Titre: Title:	Participatory Design and Development of an Artificial Intelligence-Based Personalized Chatbot for HIV Self-Management
Auteur: Author:	Yuanchao Ma
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ma, Y. (2025). Participatory Design and Development of an Artificial Intelligence-Based Personalized Chatbot for HIV Self-Management [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/67794/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/67794/
Directeurs de recherche: Advisors:	Sofiane Achiche, & Bertrand Lebouché
Programme: Program:	Génie biomédical

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Participatory Design and Development of an Artificial Intelligence-based Personalized Chatbot for HIV Self-Management

YUANCHAO MA

Institut de génie biomédical

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie biomédical

Août 2025

POLYTECHNIQUE MONTRÉAL

affiliée à l’Université de Montréal

Cette thèse intitulée :

Participatory Design and Development of an Artificial Intelligence-based Personalized Chatbot for HIV Self-Management

présentée par **Yuanchao MA**

en vue de l’obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d’examen constitué de :

Abolfazl MOHEBBI, président

Sofiane ACHICHE, membre et directeur de recherche

Bertrand LEBOUCHÉ, membre et codirecteur de recherche

Jinghui CHENG, membre

Leo Anthony CELI, membre externe

DEDICATION

To my parents, for your endless love and support.

To my love, Yichun, for always standing by me with unwavering encouragement and trust.

To myself, for your perseverance, resilience, and courage.

游戏才刚刚开始。

The adventure has only just begun.

L'aventure ne fait que commencer.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to both my thesis supervisors, collaborators, and mentors, Professor Sofiane Achiche and Doctor Bertrand Lebouché. Thank you for the invaluable opportunities you have offered me, your encouragement during periods of doubt, your guidance through challenges, and your steadfast support during critical moments. I am especially thankful for the opportunity to co-lead the MARVIN chatbot project alongside you over the past five years – a privilege that has profoundly shaped my growth as both a researcher and a leader. You have been true mentors throughout my professional journey. I feel incredibly fortunate to have had the two best supervisors one could ask for, and I sincerely hope I have lived up to your expectations. I am excited about what we will continue to build together.

I am deeply grateful to the MARVIN co-construction committee, who have supported both the project and me from the very beginning. This includes Benoit Lemire, Lina Del Balso, Nathalie Paisible, and all the patient partners I've had the privilege to work with. Your insights and feedback have enabled the MARVIN project to evolve, flourish, and become the success it is today. I would also like to acknowledge the entire MARVIN team, past and present, all 68 members in total. It has been a true privilege to collaborate with each of you. Thank you for your creativity, energy, and unwavering commitment. You brought this project to life and made this journey dynamic and inspiring.

My sincere thanks to Bertrand's team (I'm still waiting on our official team name!), especially Kim Engler, David Lessard, and Serge Vicente. Your thoughtful guidance, especially regarding methodology and strategic planning, your meticulous attention to detail, and your strong commitment to research integrity have been invaluable to my development. I am also grateful to all my colleagues from the COSIM lab at Polytechnique Montréal, who have accompanied me since the early days of graduate school. I truly appreciate the camaraderie and many joyful moments we've shared along the way.

Warm thanks go to Doctor Esli Osmanlliu. It was a pleasure meeting you three years ago. Your many insights, thoughtful suggestions, and our many engaging discussions have been a great source of inspiration and encouragement. I'm eager to see where our future collaborations will take us.

To my parents – thank you, from beginning to end, for your love, trust, and unconditional support. I am here because of everything you have given me, and everything you've stood by without asking for anything in return.

And last, but certainly not least, to my better half, Yichun – I'm endlessly grateful for your love, companionship, encouragement, and belief in me. With you, everything becomes clearer. I'm excited for all that lies ahead – the growth, the challenges, and the path we'll keep walking together.

And finally, I would like to thank Professor Abolfazl Mohebbi for serving as president of this thesis committee, Professor Jinghui Cheng for serving as a committee member, and to Doctor Leo Anthony Celi for serving as the external examiner.

RÉSUMÉ

Avec la reconnaissance du VIH comme une maladie chronique, il devient essentiel de développer des outils évolutifs et centrés sur les patients pour répondre aux besoins complexes des personnes avec le VIH (PVVIH) au-delà des milieux cliniques. Cette thèse explore comment des chatbots personnalisés basés sur l'intelligence artificielle (IA) peuvent soutenir l'autogestion chez les PVVIH. Ancrée dans les principes de la conception participative et de l'IA responsable, elle examine les considérations clés liées à la conception, le développement et la mise en œuvre d'un tel système dans des contextes de soins réels.

Dans cette thèse par articles, l'Article 1 présente un protocole maître ayant encadré la conception et le développement de MARVIN, un chatbot personnalisé de soutien à long terme pour l'autogestion des PVVIH. Ce protocole repose sur un plan d'essai adaptatif de type plateforme, conçu pour encadrer la co-conception, la mise en œuvre et l'évaluation de chatbots. Il a ainsi permis l'intégration de quatre sous-études ciblant différentes conditions de santé. Ce cadre méthodologique mobilise des méthodes mixtes et des modèles issus de la science de la mise en œuvre afin de permettre une évaluation rigoureuse des interventions basées sur des chatbots. L'Article 1 met également en valeur la démarche de conception participative de MARVIN ainsi que ses principales fonctionnalités techniques.

Dans l'Article 2, ce protocole a été appliqué dans une étude pilote en contexte réel, démontrant la faisabilité, l'utilisabilité et l'acceptabilité de MARVIN auprès des PVVIH. Les résultats quantitatifs ont atteint ou dépassé les seuils prédéfinis, tandis que les données qualitatives ont mis en évidence l'accessibilité, la confidentialité et la sécurité émotionnelle offertes par le chatbot. Les participants ont également signalé certaines limites, notamment une couverture thématique restreinte, l'absence de mémoire et le manque de fonctionnalités proactives.

Pour contribuer à combler ces lacunes, l'Article 3 porte sur le développement d'un outil de triage basé sur un grand modèle de langage (GML), visant à détecter de manière proactive les barrières à l'adhérence au traitement antirétroviral ainsi que les niveaux de risque associés à partir de messages non structurés. Les modèles ajustés ont atteint de solides performances, surpassant les GMLs génériques (par exemple, GPT-4.1) ainsi que des modèles cliniques spécialisés (par exemple, Clinical-T5 Large). L'étude a également évalué l'équité du modèle en analysant les variations de

prédiction selon des descripteurs de genre et de race, et a estimé son empreinte carbone afin d’éclairer son déploiement futur.

Pris dans leur ensemble, ces travaux pourraient faire progresser l’application de l’IA en santé dans le domaine du VIH en présentant l’ensemble du cycle de développement de MARVIN ainsi que ses fondements méthodologiques. Développé selon une approche de conception participative impliquant les PVVIH et les autres acteurs impliqués, MARVIN illustre comment des outils d’IA en santé peuvent intégrer la durabilité, une approche centrée sur l’humain, l’inclusivité, l’équité et la transparence. Les travaux futurs viseront à répondre aux enjeux de pérennisation, de rentabilité et de mise à l’échelle afin d’assurer que cette solution apporte une valeur significative et durable aux soins liés au VIH.

ABSTRACT

With HIV now recognized as a chronic condition, there is a growing need for scalable, patient-centered tools to address the complex care requirements of people with HIV (PWH) beyond clinical settings. This thesis explores how artificial intelligence (AI)-based personalized chatbots can support self-management among PWH. Grounded in participatory design and responsible AI principles, it examines the key considerations in the design, development, and implementation of one such system in real-world care settings.

In this article-based thesis, Article 1 introduces the master protocol which guided the design and assessment of MARVIN, a long-term personalized self-management chatbot for PWH. This protocol is based on an adaptive platform trial design to guide chatbot co-development, implementation, and evaluation. As such, it allowed the incorporation of four chatbot substudies targeting different health conditions. This methodological framework integrates mixed methods and implementation science models to enable rigorous testing of chatbot interventions. Article 1 also highlights MARVIN’s participatory design process and its key technical features.

In Article 2, the protocol was applied in a real-world pilot study, demonstrating the feasibility, usability, and acceptability of MARVIN among PWH. Quantitative findings met or exceeded predefined thresholds, while qualitative feedback highlighted the chatbot’s accessibility, confidentiality and emotional safety. Participants also noted limitations, including a narrow range of topics, lack of memory, and the absence of proactive features.

To help address these gaps, Article 3 focuses on developing a large language model (LLM)-based triage tool to proactively detect antiretroviral treatment adherence barriers and associated risk levels from unstructured patient messages. Fine-tuned models achieved strong performance, outperforming large scale general-purpose LLM (e.g., GPT-4.1) and clinical foundation models (e.g., Clinical-T5 Large). The study also assessed model fairness by analyzing prediction shifts across racial and gender descriptors and estimated the carbon footprint to inform future deployment.

