writing styles. The retriever uses cosine similarity, and the LLM identifies topics for information retrieval; this was done to provide facts for informational or emotional support. We used topic matching to enhance the retrieval of a wide variety of relevant documents from the facts database, including context-rich queries that challenge semantic retrieval [84] in short query databases. We manually reviewed the facts database to identify 47 broad HIV-PrEP-related topics ([Textbox 1](#)). To annotate the facts database, we instructed the LLM (Gemini-1.5-flash) to assign the most suitable topic to each query from this list. If no relevant topic is identified, we instructed the LLM to generate a closely relevant topic to have the ability to adapt to unforeseen user needs. We then manually reviewed the annotated topics to include any LLM-generated new topics in 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 retrieval-augmented generation topic document retrieval for factual support. The HIV-PrEP (pre-exposure prophylaxis) topic list was created by manually reviewing the facts database and updated using large language

model topic assignment. Topic document retrieval from the facts database was used to improve sensitivity in unstructured queries.

(1) PrEP; (2) PrEP eligibility; (3) PrEP effectiveness; (4) PrEP access and cost in Australia; (5) PrEP access and cost; (6) HIV testing; (7) HIV self-test; (8) postexposure prophylaxis (PEP); (9) PrEP administration and dosage; (10) aprepivir or injectable PrEP; (11) PrEP side effects; (12) HIV and AIDS; (13) risk and transmission; (14) HIV prevention strategies; (15) safety and precautions; (16) cabotegravir; (17) sexually transmitted infections (STIs); (18) viral load or undetectable or untransmissible; (19) HIV treatment; (20) PrEP medications; (21) emtricitabine or tenofovir disoproxil fumarate; (22) emtricitabine or tenofovir disoproxil fumarate administration; (23) emtricitabine or tenofovir disoproxil fumarate side effects; (24) drug interactions; (25) emtricitabine or tenofovir disoproxil fumarate eligibility; (26) antiretroviral or HIV drug side effects; (27) test disclosure and anonymity; (28) stigma and its impact on people living with HIV; (29) PEP access and cost; (30) PEP side effects; (31) Aprepitide or injectable PrEP side effects; (32) cabotegravir side effects; (33) PEP eligibility; (34) PrEP access and cost in the United Kingdom; (35) medication adherence; (36) Truvada side effects; (37) healthy lifestyle; (38) Descovy; (39) PrEP safety and precautions; (40) HIV prevention strategies; (41) Truvada; (42) PrEP eligibility in the United Kingdom; (43) Descovy side effects; (44) HIV or AIDS drugs; (45) Truvada access and cost; (46) Descovy access and cost; and (47) emtricitabine or tenofovir disoproxil fumarate safety and precautions.

Retrieval for Factual Support

For factual support, we retrieved the top 5 documents from the facts database with a cosine similarity score ≥ 0.65 . The cosine similarity thresholds were determined based on experiments, and we set $k=5$ for document retrieval based on previous evaluation benchmarks [85,86]. In addition, we selected 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 instructed 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 was based on the context need. If context is needed, we retrieved 3 sets of top 5 relevant documents (Figure 4) by comparing cosine similarity between

(1) the context-added query and database questions (m1), (2) the context-added query and database answers (m2), and (3) the context-free query and database questions (m3; Figure 4). Both database questions and answers were semantically searched to improve the retrieval precision. Cosine similarity between the context-free query and database answers did not improve the retrieval precision and was therefore not used in processing the context-free queries. As the majority of the facts database consists of short questions from PrEP factsheets and brochures, the retriever often failed to retrieve relevant information for queries with additional context. If no context is needed, we retrieved the top 5 documents by only comparing cosine similarity between the context-free query and the database questions.

Figure 4. Retrieval algorithm pseudocode for improving retrieval precision in factual queries using cosine similarity.

```

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

```

Retrieval for Peer Experiential Expertise and Emotional Support

For peer experiential expertise support, we retrieved 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 used both the query and context for retrieval to overcome context-dependent retrieval failure. For emotional queries, the top 5 similar documents were retrieved from the experience database in addition to providing relevant facts. Similar to document retrieval for peer experiential expertise support, we retrieved the top 5 documents from the experience database using cosine similarity with a score of at least 0.65 and, when context is needed, used both the query and context for retrieval. As information can be helpful in addressing certain HIV-PrEP emotional concerns [87], we retrieved facts relevant to the emotional query.

To improve the sensitivity, we also used up to 3 paraphrase queries with a lower cosine similarity score, between 0.6 and 0.65. For paraphrasing, we used the T5 paraphrase generator from the Transformers Hugging Face library [88] for its demonstrated superior performance over GPT-3 and ChatGPT in capturing semantic relevance between sentences [89].

LLM Generator

Once all the necessary data were retrieved, we instructed the LLM (Gemini-2.0-flash) to generate responses using the retrieved information. This process was guided by

personalized prompts (PI [personalized information], PE [personalized emotion], and PX [personalized experience]) and the user need (ie, informational, peer experiential expertise, and emotional responses) as described in Table S1 in [Multimedia Appendix 1](#). We iteratively refined 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 used the 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 used PE-1 to first identify the feelings based on the 6 Will-cox feeling categories and to generate a suitable response reflecting SSBC psychosocial constructs (ie, sympathy, understanding, empathy, relief of blame, validation, compliment, or reassurance) for the identified feeling. Identifying user feelings allows responding with 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 [90]. The RAG LLM was instructed to use the style and tone from the retrieved emotional documents, which comprise peer-to-peer emotional support. When no relevant documents were retrieved for an

emotional query, we used 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. Fact responses were validated by the fact validator. Then, the cohesive response (CR) prompt combines all subquery responses, instructing the LLM to smooth the information flow into a single CR. Finally, the LLM provides a combined response.

Fact Validator

Responses providing factual information were validated with the documents from the facts database, a step toward eliminating LLM hallucination. Our fact validation algorithm consists of 2 steps: fact identification and fact validation. To reduce overreliance on LLMs, we used a deterministic technique to identify semantically relevant facts from the retrieved documents (ie, the same top 5 documents retrieved from the information retriever process). This is done to address a known issue: that LLMs often provide imperfect fact identification when they break down complex information [91,92]. To identify the most relevant facts for validation, we (1) segmented the generated response and retrieved documents using the SAT tool and then (2) extracted the most relevant fact-supporting document segment (step 1) based on SBERT and cosine similarity (ie, highest similarity score). This creates pairs of generated response segments that correspond to their fact-supporting document segments. The LLM was instructed to use the response-document pairs and the FV-1 prompt (step 2) to identify any response segments that are not supported by the document segments identified with cosine similarity. Providing the subset of relevant fact segments to the LLM narrows its focus, improving validation accuracy. Then, using any generated response segments with inaccurate facts and the FV-2 prompt, the LLM was instructed to extract supporting facts from the retrieved documents (step 3).

To evaluate the performance of the fact validator, we manually verified the performance for 1033 fact response segments from 100 generated responses for randomly sampled query responses that the LLM generated. The fact validator achieved a 97.39% validation accuracy based on 1006 sample items. The accuracy was calculated as the percentage of fact response segments that were fully supported by the facts database, as determined by the research team. Only 3% (27/1033) of the generated response segments were not able to be validated with the facts database (95% CI 1.85-3.88). These included response segments providing general information (n=12, eg, “Other side effects [of the medication] include changes in the immune system, dizziness, diarrhea, nausea, vomiting, headache, rash, and gas.”) and personalized information (n=8, eg, “Given the recent unprotected anal sex with someone who recently learned they had syphilis, your friend should get tested for STIs as soon as possible.”). We also found false negatives (n=7), where the LLM extracted facts that could not support the generated response segment in step 3. For example, when validating a response segment (eg, “Truvada alone is not sufficient for post-exposure prophylaxis (PEP).”) with an accurate and semantically retrieved document segment, the LLM extracted

facts that could support the generated response segment (eg, “It is worth noting that Truvada must be used correctly (without skipping doses) whether it is used in combination with other ARVs in the control of HIV, or for PrEP.”). Based on these data, we opted for a 3-step validation process because the semantic retrieval (step 1) identified over 82.10% (848/1033) of accurate response segments, whereas the LLM identified (step 2 and step 3) 15.30% (158/1033) of accurate response segments. It is also important to note that we also found that LLM-based validation was less consistent compared to the deterministic techniques. For instance, when we processed the same sets of 100 samples 10 times, the LLM changed its final validation responses (step 2 and step 3) in 44% of samples. It is important to note that the 44% does not indicate that earlier responses were inaccurate. Rather, the nondeterministic nature of LLMs (as probabilistic models) relying solely on LLMs can hinder the validation process. This highlights the need for mixed validation approaches, combining deterministic techniques and LLMs to eliminate LLM hallucinations.

Internal Evaluation of the Chatbot Responses and Prompt Refinement

Internal Evaluation

The RAG chatbot was evaluated to assess the functional requirements following the iterative evaluation design process [93], with a focus on (1) accurately identifying user informational and emotional needs, (2) providing relevant and accurate information, and (3) generating personalized and emotionally suited responses. Developers conducted 10 rounds of internal evaluation of chatbot responses. Paraphrased user inputs from the experience database were used to create 2 test queries for each support need (ie, informational, peer experiential expertise, and emotional). In each of the 10 evaluation rounds, we selected (1) one simple and direct query and (2) one complex and context-dependent query for each support need. Responses were generated for the 6 test queries per round, which were then iteratively reviewed in 10 rounds of internal evaluation (n=60) by at least 2 chatbot developers and at least 1 HIV expert. This qualitative evaluation of 60 total responses helped to identify inaccuracies and refine the LLM prompts and was followed by chatbot modification to address issues identified by the reviewers. The most frequently occurring issues were as follows: (1) low retrieval sensitivity and precision for context-dependent queries, (2) contradicting information from user discussions, (3) response generation using inherent knowledge, (4) irrelevant response context, (5) chatbot fictional narratives, (6) incomprehensive responses, and (7) noisy text (eg, unwanted phrases and special characters). Generated responses were evaluated based on the following 10 criteria: clarity [94,95], accuracy [89,94,96], actionability [96], relevancy [89,94], information detail [95], tailored information [89,94], comprehensiveness (questions answered fully) [95], language suitability [94], tone [95], and empathy [89]. The responses were also manually reviewed to verify whether the chatbot generates responses only using inherent

knowledge (ie, verifying whether generated content is outside the scope of retrieved documents).