Taken as a whole, the findings could advance the application of AI in the field of HIV care by presenting the entire development lifecycle of MARVIN and its methodological underpinnings. Developed through a participatory design approach involving PWH and stakeholders, MARVIN exemplifies how healthcare AI tools can embody sustainability, human-centeredness, inclusiveness, fairness, and transparency. Future work will address long-term maintenance, cost-effectiveness, and scalability to ensure that this solution delivers meaningful and lasting value to HIV care.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LISTE OF SYMBOLS AND ABBREVIATIONS

The list of acronyms and abbreviations presents, in alphabetical order, the acronyms and abbreviations used in the thesis and their meaning. Examples:

AES	Acceptability E-Scale
AI	Artificial Intelligence
ART	AntiRetroviral Therapy
ATU	Attitude Towards Use
AWS	Amazon Web Services
BERT	Bidirectional Encoder Representations from Transformers
BIU	Behavioral Intention to Use
CI	Confidence interval
CFIR	Consolidated Framework for Implementation Research
CONSORT	Consolidated Standards of Reporting Trials
CONSORT-AI	Consolidated Standards of Reporting Trials–Artificial Intelligence
CONSORT-EHEALTH	Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CoT	Chain-of-Thoughts
FAQ	Frequently-Asked-Questions
ISCORE	I-Score study interviews
I-Score	Interference-Score
IT	Information Technology
LLM	Large Language Model
LoRA	Low-rank adaptation
LSTM	Long Short–Term Memory
MARVIN	Minimal AntiretRoViral Interference

MSM	Men who have sex with men
mHealth	mobile health
MT	MARVIN Training corpus
MU	MARVIN User conversations
MUHC	McGill University Health Centre
NLP	Natural Language Processing
NLU	Natural Language Understanding
NPS	Net Promoters Score
OF	Online Forum messages
PEU	Perceived Ease of Use
PIPEDA	The Personal Information Protection and Electronic Documents Act
PrEP	Pre-Exposure Prophylaxis
PROM	Patient-reported outcomes measure
PVSQ	Portail VIH/SIDA du Québec
PU	Perceived Usefulness
PWH	People with HIV
REB	Research Ethics Board
SD	Standard Deviation
SD	Synthetic Dataset
SEQ	Single Ease of Use Question
SHIFT	Sustainable AI, Human-centred AI, Inclusive AI, Fair AI, Transparent AI
SLM	Small Language Model
STBBI	Sexually transmitted and blood-borne infections
SUS	System Usability Scale

TAM	Technology Acceptance Model
UMUX-Lite	Usability Metric for User eXperience – Lite
UNAIDS	United Nation of AIDS
USMLE	United States Medical Licensing Examination
UTAUT	Unified Theory of Acceptance and Use of Technology
NASSS	Nonadoption, Abandonment, Scale-up, Spread, and Sustainability

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 HIV as a chronic condition: the challenge of lifelong self-management

Human immunodeficiency virus (HIV) is a virus that targets and progressively weakens the immune system [1]. As of 2023, around 40 million people were living with HIV worldwide [2]. In Canada, the estimated number of new HIV diagnoses in 2023 stood at 2,434, a 35.2% increase from 2022 [3]. Although no cure or vaccine has yet been developed more than four decades into the global HIV pandemic, the advent of effective antiretroviral therapy (ART) has significantly improved the prognosis for people with HIV (PWH). As shown in Figure 1.1, life expectancy among those on ART is now comparable to that of HIV-negative individuals, with a 20-year-old starting treatment in recent years expected to live more than 56 additional years [4]. What was once considered a terminal illness bas been transformed into a chronic condition that can be successfully managed on a long-term basis [5].

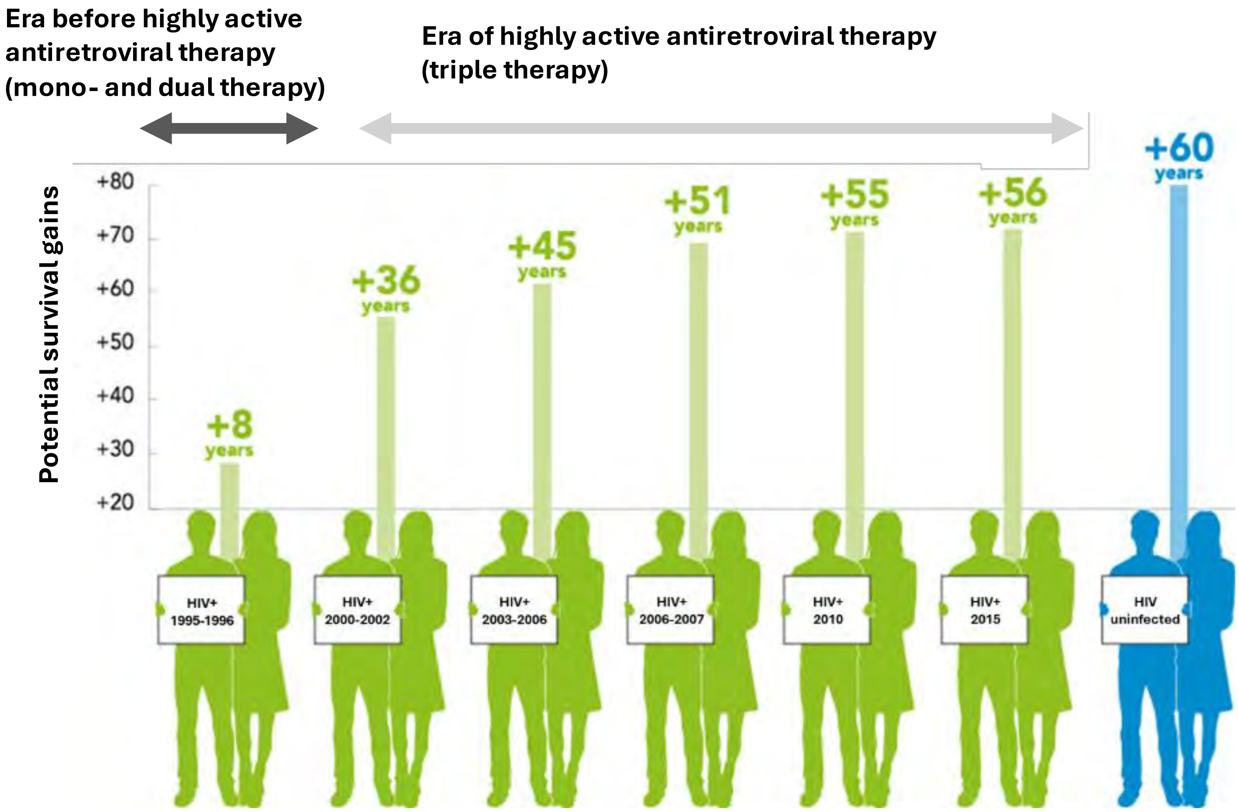


Figure 1.1 Estimated life expectancy of a 20-year-old HIV-positive individual initiating antiretrovrial therapy in a high-income country. Adapted from [4] and [6].

As with most chronic diseases, self-management refers to the ongoing process by which individuals assume primary responsibility for managing their health outside of clinical encounters [7]. For people with HIV (PWH), this encompasses responsibilities such as acquiring and interpreting health information, attending regular medical appointments, making informed treatment decisions tailored to their living conditions, and adhering to strict medication regimens to maintain viral suppression and prevent complications or onward transmission [8-10]. These tasks are further complicated by the increased risk of comorbidities and psychosocial vulnerabilities associated with living with HIV, which can challenge both the delivery of care and the individual's capacity for effective self-management, both crucial to longevity and health-related quality of life [11].

To support effective self-management, it is essential to empower PWH with appropriate knowledge, tools, and support systems to address their questions and needs. PWH now can consult a variety of actors and resources, including healthcare professionals (e.g., physicians, nurses, pharmacists), community-based organizations, peer support workers, and online health resources. However, these sources are not always readily available, and the information they provide can vary in quality and clarity [12, 13]. In some cases, content may be overly technical, outdated, or contradictory, which can create confusion and hinder informed self-management decision-making [14, 15]. These challenges were further amplified during the COVID-19 pandemic, which introduced additional barriers to access regular follow-up [16] and contributed to an "infodemic", i.e., an overabundance of information [17].

These limitations can undermine patient engagement and self-management in their care, with downstream consequences for treatment adherence [18], appointment attendance [19], and long-term health outcomes [20], ultimately increasing healthcare utilization and placing additional strain on both individuals and health systems. As HIV care shifts toward sustained, patient-driven management outside of clinical settings [21], there remains a critical need for consistent, scalable, and personalized tools that provide access to accurate, verified health information to support daily self-management.