Prompt Engineering

Overview

First, we created level-1 prompts (ie, PI, PX, PE, and CR) that were simple and direct instructions for generating a suitable response to a user query [97]. The responses generated using level-1 prompts were then internally evaluated to assess functional requirements. Initially, the LLM rarely generated inaccurate information based on the user experience database. After we finalized the RAG database architecture, we followed LLM prompt engineering techniques [49,97]. Prompts were iteratively refined following 10 rounds of internal evaluation to result in the final level-2 prompts (ie, PI, PX, PE, and CR) that were more specific and structured to enhance information comprehensiveness, accuracy, relevancy, transparency, and response coherence.

All prompts were checked sequentially 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 prompts for topic identification (PI-1), personalized information response generation (PI-2), and verifying relevancy (PI-3). Then, the PX and personalized emotion prompts (PE-1 and PE-2) were developed. Finally, we developed and refined the prompt for response cohesion (CR). The final prompts are shown in Table S1 in [Multimedia Appendix 1](#).

Prompt Engineering for Information Comprehensiveness

To improve information comprehensiveness and low retrieval sensitivity and precision, we refined the level-1 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 initial level-1 PI-2 prompt (Table S2 in [Multimedia Appendix 1](#)) simply instructed the LLM to provide a response to a user query using the retrieved documents; this generated responses that were not comprehensive (ie, the question was not answered fully).

Prompts for Response Accuracy

We refined our level-1 IC, CN-1, PI-2, PE-1, FV-1, FV-2, and PX prompts (Table S2 in [Multimedia Appendix 1](#)) to improve the accuracy of generated responses. These prompt refinements helped address issues related to low retrieval sensitivity and precision, contradictory information, response generation using inherent knowledge, and fictional narratives. We refined the zero-shot IC prompt to a few-shot IC prompt to improve the accuracy of subclassifying informational needs. To improve context handling, 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, FV-1, FV-2, and PX prompts to emphasize

(by using “double asterisk” in the LLM prompts) 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 to generate a response. The level-1 FV-1 and FV-2 prompts failed to identify facts that supported generated response segments. This was primarily due to a lack of clarity in defining factual consistency. This led to false positives, in which the LLM identified facts that only partially support the response segments. To prevent generating fictional real-world experiences, we refined the level-1 PX prompt, which instructed 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 avoid 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, which was identified when a response contained some inaccurate opinions retrieved from the experience database. For example, the experience database contained a user post “...The risk, even if he is undetectable, isn’t zero, but it’s pretty low...,” which contradicts the recognized medical information [98].

Prompts for Information Relevance

To strengthen information relevance, we developed the PI-3 prompt (refer to Table S1 in [Multimedia Appendix 1](#) for the final refined 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. This 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 the PI-2 prompt.

Prompts for Information Transparency

To prevent generation of fictional narratives, we refined the PX prompt by instructing the 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 (eg, “... This response is based on other people’s experience.”).

Prompts 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 (eg, * in “**Cost of PrEP in the US:**”) and [in “[PrEP provided...infection.]”) and logical reasoning, which included chatbot-provided explanations for prompt execution (eg, “Okay, here are the revised responses, aiming for smoother

flow and a more natural conversational tone: ...) in responses.

Exploratory User Evaluation

Study Design and Recruitment

An exploratory blinded-comparative online user study was conducted to evaluate the RAG chatbot’s output performance, where participants were presented with 3 sets of responses for each query: a real-world user response, a RAG-generated response, and a general LLM baseline (Gemini-2.0-flash). This study serves as an exploratory content validation and protocol design phase, establishing the foundational functional readiness of the system as well as future user studies. A total of 7 participants were recruited through personal networks and students at the authors’ institution.

Data Sampling

A batch of 25 queries and corresponding user responses were randomly sampled from the experience dataset (Reddit). For each query, three response types were compared: (1) user response: real-world user response from Reddit, (2) RAG chatbot response: responses generated by the proposed RAG chatbot, and (3) general LLM: responses from the Gemini-2.0-flash model (temp=0.65).

The evaluation was delivered via structured online questionnaires (Google Forms). Each participant evaluated

a randomly selected subset of 15 queries. To ensure internal validity, the 3 response types were shuffled and deidentified for every query, ensuring participants rated the content quality without knowledge of the response source.

Measures

Two types of questionnaires were developed for the study using Google Forms. The demographic questionnaire collected participant information on demographic and professional experience. The questionnaire consists of 6 deidentified questions asking participants about their age, gender, ethnicity, race, region, and professional background.

The chatbot response evaluation questionnaire was used to assess the quality of the responses based on 10 criteria (Table 1): information accuracy [89,94,96], information clarity [94,95], information actionability [96], information detail [95], relevance [89,94], tailored information [89,94], comprehensiveness-question answered fully [95], suitable tone [95], empathy [89], and suitable language [94]. The questionnaire contained instructions and definitions of each criterion along with the set of PrEP queries and corresponding responses. Participants were required to rate each response on a 5-point Likert scale for all criteria. Based on prior literature on expert-reviewed chatbot responses [99] suggesting good to excellent performance, we determined a response quality rating of above 4 on a 5-point scale for each of the 10 evaluation criteria.

Table 1. Evaluation criteria and descriptions provided in the chatbot response evaluation questionnaire.

Attribute and criterion	Description
Information quality	
Information accuracy [89,94,96]	Is the information provided factually correct and free from misinformation?
Information clarity [94,95]	Is the information presented in a clear, concise, and easy-to-understand manner?
Information actionability [96]	Does the response provide practical, actionable advice that the user can implement?
Information detail [95]	Does the response provide sufficient detail?
Response relevance and appropriateness	
Relevance [89,94]	Does the response directly address the user’s question and concerns?
Tailored information [89,94]	Is the information provided tailored to the specific needs and circumstances of the user?
Comprehensiveness: question answered fully [95]	Does the response fully answer the question without leaving significant gaps or requiring further clarification?
Language and tone	
Suitable tone [95]	Is the tone of the response appropriate for the context and the user’s overall emotional state?
Respectful, empathetic, and considerate manner [89]	Does the response demonstrate empathy and understanding for the user’s situation? Is the response respectful and non-judgmental?
Suitable language [94]	Is the language free from jargon or technical terms that may be unfamiliar to the user?

Statistical Analysis

To account for the nonindependence of observations where multiple ratings were provided by the same participants, a linear mixed-effects model was used using the *lme4* package in R (v4.3.1; R Core Team). With 7 participants evaluating 15 queries across 3 response types and 10 evaluation criteria, the model analyzed a total of 3150 response evaluations. Participants and queries were treated as random effects to control individual rater bias and item variability. Quantitative

analysis was conducted using R (R Core Team) and Microsoft Excel.

Phase III: Planned Prototype Enhancement and Heuristic Evaluation

Future Prototype Enhancements

To further improve the quality and diversity of data, we will expand the data sources to include 2 additional discussion boards from MedHelp and POZ community (HIV positive) to broaden the knowledge base. MedHelp and POZ databases

consist of PrEP-related conversational dialogues among community members. We will develop prompts specifically tailored for varying formats and structures [100], including conversational dialogues, question-answer pairs, and dialogues with quotes. We will update the chatbot architecture to maintain historical dialogues to reflect real-world multiturn conversations and include denial capabilities to set clear expectations for users and ensure safety compliance. We will also improve the fact validator's accuracy to reduce the impact of LLM's inherent inconsistencies when extracting facts to support generated responses. The final verification will focus on the accuracy of the responses rather than on the fact validator's performance using the retrieved documents.

Planned Heuristic Evaluation: Expert Evaluation

To assess chatbot response 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 (eg, HIV care providers), pharmacist, and/or social and community health worker. After providing online consent, participants will complete an online demographic questionnaire followed by the chatbot response evaluation questionnaire. The demographic questionnaire consists of 6 deidentified 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 1) 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.

To measure the consistency of ratings provided by the experts, we will calculate the intraclass correlation coefficient (ICC) [101] for the ratings of each criterion. We will then compare the chatbot-generated responses to expert responses by measuring Bilingual Evaluation Understudy (BLEU; IBM researchers Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu) [89] and Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence (ROUGE-L; Chin-Yew Lin) [102] scores. These scores computationally measure relevance and linguistic similarity, and are scalable for evaluating larger datasets. In addition, human emotions are complex and are influenced by individual perception; this makes generating text for personalized emotional support a highly challenging endeavor.

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

Future Large-Scale User Evaluation

To evaluate the system's real-world utility, we will conduct a large-scale user study involving 50 PrEP candidates. Participants will be recruited in coordination with HIV prevention experts and established PrEP programs, using public resources such as referral lists, health websites, and online directories. All study procedures will be conducted in

accordance with institutional review board (IRB) standards, with recruitment facilitated through IRB-approved flyers and digital outreach materials.