1.2 Digital health, mHealth and chatbots: potential solutions

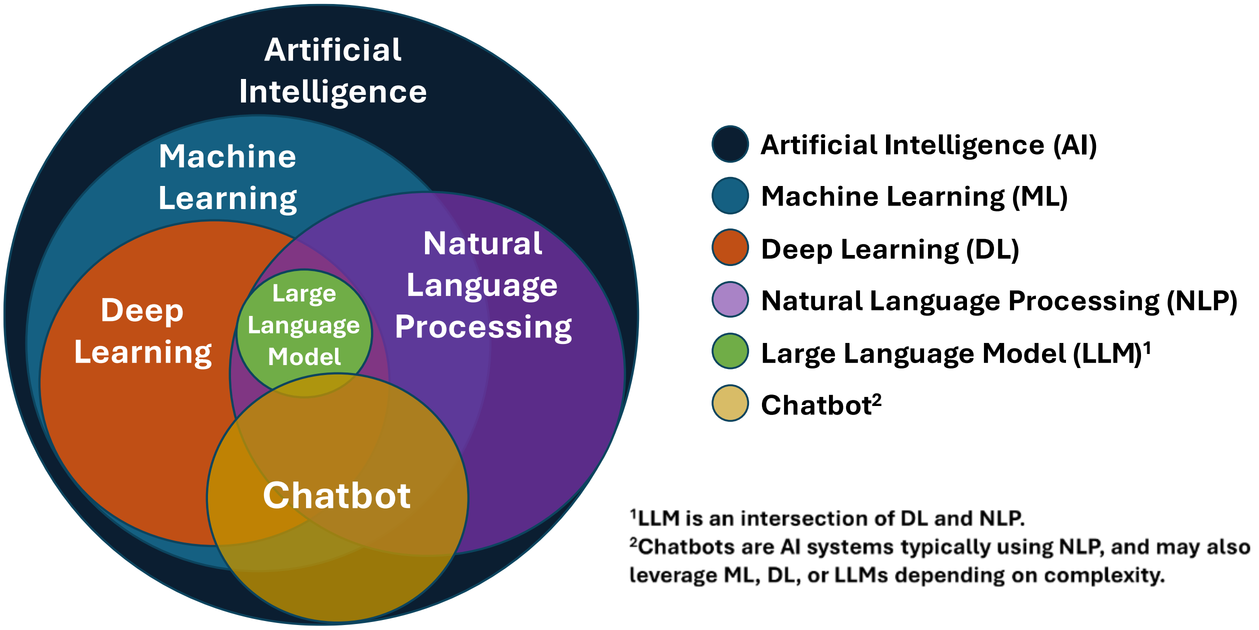
According to the World Health Organization, digital health refers to the use of information and communication technologies, including smartphone messaging, mobile health (mHealth) apps, and web-based platforms, to support population and individual health [22]. Digital health has become a salient practice area in HIV care due to its accessibility, efficiency, and patient-centeredness [23-25]. At the 2021 International AIDS Society Annual Meeting, over 100 digital interventions were showcased. A separate review reported that academic interest in HIV self-testing and digital tools has grown at an annual rate of 12.25% [26]. The scope of digital health has expanded beyond clinical settings to enable daily self-monitoring and self-management [27]. Studies have shown that remote follow-up via telephone or web platforms can improve medication adherence and viral suppression among PWH, while also reducing travel costs, saving time, and protecting privacy [8, 28]. By facilitating swift access to reliable health information, digital health solutions help address gaps in existing self-management support.

One of the most innovative developments in digital health is the emergence of chatbots. Harnessing the power of artificial intelligence to enable natural language processing as well as aid decision-making, such tools can answer questions, offer explanations, and engage users in coherent, context-sensitive dialogue. Chatbots are generally well received by patients [29] and have shown the ability to foster collaborative relationships [30, 31], offering promising opportunities to enhance patient self-management and support [32]. Since the first chatbot, ELIZA, was introduced as a simulated psychotherapist in 1966 [33], chatbots have been explored in a range of healthcare domains, including mental health [31, 34, 35], oncology [29, 36-38], and diabetes management [39, 40]. However, their application in HIV care remains limited [41], with most existing efforts focused on prevention [42-47]. Many of these chatbots are still at the pilot stage, with little evidence of real-world implementation. As a result, there is a need to develop tailored, evidence-based chatbot solutions specifically designed to support the self-management of PWH.

1.3 So, what is Artificial Intelligence and Natural Language Processing?

Artificial intelligence (AI), a term first introduced at the Dartmouth Workshop back in 1956 [48], now broadly refers to the development of computational systems capable of performing tasks that typically require human intelligence. Within AI, machine learning usually refers to statistical learning algorithms that are endowed with the ability to improve themselves from experience without explicit programming [49]. A further subset, deep learning, represents computational models that use multilayer neural networks to learn data representations with multiple levels of abstraction [50]. Figure 1.2 illustrates the conceptual relationships among these overlapping techniques.

Over the past decade, the liberation of arithmetic power resulting from the development of hardware has led to the explosive growth of AI and the popularity of machine learning and deep learning. Among the most transformative areas is natural language processing (NLP), which enables machines to interpret, generate, and analyze human language [51]. A major breakthrough came in 2017 with the release of the transformer architecture [52], which laid the groundwork for today's large language models (LLMs). One of the earliest and most influential examples was BERT (Bidirectional Encoder Representations from Transformers), developed by Google in 2018 [53].



[1] LLM is an intersection of DL and NLP.
[2] Chatbots are AI systems typically using NLP, and may also leverage ML, DL, or LLMs depending on complexity.

Figure 1.2 Relationship between AI subfields and language technologies. Inspired by [54]

Table 1.1 provides a glossary and definitions of key concepts related to LLM training and adaptation. LLMs are typically pre-trained on massive general-domain corpora using self-supervised learning, which enables the model to learn language patterns and structures at scale. They can then be fine-tuned – that is, further trained on domain- or task-specific datasets to optimize performance for specialized applications. In healthcare, clinical foundation models – developed through continued pre-training on biomedical literature or clinical text – offer domain-adapted alternatives [55, 56]. As training data and model sizes have scaled exponentially, LLMs have become significantly more powerful [57]. The release of ChatGPT in 2022 marked a turning point, showcasing models capable of fluent, context-sensitive conversation and gaining widespread popularity among the general public [58]. Today's frontier LLMs, such as GPT-4 [59], DeepSeek [60], and Claude [61], can generalize across diverse tasks with minimal instruction through prompt-tuning [62] – a technique that steers the model's behavior by optimizing the input prompts it receives, rather than changing model weights. These advances have enabled the rise of sophisticated AI chatbots like ChatGPT, capable of passing expert-level exams such as the United States Medical Licensing Examination (USMLE) [63].

Despite promising advances, the application of LLMs–and AI more broadly–is still underexplored in HIV care [41, 64]. This limited uptake may be attributed to concerns about data scarcity and representativeness in HIV-specific contexts [65], alongside broader issues affecting LLM deployment in healthcare such as hallucinated outputs, privacy risks, and environmental impact [41, 66]. Stakeholder perspectives also remained mixed: one study found that half of surveyed healthcare professionals were hesitant to use chatbots due to safety concerns and limited understanding of the technology [67], while another showed most internet users expressed willingness to use health chatbots [68]. Ensuring clinical relevance and equitable outcomes requires co-developing AI tools with PWH and other key stakeholders [41, 64]. Without careful development and stakeholder involvement, such tools risk perpetuating existing disparities rather than addressing them. Further work is needed to guide the development of AI-powered interventions that are not only technically robust, but also equitable, trustworthy, and relevant to the lived experiences of PWH.

Table 1.1 Adapting large language lodels: a comparison of training and tuning approaches.

	Pre-training[1]	**Continued Pre-training**[2]	**Fine-tuning**[3]	**Prompt Tuning**[4]
Training Data Type	General-domain text (e.g., books, web data)	Domain-specific text (e.g., PubMed)	Task-specific data (e.g., labeled clinical texts)	Task-specific prompts
Typical Dataset Size	Very large (100B+ tokens[5])	Moderate to large (1B–10B tokens)	Small to moderate (10K–1M samples)	Very small (zero-/few-shot prompting)
Domain Specificity	Low (general-purpose)	Medium to high (domain-specific)	High (task-specific)	Medium to High (task-dependent)
Example Models	GPT-4 [59], BERT [53]	BioBERT [69], Clinical-T5 [70]	Fine-tuned BERT for text classification [71, 72]	GPT-4 via prompt [62]
Typical use case	Foundation model creation	Domain adaptation	Task adaptation	Parameter-efficient task adaptation[6]

[1] Pre-training refers to training a model from scratch on very large general-domain text.

[2] Continued pre-training builds on a pre-trained model using domain-specific text.

[3] Fine-tuning adapts a pre-trained model to specific tasks using task-specific data.

[4] Prompt tuning guides model behavior by optimizing task-specific prompts rather than modifying model weights.

[5] Tokens are units of text (e.g., words or subwords) used by language models during training. One English word typically corresponds to 1–2 tokens.