Eligibility criteria include that participants must be at least 18 years of age, proficient in English, and meet clinical eligibility for PrEP. Following the provision of informed online consent, participants will complete a series of comprehensive study questionnaires followed by a concluding interview. The estimated time commitment for the session is between 90 and 120 minutes.

To ensure methodological consistency across evaluation phases, this user study will use the same 10-point evaluation criteria and questionnaires used in the exploratory user evaluation. By incorporating qualitative measures (eg, semistructured interviews), we aim to gain deeper insights into user experience, perceived trust, and the effectiveness of the chatbot-provided PrEP support. This will allow us to assess how the chatbot performs in real-world scenarios beyond research settings.

Ethical Considerations

Institutional Approval

The exploratory user evaluation was reviewed and approved by the IRB at the University of North Carolina at Charlotte (protocol number IRB-26-0423). The study was determined to pose minimal risk to participants, as the data collection was limited to subjective evaluations of chatbot responses and did not collect any personally identifiable information. For the planned heuristic evaluation and subsequent large-scale user study, separate IRB applications will be submitted for approval prior to recruitment, ensuring all ethical standards are met for that specific population.

Participant Consent and Incentives

Informed consent is obtained electronically via DocuSign (DocuSign Inc Team) from all participants, including students, community members, and HIV experts, prior to study commencement. Participants are informed about the study details, potential risks, participation rights, and incentives. They are explicitly informed that their participation was voluntary and that they could withdraw at any time.

In the exploratory user evaluation, students received the opportunity to replace one class lab activity as an incentive, while other participants received no compensation; an alternative nonresearch activity was available to students to ensure the voluntary nature of participation. Participants in the expert and large-scale user study will receive a US \$25 Amazon gift card on completion of the study, subject to IRB approval.

Privacy and Data Security

To protect participant privacy, all data were or will be collected and analyzed anonymously, ensuring that no feedback could be traced back to individual experts or professionals. The dataset is filtered to pseudonymize any

PII information such as names, email addresses, contact information, and IP addresses during the RAG process.

Results

Overview

We have developed a RAG chatbot prototype as of January 2025 and have iteratively refined the chatbot design and system based on feedback from internal evaluation.

Support Classifier Performance

To assess the reliability of the RAG chatbot in identifying support needs, we validated the 3-class support classifier in categorizing queries into informational, emotional, and contextual needs.

As reported in Tables 2 and 3, our fine-tuned Gemma 2 2B-it model demonstrated strong performance in classifying

PrEP support needs when fine-tuned on the full dataset (F_1 -score=0.86) and class-balanced subset (F_1 -score=0.89). Specifically, we found that 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 contextual queries (Table 2). Based on superior performance indicated by F_1 -score metrics, we used the class-balanced Gemma 2 2B-it model during preprocessing for classifying user support needs. In the context of classifying informational and emotional support seeking, our class-unbalanced BERT model exhibits superior performance (F_1 -information score=0.92, F_1 -emotion score=0.71) compared to a previous social support seeking dataset, CHQ-SocioEmo (F_1 -information score=0.58, F_1 -emotion score=0.62). This classification of user queries allows our chatbot to identify user needs, extract context, and generate tailored and emotionally relevant responses.

Table 2. Results of an assessment of the performance of 3 classifiers, fine-tuned on the class-balanced PrEP^a dataset. These classifiers are used for classifying user input into 3 classes.

Class	BERT-base-uncased ^b		Gemma 2B-it		Gemma 2 2B-it	
	F_1 -score	Accuracy	F_1 -score	Accuracy	F_1 -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 ^c	0.89 ^c

^aPrEP: preexposure prophylaxis.
^bBERT: Bidirectional Encoder Representations from Transformers.
^cBest-performing model.

Table 3. Results of an assessment of the performance of 3 classifiers fine-tuned on the full PrEP^a dataset. These classifiers are used for classifying user inputs into 3 classes.

Class	BERT-base-uncased ^b		Gemma 2B-it		Gemma 2 2B-it	
	F_1 -score	Accuracy	F_1 -score	Accuracy	F_1 -score	Accuracy
Information	0.92	0.929	0.92	0.927	0.92	0.92
Emotion ^c	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 ^d

^aPrEP: preexposure prophylaxis.
^bBERT: Bidirectional Encoder Representations from Transformers.
^cPoor performance compared to class-balanced fine-tuned model predictions.
^dModel with best performance.

Prompt Engineering

We manually observed that prompt engineering plays a significant role in harnessing LLM capabilities to generate desired outputs. 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 [103], starting with simple and direct instructions (Table S2 in Multimedia Appendix 1) helped 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 (eg, PI-2 prompt from Table 1). Such prompt decomposition was 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 (ie, 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 LLMs indicate low prompt adherence when a prompt has multiple instructions. In such cases, the LLM often defaults to the last instruction, resulting in partial and inconsistent execution of complex logical reasoning prompts. Another challenge was the LLM's sensitivity to linguistic variability in prompts [104]. We found that paraphrasing prompts by modifying prompt structure, using synonyms, and expanding prompt context is ineffective in context management tasks (eg, identifying the need for additional context), exhibiting inconsistent response output. We found that shorter prompts with content words summarizing the main concept or idea can reduce ambiguity in LLM decisions.

A few-shot prompting approach is effective in text classification [49]. We found that few-shot prompting did not consistently perform well across classification for informational needs, emotional needs, and context. This limitation is potentially because these categories represent high-level support needs that require diverse perspectives for inferring ambiguous context and broad scope. Consequently, fine-tuning LLMs on annotated datasets produced more reliable results than few-shot prompting. Few-shot prompting is effective in simple subclassification tasks with clear and distinct criteria for narrow scope [105].

Fact Validation

For fact validation, deterministic techniques, including semantic relevance and text similarity metrics (eg, ROUGE and BLEU), often failed due to their limited ability to capture negations [106] and numerical accuracy [107]. We found that LLMs generally struggle to validate statements that are generalized or personalized and are not directly present in the knowledge base. When required facts are scattered across documents, such generalized or personalized statements need implicit validation and cross-referencing for accurate verification of information [108].

Exploratory User Study Analysis

The exploratory user study included 7 participants who completed the full evaluation protocol; the majority of participants were female (5/7, 71.4%) and aged 25-34 years (5/7, 71.4%), with an Asian (5/7, 71.4%) ethnic background. Across 15 real-world queries evaluated against 10 quality criteria, the study generated 3144 valid observations, with 6 "NA" ratings (instances where a criterion was deemed nonapplicable by the rater).

Preliminary Comparison of RAG Responses With General LLM and Real-World User Responses

The overall performance of the 3 response sets (general LLM, RAG chatbot, and user) was evaluated across 10 distinct evaluation criteria. Due to nonindependence of data points, a linear mixed model was used to account for user bias and random effects from both participants and specific questions.

Initial observations indicate variations in response quality across the 3 groups, suggesting specific areas for improvement in the RAG chatbot architecture and prompt design based on the evaluation criteria. Both AI modalities (general

LLM and RAG) had higher preference across the evaluation criteria compared to user responses. We observed that the general LLM's communication style appeared to influence how participants rated comprehensiveness, detail, and tailoredness of the information received, though both AI models showed similar performance in terms of empathy and tone.

Given the exploratory nature of this pilot evaluation, the sample size ($n=7$) was small and skewed in terms of participant demographics (eg, primarily consisting of university students), which limits the generalizability of these findings to the broader PrEP-candidate population. However, these exploratory insights into future modifications demonstrate the viability of the evaluation framework and provide a descriptive baseline for the heuristic and user study outlined in this protocol.

Discussion

Principal Findings

We describe the development of a RAG chatbot prototype for providing personalized support to fulfill informational, peer experiential expertise, and emotional needs of PrEP candidates to facilitate PrEP care. The chatbot is designed for (1) identifying support needs and personalizing responses (tailored information, clarity, and tone); (2) providing complete, relevant, and accurate HIV and PrEP information (comprehensiveness, detail, relevancy, accuracy, and language); (3) providing peer experiential expertise support (relevancy and actionability); and (4) responding with human-like emotional support (empathy).

Performance Gap of AI Modalities

Our exploratory user study results provide initial empirical support for the proposed chatbot functionality. Both the RAG chatbot and the general LLM were preferred over user responses across all 10 evaluation criteria. This suggests that AI-generated support can offer an enhanced level of perceived utility, informational depth, and structural coherence compared to traditional peer support related to PrEP.

Addressing Technical Challenges

The analysis revealed a performance gap between the two AI architectures. The general LLM performed better than the RAG chatbot in 8 of 10 categories, with the exceptions of "empathy" and "tone." While RAG techniques are intended to reduce hallucinations [48], improve content diversity [58], user privacy [53], and personalization [54,55], our results indicate that the logical reconstruction of fragmented information remains a challenge for RAG-based systems [108].

LLMs can maintain superior structural fluidity and generate well-organized information [109]. Our results indicated higher ratings for the general LLM in stylistic criteria, such as structure and communication style. In contrast, we observed that the RAG system's output was

inherently tied to multisource retrieval and the informal nature of the retrieved peer-discussion data. However, the practical advantage of our RAG approach lies in its modular architecture, which is specifically designed to facilitate control over response accuracy and prevent the generation of fictitious experiences by relying on verifiable information. Furthermore, our RAG chatbot integrates anonymized real-world peer experiential expertise and Willcox emotional vocabulary to provide a level of cultural authenticity and empathetic alignment that a general-purpose model cannot guarantee. Another advantage of RAG models is their ability to use smaller, localized language models to achieve response quality comparable to significantly larger LLMs [110,111], thereby providing opportunities to reduce health care barriers related to data privacy and operational costs.