[6] Parameter-efficient methods adapt a model without updating all its parameters, requiring less computation and data.

1.4 Thesis outline

Chapter 2 reviews the current literature on digital health interventions for HIV self-management, the use of chatbots in HIV care, and the emerging role of LLMs in supporting chronic disease self-management, highlighting key gaps in the field and establishes the rationale for the research presented in this thesis.

Chapter 3 outlines the overarching research objectives and the overview of the methodology guiding the work presented in this thesis. It also provides an outline of the three research articles included in the thesis, each addressing a distinct aspect of chatbot development, implementation, and evaluation.

Chapter 4 introduces a master protocol based on an adaptive platform trial design to guide chatbot's co-development, implementation, and evaluation using mixed methods and implementation science frameworks. It also details the general design and development of the MARVIN chatbot, a long-term personalized self-management tool for PWH, highlighting the participatory design process and key technical features.

Chapter 5 applied this protocol in a real-world mixed-methods study, demonstrating the feasibility, usability, and acceptability of the MARVIN chatbot among PWH, while revealing limitations such as narrow topic coverage and lack of proactive features.

Chapter 6 addresses this gap by developing an LLM-based triage tool to proactively identify ART adherence barriers and associated risk levels from unstructured patient messages. Fine-tuned models achieved state-of-the-art performance, outperforming both large-scale general-purpose (e.g., GPT-4.1) and clinical foundation models (e.g., Clinical-T5 Large). The study also assessed model fairness and estimated the carbon footprint of model development and deployment to inform future implementation.

Chapter 7 provides a general discussion on the design and development of the chatbot-based personalized healthcare tool, synthesizing insights from the three articles and discussing the broader implications and contributions of the work.

Chapter 8 concludes the thesis by highlighting the results and the scientific significance of this research.

CHAPTER 2 LITERATURE REVIEW

The introduction of the BERT model in 2018 and the release of ChatGPT in 2022 have driven a sustained and rapid increase in scientific publications referencing the “healthcare chatbot” concept, as shown in Figure 2.1-A. A similar trend is observed in HIV-specific chatbot research (Figure 2.1-B), though with a smaller overall volume. This growth was further catalyzed by the COVID-19 pandemic, which accelerated the adoption of digital health solutions [73, 74]. Given this recent surge, this literature review focuses primarily on work published in the past five years.

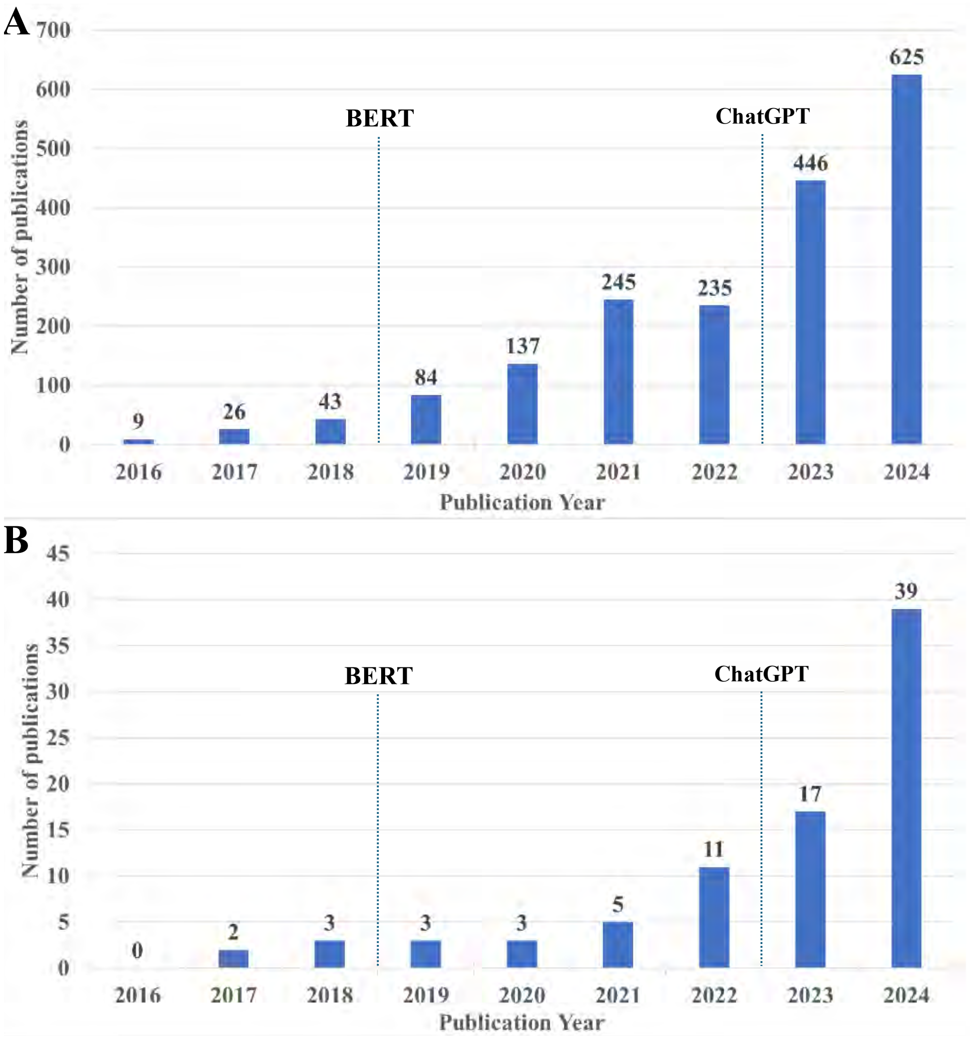


Figure 2.1 Number of scientific publications related to chatbots in health and HIV care indexed on *Web of Science* (2016–2024) (A) “All fields” search using: ("chatbot" OR "chat bot" OR "Conversational agent" OR "Intelligent Conversational Agent") AND ("health" OR "healthcare" OR "health care") (B) “All fields” search using: ("chatbot" OR "chat bot" OR "Conversational agent" OR "Intelligent Conversational Agent") AND ("HIV" OR "AIDS" OR "PrEP" OR "Antiretroviral" OR "Antiretroviral therapy") as subject terms.

2.1 Overview of digital health interventions for HIV self-management

This section provides a rapid review of digital health interventions for HIV self-management published between January 1, 2021, and the present. The aim is to identify the established benefits of these tools for self-management support, while also highlighting persistent limitations. This review provides the foundation for understanding how AI-powered chatbots may build upon, or address gaps in, earlier digital approaches. Relevant studies were identified using the following terms in the Web of Science database: "HIV" AND "Self-management" AND ("Digital Health" or "App" or "mobile health") (All fields). Inclusion criteria were limited to English-language peer-reviewed journal articles; research protocols and conference abstracts were excluded.

Several mHealth interventions have been developed to support different dimensions of HIV self-management. For instance, *Champs* [75] paired a medication adherence app with a smart bottle that recorded dosing events, supported by 12 online follow-up visits with a community health worker to facilitate implementation. The intervention was feasible and well-accepted, particularly in rural settings, but showed average retention (60%) and posed scalability concerns due to its dependence on physical devices and human labor.

Other interventions focused on specific health domains. *VIP-HANA* [76] enabled PWH to self-monitor comorbidities and receive tailored guidance based on symptom reporting. Similarly, *EmERGE* [77] implemented across five European countries, provided access to test results, medication reminders, and appointment management through clinical system integration. It achieved high engagement (87% retention) and strong user endorsement (94% recommendation rate), highlighting the value of embedding digital tools within existing care systems.

Interventions such as *Epic allies* [78] and *Bijou* [79] were developed for PWH, aiming to improve health literacy and overall self-efficacy through educational modules. Participants reported knowledge gains and increased motivation. However, both faced retention issues. Only 36% of *Epic Allies* users remained active at week 26, with many expressing a need for more personalized content. Similarly, only 52% of *Bijou* participants completed the study, often citing platform accessibility challenges, such as difficulty remembering how to log in.

A more participatory approach was taken by the *ESSC (Excellent Self Supervised HIV Care)* mobile app [80], which involved over 240 PWH in a needs assessment and went through three iterative design process. The resulting app provided multi-faceted, user-driven informational support covering medication adherence, symptoms, mental health, and sexual health. Clinical trials reported improvements in self-efficacy and reduced stigma [81, 82]. Yet, only about 10% of participants voluntarily continued daily use of the app after an 8-week period, underscoring the need for sustained engagement strategies [81]. Participants expressed a desire for a more interactive user experience and timely access to the latest information related to HIV care research.

2.1.1 Strengths, limitations, and implications for chatbot development

Digital health interventions have demonstrated strong potential to support ART adherence, promote patient engagement, and improve self-management outcomes among PWH [76, 78, 79, 81-84]. Studies generally reported favorable implementation outcomes such as feasibility, usability, and acceptability, aligning with Proctor's framework for evaluating implementation research [85]. However, long-term retention remains a persistent challenge, often due to limited personalization, interactivity, and integration into daily routines or clinical workflows.