The participant group expressed a preference for the structured and categorized communication style of the general LLM over the RAG chatbot's response format. This suggests that while RAG can be highly effective in extracting information from diverse sources to provide contextually intelligent responses [112], further refinement is needed to synthesize retrieved fragments into a clearer and more structured response format that matches the systematic response delivery of general-purpose LLMs.

User Fatigue and Rater Consistency

Our analysis revealed a residual variance of 0.662 within the linear mixed model, suggesting random noise within the rater data that warrants further investigation. We believe that this inconsistency is primarily attributable to user fatigue stemming from the study's cognitive load. Each participant was required to evaluate 15 complex query sets across 10 distinct criteria, resulting in 450 total data points per rater.

Furthermore, the high level of domain-specific expertise required to evaluate PrEP support, specifically the distinction between clinical accuracy and nuanced peer experiential support, may have been impacted by the nonexpert and nontarget user status of the participant group. As the raters were neither clinicians nor PrEP candidates, their evaluations of criteria such as "accuracy," "actionability," "clarity," "comprehensiveness," "detail," "empathetic," "language," "relevance," "tailored information," and "tone" represent perceived credibility rather than objective medical validation. For criteria related to information quality (accuracy, clarity, actionability, and detail), raters lacked the clinical training required to verify the medical correctness of PrEP information and practical guidance. Similarly, evaluating criteria on response relevance and appropriateness (relevance, tailored information, and comprehensiveness) requires an understanding of PrEP challenges such as medical mistrust, PrEP navigation, or side-effect management, commonly experienced by the target population. Furthermore, HIV experts and PrEP candidates are more familiar with PrEP terminology and community lingo, which may impact the perception of tailored information, language, and trust, as these target users can distinguish between generic AI responses and grounded, community-informed experiential support.

The low incidence of "N/A" ratings (n=6) suggests that the questionnaire structure may have lacked sufficient clarity in guiding participants to opt out when specific criteria were not relevant to a particular question sample. Improving the questionnaire structure to better highlight the specific intent of each rating and providing more robust instructions will be critical to ensure rater reliability. These exploratory observations will directly inform the next phases of our development, as detailed in the "Future Direction" section.

Implications

Overview

RAG chatbots are increasingly used to provide various support for health care use and decision-making [56,113-116]. We report the first published efforts to develop a RAG chatbot prototype for providing personalized information, peer experiential expertise, and emotional support to PrEP candidates. Our chatbot aims to enhance PrEP support by leveraging verified information, real-world experiences, and emotional considerations. It can also interpret complex context-dependent queries, and through personalized informational, peer experiential expertise, and human-like emotional support, we believe this approach has the potential to clarify misconceptions about PrEP and HIV treatment, support user confidence, and facilitate future PrEP uptake.

Data Privacy and Ethical Implications

The RAG chatbot currently leverages knowledge from Reddit, with a future plan to expand to include other online platforms. All data usage complies with the respective platforms' user agreements and policies, accessed in February 2024, and the facts dataset, collected in January 2022. Recognizing the sensitive nature of health-related discourse in these communities, our data handling process incorporates privacy safeguards; we do not attempt to identify users and protect privacy by filtering personally identifiable information (eg, name, email addresses, IP addresses, and specific unique identifiers).

The proposed RAG architecture is designed to enhance response precision; however, the deployment of personalized PrEP chatbots carries inherent ethical implications. A primary concern is the probabilistic nature of LLMs. Despite the integration of domain-specific knowledge, these models remain prone to hallucination [48,49] and may lack the deep contextual understanding required for nuanced medical inquiries. Consequently, there is a risk of users overrelying on the tool [117,118] for definitive medical advice, which could delay professional clinical consultation.

Furthermore, it is critical to distinguish between the cognitive empathy simulated by the LLM and true human-like emotional intelligence. While the system can identify and mirror user sentiments, it lacks the capacity for genuine affective responses [43]. This limitation is particularly significant in the context of HIV and PrEP, where users may present with complex psychological distress. The chatbot responses could inadvertently lead to stigma reinforcement

if the model misinterprets nuanced emotions and sensitive disclosures or provides inappropriate feedback [117,118].

To mitigate these ethical concerns, the RAG architecture implements a modular data governance strategy. The chatbot is designed to use social media data exclusively for addressing experiential queries and mirroring human-like empathetic communication styles while being guided by the Willcox Feeling Wheel to interpret user emotions. Conversely, all factual and clinical queries are strictly routed to the manually curated factual database, ensuring that health-related information is derived solely from verified public health organizations. For transparency, any user content referenced in chatbot responses is quoted anonymously to mitigate the risk of reidentification while maintaining the authenticity of the peer-to-peer perspective.

Limitations

Although recent advances in hallucination removal have improved LLM reliability, these advances do not entirely eliminate the need for validating response accuracy, especially in health care applications [92]. Due to linguistic complexity and nuances in generated responses, LLMs struggle to consistently and accurately validate factual information using only RAGs [119]. Additional methodological improvements are needed to validate broad facts outside of the database and context-dependent information.

Our system faces several inherent risks and ethical considerations. First, the complexity of human emotion makes personalized digital support a challenging endeavor, with risks of stigma reinforcement or user overreliance for clinical advice. Second, while using community-sourced data (eg, Reddit) is practical, it may introduce demographic or topical biases that limit the system's generalizability across diverse PrEP user populations. We intend to extend the datasets from MedHelp and POZ communities to mitigate these gaps.

A limitation of this study is the absence of a comprehensive quantitative analysis comparing the performance of few-shot prompting and fine-tuning for support classification. While our study provides initial qualitative insights into the effectiveness of these 2 approaches for specific tasks, these findings are exploratory. Future work should include extensive empirical comparisons to statistically generalize these results across broader datasets.

Another limitation of the current evaluation is that the user study was conducted with people who were non-target PrEP-using demographics and nonclinical experts. While this provides insight into general user perception and readability, it may not capture the technical accuracy or perceived personalization of the responses with precision. Future work will include larger-scale iterative evaluations to identify real-world failure scenarios and ensure

the model's safety profile. Furthermore, subsequent iterations will focus on demographic-aware personalization to better reflect the diverse cultural preferences, behaviors, and potential linguistic needs of the PrEP users, addressing the wider ethical implications of using sensitive, public-domain social media data for health support.

Future Direction

The current chatbot model was developed and accessed in a research setting and will be rigorously tested for safety and reliability with experts and the target population prior to real-world deployment. Building on our exploratory findings, we will conduct a larger-scale user study for gathering feedback on chatbot responses from the target population. We will also concurrently initiate a heuristic evaluation with HIV prevention experts to validate clinical safety, user trust, ethical alignment, and factual accuracy of the PrEP chatbot.

To bridge the gap between initial research product and sustainable real-world health intervention, the chatbot design will incorporate sociotechnical design considerations [120] through the Accelerated Creation-to-Sustainment (ACS) framework [121]. Through these community-representative user and expert evaluations, we will gather feedback on cultural and behavioral preferences to improve long-term adoption and user trust. This includes incorporating multi-lingual support to cater to linguistic needs. Furthermore, we will evaluate technological flexibility, including low-cost, lightweight local deployment models and multiplatform accessibility (eg, mobile and web-based interfaces), to lower entry barriers in resource-limited settings. To ensure long-term sustainability, we plan to engage with HIV health stakeholders to establish frameworks for managing scalability and routine content updates, ensuring the intervention remains clinically accurate and relevant as the HIV prevention landscape evolves.

Conclusion

This study details the implementation of a RAG chatbot prototype designed to provide personalized informational, peer experiential expertise, and emotional support to PrEP candidates. By offering personalized information and peer experience, our chatbot can provide essential information for making informed decisions about health. In addition, our chatbot's capability to provide human-like emotional support highlights its potential to address stigma related to PrEP. Although current observations are preliminary, we believe our chatbot has the potential to clarify misconceptions and provide PrEP support through these mechanisms. Future evaluation using expert and real-world feedback will validate and improve the chatbot's capability; after this, a trial to demonstrate the efficacy of the tool to increase PrEP uptake and retention will be proposed.

Acknowledgments

The authors declare the use of Google Gemini (institutional account) strictly for checking grammar and improving sentence structure (STM category 1: editing and proofreading). The tool was not used to generate intellectual content, references, or

draft manuscript text. The authors reviewed the final manuscript and are accountable for the integrity and accuracy of the manuscript content.

Funding

The authors declare no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt engineering.

[DOCX File (Microsoft Word File), 22 KB-Multimedia Appendix 1]

References

1. Kroch A, O'Byrne P, Orser L, et al. 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*. Jun 1, 2023;49(6):274-281. [doi: [10.14745/ccdr.v49i06a04](https://doi.org/10.14745/ccdr.v49i06a04)] [Medline: [38440773](https://pubmed.ncbi.nlm.nih.gov/38440773/)]
2. Bavinton BR, Grulich AE. HIV pre-exposure prophylaxis: scaling up for impact now and in the future. *Lancet Public Health*. Jul 2021;6(7):e528-e533. [doi: [10.1016/S2468-2667\(21\)00112-2](https://doi.org/10.1016/S2468-2667(21)00112-2)] [Medline: [34087117](https://pubmed.ncbi.nlm.nih.gov/34087117/)]
3. Dimitrov DT, Masse BR, Donnell D. PrEP adherence patterns strongly affect individual HIV risk and observed efficacy in randomized clinical trials. *J Acquir Immune Defic Syndr*. Aug 1, 2016;72(4):444-451. [doi: [10.1097/QAI.0000000000000993](https://doi.org/10.1097/QAI.0000000000000993)] [Medline: [26990823](https://pubmed.ncbi.nlm.nih.gov/26990823/)]
4. EHE overview. HIV.gov. URL: <https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview> [Accessed 2025-04-12]
5. Sullivan PS, DuBose SN, Castel AD, et al. Equity of PrEP uptake by race, ethnicity, sex and region in the United States in the first decade of PrEP: a population-based analysis. *Lancet Reg Health Am*. May 2024;33:100738. [doi: [10.1016/j.lana.2024.100738](https://doi.org/10.1016/j.lana.2024.100738)] [Medline: [38659491](https://pubmed.ncbi.nlm.nih.gov/38659491/)]
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*. May 2020;37(5):1778-1811. [doi: [10.1007/s12325-020-01295-0](https://doi.org/10.1007/s12325-020-01295-0)] [Medline: [32232664](https://pubmed.ncbi.nlm.nih.gov/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*. Feb 20, 2023;23(1):365. [doi: [10.1186/s12889-023-15260-3](https://doi.org/10.1186/s12889-023-15260-3)] [Medline: [36805698](https://pubmed.ncbi.nlm.nih.gov/36805698/)]
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*. 2023;5:1084657. [doi: [10.3389/frph.2023.1084657](https://doi.org/10.3389/frph.2023.1084657)] [Medline: [37152481](https://pubmed.ncbi.nlm.nih.gov/37152481/)]
9. Crooks N, Singer RB, Smith A, et al. Barriers to PrEP uptake among Black female adolescents and emerging adults. *Prev Med Rep*. Feb 2023;31:102062. [doi: [10.1016/j.pmedr.2022.102062](https://doi.org/10.1016/j.pmedr.2022.102062)] [Medline: [36467542](https://pubmed.ncbi.nlm.nih.gov/36467542/)]
10. Conner KO, McKinnon SA, Ward CJ, Reynolds CF, Brown C. Peer education as a strategy for reducing internalized stigma among depressed older adults. *Psychiatr Rehabil J*. Jun 2015;38(2):186-193. [doi: [10.1037/prj0000109](https://doi.org/10.1037/prj0000109)] [Medline: [25915057](https://pubmed.ncbi.nlm.nih.gov/25915057/)]
11. Li Y, Yan X. How could peers in online health community help improve health behavior. *Int J Environ Res Public Health*. Jan 2020;17(9):2995. [doi: [10.3390/ijerph17092995](https://doi.org/10.3390/ijerph17092995)]
12. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. *Psychol Inq*. 2000;11(4):227-268. URL: https://www.tandfonline.com/doi/abs/10.1207/s15327965pli1104_01 [Accessed 2026-08-12] [doi: [10.1207/S15327965PLI1104_01](https://doi.org/10.1207/S15327965PLI1104_01)]
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 Health*. 2019;4(3):337-356. [doi: [10.1037/sah0000148](https://doi.org/10.1037/sah0000148)]
14. Johnson J, Killelea A, Farrow K. Investing in national HIV PrEP preparedness. *N Engl J Med*. Mar 2, 2023;388(9):769-771. [doi: [10.1056/NEJMp2216100](https://doi.org/10.1056/NEJMp2216100)] [Medline: [36847476](https://pubmed.ncbi.nlm.nih.gov/36847476/)]
15. Hill M, Smith J, Elimam D, et al. 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*. 2023;18(3):e0283764. [doi: [10.1371/journal.pone.0283764](https://doi.org/10.1371/journal.pone.0283764)] [Medline: [36996143](https://pubmed.ncbi.nlm.nih.gov/36996143/)]
16. Sullivan PS, Siegler AJ. Getting pre-exposure prophylaxis (PrEP) to the people: opportunities, challenges and emerging models of PrEP implementation. *Sex Health*. Nov 2018;15(6):522-527. [doi: [10.1071/SH18103](https://doi.org/10.1071/SH18103)] [Medline: [30476461](https://pubmed.ncbi.nlm.nih.gov/30476461/)]
17. Maita KC, Maniaci MJ, Haider CR, et al. The impact of digital health solutions on bridging the health care gap in rural areas: a scoping review. *Perm J*. Sep 16, 2024;28(3):130-143. [doi: [10.7812/TPP/23.134](https://doi.org/10.7812/TPP/23.134)] [Medline: [39135461](https://pubmed.ncbi.nlm.nih.gov/39135461/)]

18. Liu A, Coleman K, Bojan K, et al. 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*. Jan 25, 2019;8(1):e10659. [doi: [10.2196/10659](https://doi.org/10.2196/10659)] [Medline: [30681964](https://pubmed.ncbi.nlm.nih.gov/30681964/)]
19. 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 Res Protoc*. May 31, 2023;12(1):e42983. [doi: [10.2196/42983](https://doi.org/10.2196/42983)]
20. 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*. Dec 12, 2022;24(12):e39816. [doi: [10.2196/39816](https://doi.org/10.2196/39816)] [Medline: [36508248](https://pubmed.ncbi.nlm.nih.gov/36508248/)]
21. Chen S, Zhang Q, Chan CK, et al. 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*. Jun 30, 2023;12:e48447. [doi: [10.2196/48447](https://doi.org/10.2196/48447)] [Medline: [37389935](https://pubmed.ncbi.nlm.nih.gov/37389935/)]
22. Liu AY, Alleyne CD, Doblecki-Lewis S, et al. 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 Form Res*. Mar 27, 2024;8(1):e54073. [doi: [10.2196/54073](https://doi.org/10.2196/54073)] [Medline: [38536232](https://pubmed.ncbi.nlm.nih.gov/38536232/)]
23. Massa P, de Souza Ferraz DA, Magno L, et al. 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*. Jun 23, 2023;25:e41881. [doi: [10.2196/41881](https://doi.org/10.2196/41881)] [Medline: [37351920](https://pubmed.ncbi.nlm.nih.gov/37351920/)]
24. 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 Res Protoc*. Aug 13, 2024;13(1):e59975. [doi: [10.2196/59975](https://doi.org/10.2196/59975)] [Medline: [39137028](https://pubmed.ncbi.nlm.nih.gov/39137028/)]
25. Muessig KE, Knudtson KA, Soni K, et al. "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. [Medline: [30123342](https://pubmed.ncbi.nlm.nih.gov/30123342/)]
26. Hightow-Weidman LB, Muessig K, Soberano Z, et al. Tough talks virtual simulation HIV disclosure intervention for young men who have sex with men: development and usability testing. *JMIR Form Res*. Sep 8, 2022;6(9):e38354. [doi: [10.2196/38354](https://doi.org/10.2196/38354)] [Medline: [36074551](https://pubmed.ncbi.nlm.nih.gov/36074551/)]
27. Galea JT, Vasquez DH, Rupani N, et al. 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*. May 7, 2024;13:e55559. [doi: [10.2196/55559](https://doi.org/10.2196/55559)] [Medline: [38713501](https://pubmed.ncbi.nlm.nih.gov/38713501/)]
28. 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*. Nov 29, 2021;7(4):e27850. [doi: [10.2196/27850](https://doi.org/10.2196/27850)] [Medline: [34847056](https://pubmed.ncbi.nlm.nih.gov/34847056/)]
29. Ardiana DPY, Joni IDMAB, Udayana IPAED. Mobile based chatbot application for HIV/AIDS counseling using artificial intelligence markup language approach. *J Phys: Conf Ser*. Feb 1, 2020;1469(1):012041. [doi: [10.1088/1742-6596/1469/1/012041](https://doi.org/10.1088/1742-6596/1469/1/012041)]
30. Moreno JC, Sánchez-Anguix V, Alberola JM, Julián V, Botti V. An intelligent conversational agent for educating the general public about HIV. *Neurocomputing*. Jan 2024;563:126902. [doi: [10.1016/j.neucom.2023.126902](https://doi.org/10.1016/j.neucom.2023.126902)]
31. 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 Uhealth*. Apr 3, 2020;8(4):e15806. [doi: [10.2196/15806](https://doi.org/10.2196/15806)] [Medline: [32242820](https://pubmed.ncbi.nlm.nih.gov/32242820/)]
32. Wang T, Oliver D, Msosa Y, et al. Implementation of a real-time psychosis risk detection and alerting system based on electronic health records using CogStack. *J Vis Exp*. May 15, 2020;PMID(159). [doi: [10.3791/60794](https://doi.org/10.3791/60794)] [Medline: [32478737](https://pubmed.ncbi.nlm.nih.gov/32478737/)]
33. 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 Ment Health*. Dec 13, 2018;5(4):e64. [doi: [10.2196/mental.9782](https://doi.org/10.2196/mental.9782)] [Medline: [30545815](https://pubmed.ncbi.nlm.nih.gov/30545815/)]
34. 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 Interv*. Mar 2022;27:100495. [doi: [10.1016/j.invent.2022.100495](https://doi.org/10.1016/j.invent.2022.100495)] [Medline: [35059305](https://pubmed.ncbi.nlm.nih.gov/35059305/)]
35. 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. *Int J Med Inform*. Jun 2021;150:104440. [doi: [10.1016/j.ijmedinf.2021.104440](https://doi.org/10.1016/j.ijmedinf.2021.104440)] [Medline: [33799055](https://pubmed.ncbi.nlm.nih.gov/33799055/)]