Among these factors, content relevance and breadth were strong determinants of usability and acceptability. Even narrowly focused tools, such as those addressing ART adherence [75] or comorbidity management [76], were well-received when content was clearly relevant. However, broader needs assessments consistently emphasized the value of comprehensive, user-driven content that spans multiple self-management domains [80, 86]. Greater content variety not only supports a wider range of concerns but also enables more nuanced personalization, an area where many interventions underperformed [78, 81, 86]. For chatbots to effectively support HIV self-management, they should deliver both diverse and personalized content that reflects the complex needs of PWH as they navigate daily self-management responsibilities.

Accessibility and interface design also shaped usability and long-term engagement. Some users reported difficulty accessing platforms [79], while others highlighted that ease of use is a key determinant of intervention usability [81]. Clear, simple interfaces are essential to ensure accessibility [86]. While human facilitation can improve access, reliance on such support limits scalability and automation – factors that are especially critical in resource-limited settings [75]. Chatbots, with their capacity for automated, natural language interaction, are well-positioned to address both usability and scalability through more intuitive, self-directed engagement.

Interactivity was another frequently cited gap [78, 82, 86]. Although many tools employed SMS or app notifications, participants often perceived a lack of responsiveness and wanted more dynamic, two-way communication [82, 86]. Prior work suggests that interactive features, particularly those that offer personalized feedback and emotional support, can enhance motivation and promote self-management behaviors [87]. In this context, chatbots offer a distinct advantage by enabling real-time, conversational interactions that support engagement over time.

Finally, the process of development itself played a critical role. Participatory design, especially the involvement of PWH in early-stage input and iterative prototyping, proved instrumental in enhancing usability and acceptability [77, 79-82]. However, the depth of involvement varied widely across studies, and limited co-design activities may have contributed to usability gaps in some cases [88]. These insights highlight the need for meaningful participatory design in chatbot development to ensure that tools are usable, acceptable, and responsive to the lived experiences of PWH.

2.2 Chatbot use in HIV care

This section synthesizes research on the development of HIV care-related chatbots published between January 2018 and June 2025. Further, the analysis is structured around three primary frameworks: the digital health intervention lifecycle [89], a chatbot design taxonomy summarized from the work of Janssen et al. [90] and Nißen et al. [91], and the SHIFT (Sustainable AI, Human-centred AI, Inclusive AI, Fair AI, Transparent AI) framework for responsible healthcare AI [92]. This approach provides a comprehensive understanding to both development practices and ethical considerations relevant to the design of chatbots for HIV self-management. Relevant studies were identified through the Web of Science database using the following terms: ("chatbot" OR "chat bot" OR "conversational agent" OR "intelligent conversational agent") AND ("HIV" OR "AIDS") (All fields). The inclusion criteria were limited to English-language peer-reviewed journal articles and research protocols; conference abstracts were excluded.

2.2.1 Review approach

2.2.1.1 Conceptual framework – Digital health intervention lifecycle

While AI-based digital health interventions have seen rapid progress in model development, their real-world integration into routine care remains a significant challenge, often staying within the realm of "innovation" rather than core care delivery processes [93]. Successfully integrating AI into healthcare requires a systematic approach to development and implementation within the anticipated operational environment.

According to Li et al. [89] and Bitomsky et al. [94], the iterative development model for AI-based digital health systems consists of four steps: 1) needs assessment; 2) design and development; 3) implementation; 4) evaluation and dissemination. Understanding the focus of each chatbot intervention study within this lifecycle helps us assess its maturity and readiness for integration into HIV care.

2.2.1.2 Conceptual framework – Chatbot design taxonomy

Chatbots are complex, modular systems that require thoughtful design principles and architecture. Taxonomies are valuable for understanding the scientific principles behind the observed artifacts, including digital service products. Several taxonomies for chatbot design have been proposed [90, 91, 95-98]. The analysis in this review is mainly based on the taxonomy proposed by Janssen et al. [90] and Nißen et al. [91] (see original taxonomies in Annexe A), focusing on the following criteria:

- Time horizon: This criterion describes the expected duration of the user-chatbot relationship:
    - Short-term – limited to single or occasional interactions.
    - Medium-term – spanning multiple interactions over several days or weeks.
    - Long-term – sustained over an extended period, such as during a weight loss program.
    - Lifelong – designed to accompany the user through various stages of life, offering continuous support and companionship.
- Primary communication style: This refers to the main purpose of user-chatbot interaction:
    - Socially-oriented – focusing on chat-based or relational engagement.
    - Task-oriented – aiming to complete specific tasks.
    - Informative – delivering structured information or answering frequently-asked-questions (FAQs) in an interactive manner [98].
- Intelligence framework: This criterion reflects the chatbot’s underlying technological capabilities:
    - Rule-based – relying on predefined "if-then" rules.
    - Text understanding – capable of natural language understanding (NLU).
    - Text understanding+ – incorporating additional functions such as logical inference, mathematical calculation, or image recognition.

- Front-end user interface: This indicates the platforms through which users access the chatbot:
    - Standalone applications.
    - Social media messengers (e.g., Facebook Messenger).
    - Websites.
    - Communication platforms (e.g., WhatsApp).
    - Combinations of these interfaces.
- User assistance design: This defines how conversational control is managed during interaction:
    - Reactive – the chatbot responds only to user inputs.
    - Proactive – the chatbot steers the conversation.
    - Reciprocal – the chatbot can alternate.
- Personalization: This refers to the chatbot’s ability to tailor interactions to the user:
    - Static – responses remain consistent regardless of user history.
    - Adaptive – the chatbot adjusts its responses based on prior interactions and user-specific data.

2.2.1.3 Conceptual framework – Responsible AI: The SHIFT Framework

Ethical considerations are paramount for best practices in healthcare AI. Currently, there are several ethical frameworks and guidelines available for reference. We have summarized the key points based on the comprehensive SHIFT framework for assessing responsible AI in healthcare proposed by Siala et al. [92] to guide our analysis:

- Sustainable AI: Promoting responsible leadership and long-term viability of AI initiatives, while addressing social, economic, and environmental impacts to support human well-being.
- Human-centered AI: Ensuring appropriate integration of chatbots into healthcare by clarifying their roles in relation to providers, i.e., assistive or substitutive functions.
- Inclusive AI: Incorporating participatory design, stakeholder engagement, and cultural tailoring (e.g., language adaptation) to meet the needs of diverse populations and reduce potential disparities.
- Fair AI: Promoting equity through attention to fairness in data collection and model development, addressing health disparities among all user groups.

- Transparent AI: Upholding data privacy and security, ensuring informed consent for data use, enhancing model interpretability, and clearly communicating system limitations to users and stakeholders.

This framework allows us to assess not only technical and clinical rigor, but also the degree to which chatbot interventions align with ethical best practices.

2.2.1.4 Review questions and synthesis approach

In this review, we pose the following five questions to assess the current development of HIV care-related chatbots:

1) Who are the target users of the chatbots and what features are expected to meet user needs?
2) How are chatbots technically designed and developed?
3) How have these chatbots been implemented and evaluated?
4) What are the reported outcomes and challenges of chatbot interventions?
5) How is the concept of responsible AI addressed across chatbot studies?

To answer these questions, we extracted and synthesized detailed information from each reviewed articles based on the following structured categories:

- Needs assessment: Target populations and clinical aim, identified needs and expected features
- Design and development: Time horizon, primary communication style, intelligence framework, front-end user interface, user assistance design, personalization
- Implementation, evaluation and dissemination: Evaluation outcomes (technical outcomes, implementation outcomes and service outcomes), identified challenges and future plan
- SHIFT Dimensions: Explicit actions or assessments related to the concepts of Sustainable AI, Human-centric AI, Inclusive AI, Fair AI, and Transparent AI

Specifically, information on needs assessment, implementation, evaluation and dissemination, and SHIFT dimensions was extracted from article content, while design and development details were synthesized from the identified chatbots, given our technical focus.

2.2.2 Review findings

2.2.2.1 Overview

Table 2.1 summarizes the characteristics of the 19 articles that met the inclusion criteria, including 15 research studies and 4 research protocols. These articles were originated from United States (n=4), Brazil (n=2), Malaysia (n=2), Peru (n=2), Spain (n=2), South Africa (n=2), Canada (n=1), China (n=1), Nigeria (n=1), United Kingdom (n=1), and Zambia (n=1). According to the World Bank[99], most articles were from high-income (n=8) or upper-middle-income countries (n=9). Only two were from low-income countries (n=2). Most articles were published in 2024 (n=7), followed by 2025 (n=4), 2022 (n=4), 2023 (n=3) and 2018 (n=1).

2.2.2.2 Needs assessment

Four of the reviewed articles (4/19) primarily focused on conducting a comprehensive needs assessment study [100-103]. The remaining articles (15/19), while not explicitly centered on this step, briefly addressed various components of needs assessment throughout the paper.