36. Hamid MS, Valicevic A, Brennen B, Niziol LM, Stein JD, Newman-Casey PA. Text parsing-based identification of patients with poor glaucoma medication adherence in the electronic health record. *Am J Ophthalmol*. Feb 2021;222:54-59. [doi: [10.1016/j.ajo.2020.09.008](https://doi.org/10.1016/j.ajo.2020.09.008)] [Medline: [32926847](https://pubmed.ncbi.nlm.nih.gov/32926847/)]
37. De Nieva JO, Joaquin JA, Tan CB, Marc Te RK, Ong E. Investigating students' use of a mental health chatbot to alleviate academic stress. Presented at: CHIUXID '20: 6th International ACM In-Cooperation HCI and UX Conference; Oct 21-23, 2020; Jakarta & Bandung Indonesia. URL: <https://dl.acm.org/doi/proceedings/10.1145/3431656> [Accessed 2026-08-12]
38. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513. [doi: [10.1136/jamia.2009.001560](https://doi.org/10.1136/jamia.2009.001560)] [Medline: [20819853](https://pubmed.ncbi.nlm.nih.gov/20819853/)]
39. Ali MR, Razavi SZ, Langevin R, et al. A virtual conversational agent for teens with autism spectrum disorder: experimental results and design lessons. Presented at: IVA '20: 20th ACM International Conference on Intelligent Virtual Agents; Oct 19-23, 2020. URL: <https://dl.acm.org/doi/10.1145/3383652.3423900> [Accessed 2026-09-19] [doi: [10.1145/3383652.3423900](https://doi.org/10.1145/3383652.3423900)]
40. Peng ML, Wickersham JA, Altice FL, et al. Formative evaluation of the acceptance of HIV prevention artificial intelligence chatbots by men who have sex with men in Malaysia: focus group study. *JMIR Form Res*. Oct 6, 2022;6(10):e42055. [doi: [10.2196/42055](https://doi.org/10.2196/42055)] [Medline: [36201390](https://pubmed.ncbi.nlm.nih.gov/36201390/)]
41. Nadarzynski T, Puentes V, Pawlak I, et al. Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: a qualitative analysis. *Sex Health*. Nov 2021;18(5):385-393. [doi: [10.1071/SH21123](https://doi.org/10.1071/SH21123)] [Medline: [34782055](https://pubmed.ncbi.nlm.nih.gov/34782055/)]
42. Luo M, Warren CJ, Cheng L, Abdul-Muhsin HM, Banerjee I. Assessing empathy in large language models with real-world physician-patient interactions. Presented at: 2024 IEEE International Conference on Big Data (BigData); Dec 15-18, 2024:6510-6519; Washington, DC, USA. 2024. URL: <https://ieeexplore.ieee.org/document/10825307> [Accessed 2026-09-19] [doi: [10.1109/BigData62323.2024.10825307](https://doi.org/10.1109/BigData62323.2024.10825307)]
43. Vzorin GD, Bukinich AM, Sedykh AV, Vetrova II, Sergienko EA. The emotional intelligence of the GPT-4 large language model. *Psychol Russ*. 2024;17(2):85-99. [doi: [10.11621/pir.2024.0206](https://doi.org/10.11621/pir.2024.0206)] [Medline: [39552777](https://pubmed.ncbi.nlm.nih.gov/39552777/)]
44. Sorin V, Brin D, Barash Y, et al. Large language models and empath: systematic review. *J Med Internet Res*. Dec 2024;26(1):e52597. [doi: [10.2196/52597](https://doi.org/10.2196/52597)] [Medline: [39661968](https://pubmed.ncbi.nlm.nih.gov/39661968/)]
45. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI Soc*. 2022;37(4):1353-1359. [doi: [10.1007/s00146-021-01230-z](https://doi.org/10.1007/s00146-021-01230-z)] [Medline: [34054228](https://pubmed.ncbi.nlm.nih.gov/34054228/)]
46. Yu HQ, McGuinness S. An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. *J Med Artif Intell*. Jun 4, 2024;7:16-16. [doi: [10.21037/jmai-23-136](https://doi.org/10.21037/jmai-23-136)]
47. Ye H, Liu T, Zhang A, Hua W, Jia W. Cognitive mirage: a review of hallucinations in large language models. *arXiv*. Preprint posted online on Sep 13, 2023. [doi: [10.48550/arXiv.2309.06794](https://doi.org/10.48550/arXiv.2309.06794)]
48. Gilson A, Ai X, Arunachalam T, et al. Enhancing large language models with domain-specific retrieval augment generation: a case study on long-form consumer health question answering in ophthalmology. *arXiv*. Preprint posted online on Sep 20, 2024. [doi: [10.48550/arXiv.2409.13902](https://doi.org/10.48550/arXiv.2409.13902)]
49. Wang J, Shi E, Yu S, et al. Prompt engineering for healthcare: methodologies and applications. *Meta-Radiology*. Mar 2026;4(1):100190. [doi: [10.1016/j.metrad.2025.100190](https://doi.org/10.1016/j.metrad.2025.100190)]
50. 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 Netw Open*. Aug 1, 2024;7(8):e2425981. [doi: [10.1001/jamanetworkopen.2024.25981](https://doi.org/10.1001/jamanetworkopen.2024.25981)] [Medline: [39136946](https://pubmed.ncbi.nlm.nih.gov/39136946/)]
51. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. Jan 11, 2024;7(1):6. [doi: [10.1038/s41746-023-00970-0](https://doi.org/10.1038/s41746-023-00970-0)] [Medline: [38200151](https://pubmed.ncbi.nlm.nih.gov/38200151/)]
52. Yu H, Guo P, Sano A. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. Presented at: Proceedings of the 3rd Machine Learning for Health Symposium, PMLR. 650-663. 2023.
53. Zeng S, Zhang J, He P, et al. Mitigating the privacy issues in retrieval-augmented generation (RAG) via pure synthetic data. Presented at: 2025 Conference on Empirical Methods in Natural Language Processing; Nov 4-7, 2025; Suzhou, China. URL: <https://aclanthology.org/2025.emnlp-main> [Accessed 2026-08-12] [doi: [10.18653/v1/2025.emnlp-main.1247](https://doi.org/10.18653/v1/2025.emnlp-main.1247)]
54. Salemi A, Mysore S, Bendersky M, Zamani H. LaMP: when large language models meet personalization. Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Aug 11-16, 2024; Bangkok, Thailand. URL: <https://aclanthology.org/2024.acl-long.399/> [Accessed 2026-09-19] [doi: [10.18653/v1/2024.acl-long.399](https://doi.org/10.18653/v1/2024.acl-long.399)]
55. Chen J, Liu Z, Huang X, et al. When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web*. Jul 2024;27(4):42. [doi: [10.1007/s11280-024-01276-1](https://doi.org/10.1007/s11280-024-01276-1)]

56. 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. Presented at: BCB'23: 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; Sep 3-6, 2023:1-10; Houston, TX. URL: <https://dl.acm.org/doi/proceedings/10.1145/3584371> [Accessed 2026-09-19] [doi: [10.1145/3584371.3612956](https://doi.org/10.1145/3584371.3612956)]
57. Ramjee P, Sachdeva B, Golechha S, et al. CataractBot: an LLM-powered expert-in-the-loop chatbot for cataract patients. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. Jun 9, 2025;9(2):1-31. [doi: [10.1145/3729479](https://doi.org/10.1145/3729479)]
58. 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*. Jan 2025;2(1):AIra2400380. [doi: [10.1056/AIra2400380](https://doi.org/10.1056/AIra2400380)]
59. Zhou Q, Liu C, Duan Y, et al. GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. *Front Med (Lausanne)*. May 22, 2024;11:1392555. [doi: [10.3389/fmed.2024.1392555](https://doi.org/10.3389/fmed.2024.1392555)] [Medline: [38841582](https://pubmed.ncbi.nlm.nih.gov/38841582/)]
60. Gordillo V, Fekete E, Platteau T, et al. Emotional support and gender in people living with HIV: effects on psychological well-being. *J Behav Med*. Dec 2009;32(6):523-531. [doi: [10.1007/s10865-009-9222-7](https://doi.org/10.1007/s10865-009-9222-7)] [Medline: [19543823](https://pubmed.ncbi.nlm.nih.gov/19543823/)]
61. 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*. Oct 2024;11(5):2649-2662. [doi: [10.1007/s40615-023-01729-9](https://doi.org/10.1007/s40615-023-01729-9)]
62. Duthely LM, Sanchez-Covarrubias AP, Brown MR, et al. Pills, PrEP, and Pals: adherence, stigma, resilience, faith and the need to connect among minority women with HIV/AIDS in a US HIV epicenter. *Front Public Health*. 2021;9:667331. [doi: [10.3389/fpubh.2021.667331](https://doi.org/10.3389/fpubh.2021.667331)] [Medline: [34235129](https://pubmed.ncbi.nlm.nih.gov/34235129/)]
63. Hartzler A, Pratt W. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. *J Med Internet Res*. Aug 16, 2011;13(3):e62. [doi: [10.2196/jmir.1728](https://doi.org/10.2196/jmir.1728)] [Medline: [21846635](https://pubmed.ncbi.nlm.nih.gov/21846635/)]
64. Dennis CL. Peer support within a health care context: a concept analysis. *Int J Nurs Stud*. Mar 2003;40(3):321-332. [doi: [10.1016/s0020-7489\(02\)00092-5](https://doi.org/10.1016/s0020-7489(02)00092-5)] [Medline: [12605954](https://pubmed.ncbi.nlm.nih.gov/12605954/)]
65. Wang YC, Kraut R, Levine JM. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. Presented at: CSCW'12: Proceedings of the ACM Conference on Computer Supported Cooperative Work; Feb 11-15, 2012:833-842; Seattle, WA. URL: <https://dl.acm.org/doi/10.1145/2145204.2145329> [Accessed 2026-09-19] [doi: [10.1145/2145204.2145329](https://doi.org/10.1145/2145204.2145329)]
66. Castro EM, Van Regenmortel T, Sermeus W, Vanhaecht K. Patients' experiential knowledge and expertise in health care: a hybrid concept analysis. *Soc Theory Health*. Sep 2019;17(3):307-330. [doi: [10.1057/s41285-018-0081-6](https://doi.org/10.1057/s41285-018-0081-6)]
67. Burleson BR. Emotional support skills. In: Greene JO, Burleson BR, editors. *Handbook of Communication and Social Interaction Skills* Routledge. Lawrence Erlbaum Associates, Inc; 2003. [doi: [10.4324/9781410607133-22](https://doi.org/10.4324/9781410607133-22)]
68. 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. *Am Sci*. 2001;89(4):344-350. [doi: [10.1511/2001.28.344](https://doi.org/10.1511/2001.28.344)]
69. Willcox G. The feeling wheel: a tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Trans Anal J*. Oct 1, 1982;12(4):274-276. [doi: [10.1177/036215378201200411](https://doi.org/10.1177/036215378201200411)]
70. 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*. Dec 12, 2017;6(2):e24. [doi: [10.2196/ijmr.7000](https://doi.org/10.2196/ijmr.7000)] [Medline: [29233802](https://pubmed.ncbi.nlm.nih.gov/29233802/)]
71. Aliannejadi M, Chakraborty M, Rissola EA, Crestani F. Harnessing evolution of multi-turn conversations for effective answer retrieval. Presented at: CHIR '20: Conference on Human Information Interaction and Retrieval; Aug 14-18, 2020:33-42; Vancouver, BC, Canada. URL: <https://dl.acm.org/doi/proceedings/10.1145/3343413> [Accessed 2026-08-12] [doi: [10.1145/3343413.3377968](https://doi.org/10.1145/3343413.3377968)]
72. Strine TW, Chapman DP, Balluz L, Mokdad AH. Health-related quality of life and health behaviors by social and emotional support. *Soc Psychiat Epidemiol*. Feb 2008;43(2):151-159. [doi: [10.1007/s00127-007-0277-x](https://doi.org/10.1007/s00127-007-0277-x)]
73. Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. *J Med Internet Res*. Nov 8, 2024;26:e58041. [doi: [10.2196/58041](https://doi.org/10.2196/58041)] [Medline: [39046096](https://pubmed.ncbi.nlm.nih.gov/39046096/)]
74. Cutrona CE, Suhr JA. Controllability of stressful events and satisfaction with spouse support behaviors. *Communic Res*. Apr 1992;19(2):154-174. [doi: [10.1177/009365092019002002](https://doi.org/10.1177/009365092019002002)]
75. Suhr JA, Cutrona CE, Krebs KK, Jensen SL. The social support behavior code (SSBC). In: *Couple Observational Coding Systems*. Routledge; 2004. URL: <https://www.taylorfrancis.com/chapters/edit/10.4324/9781410610843-24/social-support-behavior-code-ssbc-julie-suhr-carolyn-cutrona-krista-krebs-sandra-jensen> [Accessed 2026-08-10]
76. Rousseau E, Katz AWK, O'Rourke S, et al. Adolescent girls and young women's PrEP-user journey during an implementation science study in South Africa and Kenya. *PLoS One*. 2021;16(10):e0258542. [doi: [10.1371/journal.pone.0258542](https://doi.org/10.1371/journal.pone.0258542)] [Medline: [34648589](https://pubmed.ncbi.nlm.nih.gov/34648589/)]