2.2.2.2.1 Target populations and clinical aim

Most studies (12/15) and protocols (3/4) focused on chatbots for HIV prevention [42-47, 100, 103-110]. Within this aim, two articles specifically targeted HIV self-testing [43, 105], while three addressed prevention at a broader STI/HIV education level [47, 104, 108]. Among the chatbot user populations considered were:

- General population at risk (5 articles) [46, 104-106, 108].
- Men who have sex with men (MSM) (5 articles) [42-44, 100, 103], with one article targeting Black MSM, and one targeting Chinese MSM.
- Women at risk (3 articles) [47, 107, 110], including one specifically on Black cisgender women [110].
- Adolescents and young adults at risk (2 articles) [45, 109].

Beyond prevention, two articles focused on chatbot use to support the mental health of adolescents with HIV [111, 112]. One article aimed to facilitate communication between HIV-related study participants and research staff [101], and another focused on assisting community pharmacists treating PWH [102].

Table 2.1 Characteristics of the articles and chatbots inlcuded in the review.

Team	Article	Publication year	Article type	Country	Chatbot	Chabot language
Van Heerden et al.	[46]	2018	Research study	South Africa	Nolwazi	English
	[105]	2022	Research study	South Africa	Nolwazi_bot	English & isiZulu
Nadarzynski et al.	[106]	2024	Research study	United Kingdom	Casa	English
Botti et al.	[104]	2022	Research study	Spain	VIHrtual-App	Spanish
	[108]	2024	Research study	Spain		
Population Council	[47]	2022	Research study	Zambia	HIV/FP Chatbot	English
	[110]	2025	Research study	Nigeria	Let's chat!	English & Pidgin English
Zhao et al.	[100]	2022	Research study	Malaysia	/	/
	[44]	2024	Research study	Malaysia	Haris	English
	[103]	2025	Research study	United States	/	/
Comulada et al.	[42]	2023	Research protocol	United States	TelePrEP	English
	[101]	2024	Research study	United States	/	English
Massa et al.	[45]	2023	Research study	Brazil	Amanda Selfie	Pajubá
	[109]	2025	Research study	Brazil		
Galea et al.	[90]	2024	Research protocol	Peru	E.V.A.	Spanish
	[112]	2025	Research study	Peru		
Lebouché et al.	[102]	2024	Research study	Canada	/	/
Wang et al.	[43]	2023	Research protocol	China	HIVST-Chatbot	Chinese
Zhang et al.	[107]	2024	Research protocol	United States	/	/

2.2.2.2.2 Identified needs and expected features

Across diverse target populations, a consistent set of user needs emerged, reflecting critical gaps in current HIV care and aligning with broader digital health priorities. Chief among these was the need for accessible, on-demand health information and services. For instance, users seeking HIV prevention information hoped that chatbots could answer their questions and provide diverse, relevant educational content, including HIV counseling and information on testing, safe sexual practices, and pre-exposure prophylaxis (PrEP) treatment [43, 45, 100, 103, 105, 110]. This information was expected to be professionally verified for accuracy and reliability [42-45, 100, 104, 106, 108], and presented in a clear and comprehensive manner [42, 44-46, 104, 105, 108, 110, 112], with multimodal formats (e.g., text, images, videos, audio) suggested to aid understanding [46, 104, 108]. Support for multiple languages was also noted as important for accessibility [45, 47, 105, 106, 110].

Beyond basic question-and-answer functionality, users expected advanced language understanding for handling complex scenarios [42, 44, 100, 103, 105], as well as task support through interactive prompts [42, 44, 46, 100, 102, 104, 105]. Such tasks included automated appointment scheduling [101], setting medication reminders [100], and completing surveys [101]. When necessary, chatbots should also facilitate referrals to real external services, such as real-time professional or community-based support [43-45, 100, 101, 103, 111, 112].

Ease of use was a key user requirement, with users calling for intuitive, user-friendly interfaces [44, 47, 100, 102-104, 108, 110, 112]. Suggestions included onboarding guides for new users [100, 104, 110, 112], and ongoing technical support for troubleshooting [44, 100, 110]. For pharmacists supporting PWH, seamless integration into existing workflows was emphasized to avoid creating additional technical burdens[102]. Cross-platform accessibility was also considered essential [45, 47, 100, 101, 108], particularly through channels that are inclusive and friendly to vulnerable groups, e.g., MSM [44, 100, 103].

Privacy and anonymity were especially critical in stigmatized contexts surrounding HIV and sexual orientation [43, 44, 46, 47, 100, 103-106, 108, 110, 112]. Users favored anonymous chatbot interactions, [44, 46, 100, 103], with some explicitly preferring deployment on neutral platforms to avoid disclosing identities or linking to social media accounts [100].

Finally, emotional support was a key dimension of user expectations [42-46, 100, 103-106, 108, 110-112]. Many valued the non-judgmental nature of chatbot interactions, especially in contexts where stigma is prevalent [42, 44, 45, 100, 103, 105-107]. However, they also raised concerns about impersonal or robotic tones in text-based interactions [100, 103], and emphasized the importance of engaging end-users in the design process to ensure that communication styles are appropriately tailored to their emotional and relational needs [45, 103, 104].

2.2.2.3 Design and development

Across the 19 reviewed articles, eleven distinct chatbots were identified: 1) *Nolwazi* [46] and 2) its updated version, *Nolwazi_bot* [105], 3) *Casa* [106], 4)*VIHrtual-App* [104, 108], 5) a chatbot developed by Yam et al. [47], which was further developed into 6) *Let's chat!* by the same team [110], 7) *Haris* [44], 8) *TelePrEP* [42], 9) *Amanda Selfie* [45, 109], 10) *E.V.A.* [111, 112], and 11) *HIVST-Chatbot* [43]. The four needs assessment studies and one protocol [107] did not involve the development of a specific chatbot product. Table 2.2 summarizes the current status of these chatbots based on the selected taxonomy dimensions.

2.2.2.3.1 Chatbot design taxonomy – Time horizon

Most chatbots reviewed were designed for short-term use, offering one-time interactions focused on specific topics. For instance, *Nolwazi* [46], *Nolwazi_bot* [105], *Haris* [44] and *TelePrEP* [42] guides users through HIV risk screening questions and provide tailored preventive measures based on the assessment. The chatbot developed by Yam et al. [47] and *Let's chat!* [110] deliver sexual health and HIV-related information to patients during clinic wait times. Similarly, *Casa* [106] and *VIHrtual-App* [104, 108] focus on answering users' questions related to HIV. And *E.V.A.* [112] guides users through mental health screening and assessment video modules within a single session.

In contrast, only two chatbots support longer-term engagement. *HIVST-Chatbot* [43], for example, is supported by a larger knowledge database to answer users' ongoing questions about HIV risk in everyday life. In addition, it recommends and supports HIV self-testing, offering follow-up interactions based on test results. *Amanda Selfie* [45] also operates with a long-term focus. Beyond providing HIV prevention-related knowledge, it allows users to freely choose when to receive PrEP reminders and proactively asks users if they have picked up their medication from the pharmacy.

Table 2.2 HIV care chatbots: current status based on chatbot taxonomy proposed by Janssen et al. [89] and Nißen et al. [90].

Chatbot	**Article**	**Time horizon**	**Primary communication style**	**Intelligence framework**	**Front-end user interface**	**User assistance design**	**Personalization**
Nolwazi	[46]	Short-term	Task-oriented	Text understanding	Telegram	Reactive	Adaptive
Nolwazi_bot	[105]	Short-term	Task-oriented	Text understanding	Telegram	Reactive	Adaptive
Casa	[106]	Short-term	Task-oriented	Rule-based	Website	Reactive	Adaptive
VIHrtual-App	[104, 108]	Short-term	Informative	Text understanding	Website	Reactive	Adaptive
HIV/FP Chatbot	[47]	Short-term	Informative	Rule-based	Website	Proactive	Static
Let’s chat!	[110]	Short-term	Informative	Rule-based	Website	Proactive	Static
Haris	[44]	Short-term	Task-oriented	/	Website	/	/
TelePrEP	[42]	Short-term	Informative	Text understanding	Website	Reactive	Static
Amanda Selfie	[45] [109]	Long-term	Task-oriented	Text understanding +	Facebook Messenger	Reciprocal	Adaptive
E.V.A.	[90] [112]	Short-term	Task-oriented	Rule-based	/	Reactive	Static
HIVST-Chatbot	[43]	Medium-term	Task-oriented	Text understanding	WhatsApp	Reactive	Adaptive

2.2.2.3.2 Chatbot design taxonomy – Primary communication style

More than half of the reviewed chatbots (7/11) were task-oriented, designed to guide users through specific actions. Two focused on promoting HIV risk screening and counseling [46, 106], and three further facilitated HIV self-testing [43, 44, 105]. *Amanda Selfie* focused on promoting HIV prevention and pre-exposure prophylaxis (PrEP) administration, and *E.V.A.* supported mental health screening and assessment for youth with HIV [112].

The remaining chatbots (4/11) provided primarily informative support, either by answering users' questions about sexual health and HIV prevention [42, 104, 108], or by delivering general educational content on these topics [47, 110].

2.2.2.3.3 Chatbot design taxonomy – Intelligence framework

Four chatbots used only rule-based frameworks [47, 106, 110, 112], while most relied on NLP to interpret user input [42, 43, 46, 104, 105, 108]. *Amanda Selfie* stood out by combining semantic understanding with additional capabilities, i.e., calculating time intervals and medication pill counts [45]. No technical details on *Haris'* intelligence framework were provided [44].

2.2.2.3.4 Chatbot design taxonomy – Front-end user interface

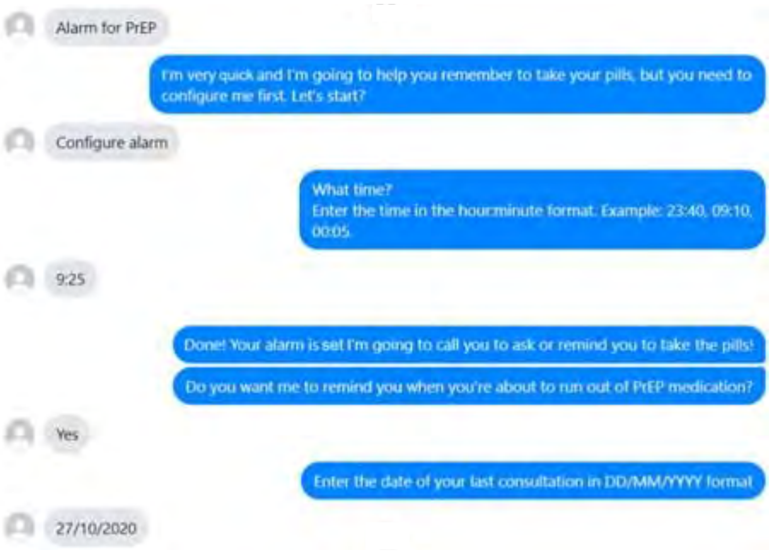
A wide range of interface options were used across the reviewed chatbots: *Amanda Selfie* [45] was deployed via Facebook Messenger; *Nolwazi* [46], *Nolwazi_bot* [105] and *HIVST-Chatbot* [43] used Telegram and WhatsApp, respectively; and six chatbots were accessed through their own websites [42, 44, 47, 104, 106, 108, 110]. No interface information was available for *E.V.A.* [112]. Notably, none of these chatbots were deployed on more than one platform.

2.2.2.3.5 Chatbot design taxonomy – User assistance design

Most reviewed chatbots (7/11) employed a reactive user assistance design, requiring users to initiate and navigate the conversation [42, 43, 46, 104-106, 108, 112]. In contrast, the chatbot developed by Yam et al. [47] and *Let's chat!* [112] adopted a proactive approach, where the chatbot guided the interaction. *Amanda Selfie* used a reciprocal model, allowing for dynamic, two-way exchanges between user and bot [45]. No technical details on *Haris*' user assistance design approach were reported [44].

2.2.2.3.6 Chatbot design taxonomy – Personalization

Four chatbots were static, offering no adaptation based on prior user interactions [42, 47, 110, 112]. Except for *Haris* [44], for which no personalization details were reported, all others demonstrated some ability to personalize conversations. *Nolwazi* [46], *Nolwazi_bot* [105], and *HIVST-Chatbot* [43] adjusted responses based on users' HIV risk assessments. *VIHrtual-App* [104, 108], *Casa* [106], and *Amanda Selfie* [45] used NLP to detect user intent and deliver tailored responses, with *Amanda Selfie* further personalizing interactions through customized medication reminders [45] (See Figure 2.2).



Alarm for PrEP

I'm very quick and I'm going to help you remember to take your pills, but you need to configure me first. Let's start?

Configure alarm

What time?
Enter the time in the hour:minute format. Example: 23:40, 09:10, 00:05.

9:25

Done! Your alarm is set I'm going to call you to ask or remind you to take the pills!

Do you want me to remind you when you're about to run out of PrEP medication?

Yes

Enter the date of your last consultation in DD/MM/YYYY format

27/10/2020

Figure 2.2 Cuztomized medication reminders by *Amanda Selfie* (right). PrEP: pre-exposure prophylaxis. Figure extracted from [45].

2.2.2.4 Implementation and evaluation approaches

This section reviews the 11 studies and 4 protocols that involved chatbot development, excluding the four needs assessment studies.

Across the reviewed work, study designs and evaluation approaches varied considerably. One study focused solely on technical development, reporting prediction metrics such as precision, recall, and F1-score [104]. One protocol did not specify any implementation strategy or study design [42]. Among the remaining articles (n = 13), three were pilot observational studies conducted in simulated environments with fewer than 30 participants [44, 46, 112]. Six observational studies were implemented in real-world settings [45, 47, 105, 106, 108, 110], including one that was followed by a cost analysis [109]. Of the three protocols, one presented a randomized controlled trial [43], while the other two outlined observational designs [107, 111].

In terms of evaluation approach, two studies employed qualitative methods only [46, 112]; four studies and two protocols used quantitative methods [43, 45, 106, 108, 109, 111]; and three studies along with one protocol adopted a mixed-methods design [44, 47, 107, 110].

Finally, theory-based implementation is essential for understanding and improving the success of evidence-based practice [113, 114]. However, only two studies explicitly referenced theoretical frameworks to guide the outcomes analysis. For example, *Haris* was evaluated through the Unified Theory of Acceptance and Use of Technology (UTAUT) framework to guide the analysis of user experience [44]. *E.V.A.* was evaluated through Sekhon's framework of healthcare intervention acceptability [112].

2.2.2.5 Evaluation outcomes and challenges

In terms of evaluation outcomes, most articles (10/13) focused, or were planning, on assessing implementation outcomes, mainly feasibility (5 studies, 1 protocol) [44, 46, 47, 105, 110, 111], acceptability (5 studies, 1 protocol) [44, 47, 105, 110-112], and usability (3 studies, 1 protocol) [44, 107, 108, 110]. One study did a cost analysis [109].

Feasibility refers to the extent to which an intervention can be successfully implemented within a given context or setting [85]. All studies assessing feasibility concluded that the chatbot interventions were feasible within their respective settings. For instance, 7 out of 10 testers indicated they would use *Nolwazi* in the future for HIV counseling and testing support [46], and 11 out of 14 expressed similar intentions for *Haris [44]*. *Let's chat!* was considered feasible for use in clinical waiting areas, with 77% of users reporting that the average session duration (approximately 25 minutes) was appropriate and informative.

Similar positive outcomes were reported in terms of acceptability, which refers to the extent to which an intervention is perceived as satisfactory, appropriate, or agreeable by its intended users [115]. Most participants (95/120, 79.2%) preferred the *Nolwazi_bot* over human counselors and felt as if they were talking to a real person [105]. All participants who tested *Haris* (14/14) expressed confidence in using the chatbot, and believed others could also quickly learn to use it [44]. *E.V.A.* was also deemed acceptable, with caregiver participants valuing its educational content, self-help tools, and potential to enhance communication with their children, especially around HIV diagnosis [112].

Usability refers to the degree to which a product enables users to achieve specific goals effectively, efficiently, and with satisfaction [116]. Regarding usability, both *VIHrtual-App* [108] and *Haris* [44] above-threshold System Usability Scale (SUS) [Ref] scores of 85 and 76, respectively, surpassing the recommended threshold of 68. *Let's chat!* [110] also demonstrated strong usability, with 97% of users finding the chatbot content easy to understand. Participants also highlighted its user-friendliness, attributing it to its interactive nature, guided response options, and seamless operations.

A limited number of articles (5/13) evaluated or planned to evaluate service outcomes [43, 45, 106, 108, 110]. For instance, users of *VIHtural-App* [108] and *Let's chat!* [110] demonstrated increased knowledge of HIV prevention following chatbot interaction. The *HIVST-Chatbot* protocol planned to assess HIV self-testing uptake and subsequent PrEP linkage following chatbot use [43].

Amanda Selfie was the only chatbot intervention evaluated through a formal cost-effectiveness analysis [109]. Among six youth-targeted HIV prevention strategies in Brazil, it had the highest average cost per participant (US$5,572), primarily due to software development costs. Despite this, only 16.2% of users who initially interacted with the chatbot accessed PrEP services, indicating lower effectiveness in promoting PrEP initiation compared to other digital strategies such as social media and dating apps [45].