77. Frohmann M, Sterner I, Vulić I, Minixhofer B, Schedl M. Segment any text: a universal approach for robust, efficient and adaptable sentence segmentation. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, Florida, USA. URL: <https://aclanthology.org/2024.emnlp-main.665/> [Accessed 2026-06-19] [doi: [10.18653/v1/2024.emnlp-main.665](https://doi.org/10.18653/v1/2024.emnlp-main.665)]
78. Team G, Riviere M, Pathak S, et al. Gemma 2: improving open language models at a practical size. arXiv. Preprint posted online on Oct 2, 2024. [doi: [10.48550/arXiv.2408.00118](https://doi.org/10.48550/arXiv.2408.00118)]
79. Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: a guide to parameter-efficient fine-tuning. arXiv. Preprint posted online on Mar 28, 2023. [doi: [10.48550/arXiv.2303.15647](https://doi.org/10.48550/arXiv.2303.15647)]
80. Malladi S, Gao T, Nichani E, et al. Fine-tuning language models with just forward passes. Presented at: The 37th Annual Conference on Neural Information Processing Systems (NeurIPS 2023); Dec 10-16, 2023; New Orleans, Louisiana, USA. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/a627810151be4d13f907ac898ff7e948-Paper-Conference.pdf [Accessed 2026-09-19] [doi: [10.52202/075280-2308](https://doi.org/10.52202/075280-2308)]
81. Maatouk A, Ampudia KC, Ying R, Tassiulas L. Tele-LLMs: a series of specialized large language models for telecommunications. arXiv. Preprint posted online on Sep 9, 2024. [Accessed 2026-09-21] [doi: [10.48550/arXiv.2409.05314](https://doi.org/10.48550/arXiv.2409.05314)]
82. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019; Hong Kong, China. URL: <https://www.aclweb.org/anthology/D19-1> [Accessed 2026-08-12] [doi: [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410)]
83. 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. Presented at: Findings of the Association for Computational Linguistics; Nov 12-16, 2024:7393-7428; Miami, Florida, USA. URL: <https://aclanthology.org/2024.findings-emnlp.436/> [Accessed 2026-09-19] [doi: [10.18653/v1/2024.findings-emnlp.436](https://doi.org/10.18653/v1/2024.findings-emnlp.436)]
84. Tamine L, Goeuriot L. Semantic information retrieval on medical texts. *ACM Comput Surv*. Sep 30, 2022;54(7):1-38. [doi: [10.1145/3462476](https://doi.org/10.1145/3462476)]
85. Katsis Y, Rosenthal S, Fadnis K, et al. mt RAG: a multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Trans Assoc Comput Linguist*. Jul 29, 2025;13:784-808. [doi: [10.1162/TACL.a.19](https://doi.org/10.1162/TACL.a.19)]
86. Radlinski F, Craswell N. Comparing the sensitivity of information retrieval metrics. Presented at: The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10); Jul 19-23, 2010:667-674; Geneva, Switzerland. [doi: [10.1145/1835449.1835560](https://doi.org/10.1145/1835449.1835560)]
87. Park J, Saha S, Han D, et al. Emotional communication in HIV care: an observational study of patients' expressed emotions and clinician response. *AIDS Behav*. Oct 2019;23(10):2816-2828. [doi: [10.1007/s10461-019-02466-z](https://doi.org/10.1007/s10461-019-02466-z)] [Medline: [30895426](https://pubmed.ncbi.nlm.nih.gov/30895426/)]
88. Vamsi/t5_paraphrase_paws. Hugging Face. URL: https://huggingface.co/Vamsi/T5_Paraphrase_Paws [Accessed 2025-03-09]
89. Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digit Med*. Mar 29, 2024;7(1):82. [doi: [10.1038/s41746-024-01074-z](https://doi.org/10.1038/s41746-024-01074-z)] [Medline: [38553625](https://pubmed.ncbi.nlm.nih.gov/38553625/)]
90. Huang J, Lam MH, Li EJ, et al. Emotionally numb or empathetic? Evaluating how LLMs feel using EmotionBench. arXiv. Preprint posted online on Nov 16, 2023. [Accessed 2026-09-21] [doi: [10.48550/arXiv.2308.03656](https://doi.org/10.48550/arXiv.2308.03656)]
91. Wang Y, Wang M, Manzoor MA, et al. Factuality of large language models: a survey. Presented at: 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, FL. URL: <https://aclanthology.org/2024.emnlp-main> [Accessed 2026-08-12] [doi: [10.18653/v1/2024.emnlp-main.1088](https://doi.org/10.18653/v1/2024.emnlp-main.1088)]
92. 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*. Jul 1, 2023:2023.04.18.23288752. [doi: [10.1101/2023.04.18.23288752](https://doi.org/10.1101/2023.04.18.23288752)] [Medline: [37398329](https://pubmed.ncbi.nlm.nih.gov/37398329/)]
93. Hewett TT. The role of iterative evaluation in designing systems for usability. In: *Proceedings of the Second Conference of the British Computer Society, Human Computer Interaction Specialist Group on People and Computers: Designing for Usability*. Cambridge University Press; 1986:196-214. [doi: [10.5555/17324.24085](https://doi.org/10.5555/17324.24085)]
94. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol*. Jun 30, 2024;15(3):1-45. [doi: [10.1145/3641289](https://doi.org/10.1145/3641289)]
95. van der Lee C, Gatt A, van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. Presented at: 12th International Conference on Natural Language Generation; Oct 29 to Nov 1, 2019; Tokyo, Japan. URL: <https://aclanthology.org/W19-86> [Accessed 2026-08-12] [doi: [10.18653/v1/W19-8643](https://doi.org/10.18653/v1/W19-8643)]

96. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncol.* Oct 1, 2023;9(10):1437-1440. [doi: [10.1001/jamaoncol.2023.2947](https://doi.org/10.1001/jamaoncol.2023.2947)] [Medline: [37615960](https://pubmed.ncbi.nlm.nih.gov/37615960/)]
97. Heston T, Khun C. Prompt engineering in medical education. *Int Med Educ.* 2023;2(3):198-205. [doi: [10.3390/ime2030019](https://doi.org/10.3390/ime2030019)]
98. Cohen MS, Chen YQ, McCauley M, et al. Antiretroviral therapy for the prevention of HIV-1 transmission. *N Engl J Med.* Sep 1, 2016;375(9):830-839. [doi: [10.1056/NEJMoa1600693](https://doi.org/10.1056/NEJMoa1600693)] [Medline: [27424812](https://pubmed.ncbi.nlm.nih.gov/27424812/)]
99. Ayers JW, Poliak A, Dredze M. Responses to patient questions posted to a public social media forum. *JAMA Intern Med.* Apr 28, 2023;71(183):589-596. [doi: [10.1001/jamainternmed.2023.1838](https://doi.org/10.1001/jamainternmed.2023.1838)]
100. Zhao S, Huang Y, Song J, Wang Z, Wan C, Ma L. Understanding the fundamental design decisions of retrieval-augmented generation systems. *arXiv.* Preprint posted online on Nov 29, 2024. [doi: [10.48550/arXiv.2411.19463](https://doi.org/10.48550/arXiv.2411.19463)]
101. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* Jun 2016;15(2):155-163. [doi: [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012)] [Medline: [27330520](https://pubmed.ncbi.nlm.nih.gov/27330520/)]
102. Lin CY. ROUGE: a package for automatic evaluation of summaries. Presented at: Text Summarization Branches Out (A Post-Conference Workshop of ACL 2004); Jul 25, 2004:74-81; Barcelona, Spain. URL: <https://aclanthology.org/W04-1013/> [Accessed 2025-03-09]
103. Meincke L, Mollick ER, Mollick L, Shapiro D. Prompting science report 1: prompt engineering is complicated and contingent. *Social Science Research Network*; 2025. URL: <https://papers.ssrn.com/abstract=5165270> [Accessed 2026-08-10] [doi: [10.2139/ssrn.5165270](https://doi.org/10.2139/ssrn.5165270)]
104. Leidinger A, van Rooij R, Shutova E. The language of prompting: what linguistic properties make a prompt successful? Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; Dec 6-10, 2023:9210-9232; Singapore. URL: <https://aclanthology.org/2023.findings-emnlp.618.pdf> [Accessed 2026-09-20] [doi: [10.18653/v1/2023.findings-emnlp.618](https://doi.org/10.18653/v1/2023.findings-emnlp.618)]
105. Lamichhane B. Evaluation of chatgpt for NLP-based mental health applications. *arXiv.* Preprint posted online on Mar 28, 2023. [doi: [10.48550/arXiv.2303.15727](https://doi.org/10.48550/arXiv.2303.15727)]
106. Tay W, Joshi A, Zhang X, Karimi S, Wan S. Red-faced ROUGH: examining the suitability of ROUGE for opinion summary evaluation. In: Mistica M, Piccardi M, MacKinlay A, editors. Presented at: The 17th Annual Workshop of the Australasian Language Technology Association (ALTA 2019); Dec 4-6, 2019; Sydney, Australia. URL: <https://aclanthology.org/U19-1008/> [Accessed 2026-09-20]
107. Wallace E, Wang Y, Li S, Singh S, Gardner M. Do NLP models know numbers? probing numeracy in embeddings. Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019; Hong Kong, China. URL: <https://www.aclweb.org/anthology/D19-1> [Accessed 2026-08-12] [doi: [10.18653/v1/D19-1534](https://doi.org/10.18653/v1/D19-1534)]
108. Cheng F, Li H, Liu F, van Rooij R, Zhang K, Lin Z. Empowering LLMs with logical reasoning: a comprehensive survey. Presented at: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence; Aug 16-22, 2025; Montreal, Canada. URL: <https://www.ijcai.org/proceedings/2025> [Accessed 2026-09-20] [doi: [10.24963/ijcai.2025/1155](https://doi.org/10.24963/ijcai.2025/1155)]
109. Zamaraeva O, Flickinger D, Bond F, Gómez-Rodríguez C. Comparing LLM-generated and human-authored news text using formal syntactic theory. Presented at: 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 27 to Aug 1, 2025; Vienna, Austria. URL: <https://aclanthology.org/2025.acl-long> [Accessed 2026-08-12] [doi: [10.18653/v1/2025.acl-long.443](https://doi.org/10.18653/v1/2025.acl-long.443)]
110. Liu S, Yu Z, Huang F, Bulbulia Y, Bergen A, Liut M. Can small language models with retrieval-augmented generation replace large language models when learning computer science? Presented at: 2024 on Innovation and Technology in Computer Science Education V 1 (ITiCSE 2024); Jul 8-10, 2024:388-393; Milan, Italy. URL: <https://dl.acm.org/doi/proceedings/10.1145/3649217> [Accessed 2026-08-12] [doi: [10.1145/3649217.3653554](https://doi.org/10.1145/3649217.3653554)]
111. Shandilya B, Palmer A. Boosting the capabilities of compact models in low-data contexts with large language models and retrieval-augmented generation. Presented at: 31st International Conference on Computational Linguistics; Jan 21-24, 2025; Abu Dhabi, UAE. URL: <https://aclanthology.org/2025.coling-main.499/> [Accessed 2026-09-06]
112. Lahiri AK, Hu QV. AlzheimerRAG: multimodal retrieval-augmented generation for clinical use cases. *Mach Learn Knowl Extr.* Aug 27, 2025;7(3):89. [doi: [10.3390/make7030089](https://doi.org/10.3390/make7030089)]
113. Zhou S, Luo X, Chen C, et al. The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries. *Int J Surg.* Oct 1, 2024;110(10):6509-6517. [doi: [10.1097/JS9.0000000000001850](https://doi.org/10.1097/JS9.0000000000001850)] [Medline: [38935100](https://pubmed.ncbi.nlm.nih.gov/38935100/)]
114. Arasteh ST, Lotfinia M, Bressemer K, et al. RadioRAG: factual large language models for enhanced diagnostics in radiology using online retrieval augmented generation. *Radiol Artif Intell.* 2025. [doi: [10.1148/ryai.240476](https://doi.org/10.1148/ryai.240476)]
115. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl.* Jan 2023;82(3):3713-3744. [doi: [10.1007/s11042-022-13428-4s](https://doi.org/10.1007/s11042-022-13428-4s)]

116. 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. Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 15-19, 2024:1-4; Orlando, FL, USA. URL: <https://ieeexplore.ieee.org/document/10781509> [Accessed 2026-09-20] [doi: [10.1109/EMBC53108.2024.10781509](https://doi.org/10.1109/EMBC53108.2024.10781509)]
117. Aghaziarati A, Rahimi H. The future of digital assistants: human dependence and behavioral change. J Foresight Health Gov. 2025;2(1):52-61. URL: <https://journalfhg.com/index.php/jfph/article/view/6> [Accessed 2026-08-10]
118. Hipgrave L, Goldie J, Dennis S, Coleman A. Balancing risks and benefits: clinicians' perspectives on the use of generative AI chatbots in mental healthcare. Front Digit Health. 2025;7:1606291. [doi: [10.3389/fdgth.2025.1606291](https://doi.org/10.3389/fdgth.2025.1606291)] [Medline: [40510413](https://pubmed.ncbi.nlm.nih.gov/40510413/)]
119. Mansurova A, Mansurova A, Nugumanova A. QA-RAG: exploring LLM reliance on external knowledge. Big Data Cogn Comput. Sep 2024;8(9):115. [doi: [10.3390/bdcc8090115](https://doi.org/10.3390/bdcc8090115)]
120. Li DH, Brown CH, Gallo C, et al. Design considerations for implementing eHealth behavioral interventions for HIV prevention in evolving sociotechnical landscapes. Curr HIV/AIDS Rep. Aug 2019;16(4):335-348. [doi: [10.1007/s11904-019-00455-4](https://doi.org/10.1007/s11904-019-00455-4)] [Medline: [31250195](https://pubmed.ncbi.nlm.nih.gov/31250195/)]
121. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res. May 2017;19(5):e153. [doi: [10.2196/jmir.7725](https://doi.org/10.2196/jmir.7725)] [Medline: [28490417](https://pubmed.ncbi.nlm.nih.gov/28490417/)]

Abbreviations

ACS: Accelerated Creation-to-Sustainment
BERT: Bidirectional Encoder Representations from Transformers
BLEU: Bilingual Evaluation Understudy
CN-1: context handling prompt 1
CN-2: context handling prompt 2
CR: cohesive response
EHE: Ending the HIV Epidemic
IC : information subclassification
ICC: intraclass correlation coefficient
IRB: institutional review board
LLM: large language model
LoRA: Low-Rank Adaptation
PE: personalized emotion
PEFT: Parameter-Efficient Fine-Tuning
PI: personalized information
PrEP: preexposure prophylaxis
PX: personalized experience
RAG: retrieval-augmented generation
ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence
SAT: Segment Any Text
SBERT: Sentence-Bidirectional Encoder Representations from Transformers
SSBC: Social Support Behavior Code

Edited by Javad Sarvestan; peer-reviewed by Maria Chatzimina, Sook Ha; submitted 01 Jul.2025; final revised version received 02 Jun.2026; accepted 03 Jun.2026; published 01 Oct.2026

Please cite as:

Sayed F, Park A, Sullivan PS, Ge Y

A Conversational Agent for Providing Personalized Preexposure Prophylaxis (PrEP) Support: Protocol for Chatbot Implementation and Evaluation

JMIR Res Protoc 2026;15:e79966

URL: <https://www.researchprotocols.org/2026/1/e79966>

doi: [10.2196/79966](https://doi.org/10.2196/79966)

© Fatima Sayed, Albert Park, Patrick Sean Sullivan, Yaorong Ge. Originally published in JMIR Research Protocols (<https://www.researchprotocols.org>), 01.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Research Protocols, is properly cited. The complete

bibliographic information, a link to the original publication on <https://www.researchprotocols.org>, as well as this copyright and license information must be included.