Despite reported positive outcomes of chatbot-based interventions from the reviewed studies (n=11), several challenges were also consistently identified. A key technical challenge reported across studies is the limited conversational capacity of chatbots, stemming from issues such as misunderstanding user input [45, 46, 108, 110], restricted conversation topics [46, 108, 110], and reliance on simple decision models [46, 108, 110]. These limitations reduce interactivity, personalization, and overall user satisfaction. One-way interaction models, which prevent users from freely asking questions, were also noted to diminish engagement [45, 105, 110]. To address these limitations, several studies called for more intelligent features, including advanced NLP, personalized responses, and mood-adaptive interaction to better meet users' informational and emotional needs [108, 110].

Enhance accessibility and user experience was another priority, several studies emphasized user-centered design strategies [104, 108, 110] such as accommodating varied digital literacy levels [106, 107, 110, 112], simplifying technical language [45], and providing multilingual support [44, 45]. Additionally, support for external resource links – often critical to guiding users toward care – was inconsistently implemented or absent [45, 105]. In addition to conversational improvements, users expressed a desire for chatbots to be accessible across different user interface platforms [44, 45].

From an implementation perspective, multiple challenges hinder scalability and real-world integration. Insufficient team technical capabilities [42, 107, 112] and budget constraints [42] were cited as implementation barriers, particularly in low-resource settings. Multiple studies also highlighted the need for better dissemination strategies to facilitate recruitment [44, 45, 111]. Additionally, studies often relied on pilot-scale samples, raising concerns about generalizability to broader populations [43, 44, 47, 107, 110-112]. Looking ahead, many studies called for larger-scale trials with ongoing monitoring and optimization to assess long-term implementation outcomes and clinical effectiveness [42-44, 47, 106, 107, 110-112].

2.2.2.6 Responsible AI – SHIFT framework

On the topic of sustainable AI, eight research teams demonstrated long-term commitment to their chatbot initiatives by contributing to two or more related articles (see Table 2.1). A good example is the van Heerden team, which completed two versions of the *Nolwazi_bot* chatbot development and conducted tests in 2017 and 2022. This continuity reflects responsible leadership and enhances the long-term viability of their AI interventions. Additionally, all chatbot interventions included in this review were designed to promote health equity and improve access to HIV prevention and care services. These efforts implicitly support social sustainability, particularly in settings with limited healthcare resources and among underserved populations. However, economic and environmental sustainability were rarely addressed across the reviewed articles. Only one study conducted a formal cost-effectiveness analysis of its chatbot intervention and reported the highest cost per participant among all evaluated strategies [109]. No articles assessed or planned to assess the environmental impact associated with chatbot development or deployment.

For human-centeredness, nearly all articles (13/19) explicitly positioned chatbots as assistive technologies designed to support, rather than replace, human care. Most chatbot interventions aimed to enhance access to information [46], initiate dialogue on sensitive health topics such as sexual health or HIV prevention [45, 47, 105, 110], or guide users toward appropriate services [42, 43, 45, 107, 111, 112]. Several included safeguards for human follow-up, particularly in crisis situations involving self-harm or suicidal ideation [47, 105, 110], underscoring a commitment to maintaining human oversight. However, a small number of articles [104, 106, 108] did not clearly define the chatbot's role in relation to healthcare providers.

Inclusive AI emphasizes the importance of participatory design and stakeholder engagement, and cultural adaptation to ensure technologies are responsive to the needs of diverse populations. In this review, several articles integrated participatory design approaches during chatbot development [44, 45, 47, 107, 110-112]. These included iterative workshop sessions [47] and the establishment of advisory committees [42, 45, 47, 107, 110-112] to guide decision-making for chatbot development. Further, Nadarzynski et al. explicitly implemented a co-development process achieved through the patient public involvement approach [106]. Beyond this, two studies and one protocol adopted a user-centered approach [42, 104, 108], while Mathur et al. suggested a user-centric refinement strategy for future development [110]. In contrast, three other articles did not report the use of relevant methods [43, 46, 105]. Multiple chatbots (6/11) have been tailored to local linguistic and cultural contexts to enhance content relevance and accessibility. For example, using isiZulu and English for different age groups in South Africa [105], and Pajubá, a dialect used by Brazilian transgender and queer communities [45]. Other chatbots were developed in Spanish, or in English with plans for future cultural adaptation [104, 108].

No articles reported conducting a formal fairness assessment during data collection or chatbot development processes. Despite the lack of data/model fairness assessments, many chatbots were explicitly designed to reduce health disparities, e.g., by improving access to HIV testing and care among marginalized populations [43-45]. Rupani et al. focused on reducing barriers to mental health care among adolescents with HIV[112], while Nadarzynski et al. used the chatbot to address disparities in STI prevalence and aimed to advance sexual health equity [106].

As Arrieta et al. described, an explainable AI is one that provides details or reasons that make its functioning clear or easy to understand for a given audience [117]. Across the reviewed articles, model explainability assessment was largely absent, with no studies providing detailed explanations of chatbot comprehension or decision errors. Furthermore, data privacy and security practices varied significantly. Some articles explicitly referenced compliance with international standards, such as the General Data Protection Regulation [104, 108], or the Health Insurance Portability and Accountability Act [107]. Others, however, failed to specify any privacy policy or security protocols [42-44, 46, 47, 110-112], raising concerns about the protection of sensitive user data.

2.2.3 Discussion: Key gaps and future priorities

The use of chatbots in HIV care is rapidly expanding. This review identified a modest upward in publications over the past four years (2022: 4; 2023: 3; 2024: 7; 2025 [January to June]: 4). However, only two studies were conducted in low-income countries, despite these regions bearing the greatest HIV burden and standing to benefit most from scalable, low-cost digital health tools such as chatbots [118]. Furthermore, the field skewed toward HIV prevention-focused applications, with only two articles addressing mental health support for adolescents with HIV. This highlights a clear gap across the HIV care cascade [119], particularly in supporting long-term self-management.

The diversity among people living with or at risk of infection – including variation in race, gender identity, age and sexual orientation – underscores the need to design chatbots that are culturally, linguistically, and educationally tailored to marginalized communities most vulnerable to HIV. Yet persistent barriers remain. For instance, despite prior needs assessments, Malaysian users reported that *Haris* still required improvement in incorporating Manglish (Malaysian English) [44]. Similarly, the use of complex medical terminology has been identified as a barrier to *Amanda Selfie*'s accessibility [45]. These challenges emphasize the importance of inclusive AI systems that accommodate diverse literacy levels and language contexts. Participatory design is central to this goal. As observed in many chatbot development processes, deepened patient and stakeholder involvement, such as co-developing and validating content or participating in iterative testing workshops, can improve chatbot relevance, responsiveness, and cultural appropriateness [45, 105, 110].

While most chatbots demonstrated strong feasibility and usability, these strengths were largely driven by their ability to deliver accessible, reliable information. Functionality, however, remains limited. Most chatbots offered shallow, narrowly scoped conversations, which may be insufficient for the needs of long-term self-management. Chronic conditions like HIV require ongoing self-care, behavior change, and psychosocial support that extend beyond episodic, user-initiated dialogues [120]. From a technical perspective, most chatbots in this review were short-term, one-way, or even purely reactive tools with limited NLP capabilities. *Amanda Selfie* was the only chatbot explicitly designed for long-term interaction, with two-way communication and proactive outreach. Still, users reported discomfort when discussing sensitive topics such as daily PrEP use and family dynamics [109]. This suggested that simple proactivity – for example, through PrEP-related reminders – is not enough. Long-term engagement requires more nuanced, context-aware dialogue and greater personalization. Without these features, novelty effects fade, and interactions become predictable or disengaging [121, 122]. Future chatbot development must prioritize broader topic coverage and the integration of more sophisticated AI, e.g., LLMs, to support individualized care.

Empathy is another critical but underdeveloped dimension. Given the stigma and emotional burden associated with HIV, chatbots must convey emotional sensitivity in their interactions. Future development could incorporate sentiment analysis or emotion recognition algorithms to tailor empathetic responses. However, chatbots should not be viewed as substitutes for human care. Clear escalation protocols are essential to ensure appropriate human intervention in emergencies, reinforcing the principle of human-centered AI.

Figure 2.3 maps the reviewed articles across the digital health intervention lifecycle. While implementation and evaluation efforts are increasing, most initiatives remain in early stages, including three pilot studies conducted in simulated environments. Most reviewed articles (10/19) focused on short-term feasibility or usability, with few progressing to large-scale implementation. Future research should prioritize the reporting of longitudinal implementation metrics (e.g., fidelity, adoption, cost-effectiveness [79]) and clinical outcomes. Moreover, the limited use of standardized frameworks to guide implementation assessment and the over-reliance on quantitative methods have constrained insight into real-world effectiveness. Mixed-methods approaches that integrate objective metrics with participant perspectives would provide a more comprehensive understanding of chatbot performance [123, 124]. Theory-informed, longitudinal study designs would be essential to facilitate the sustained assessment of chatbot interventions and support translation of such innovations into routine HIV care.

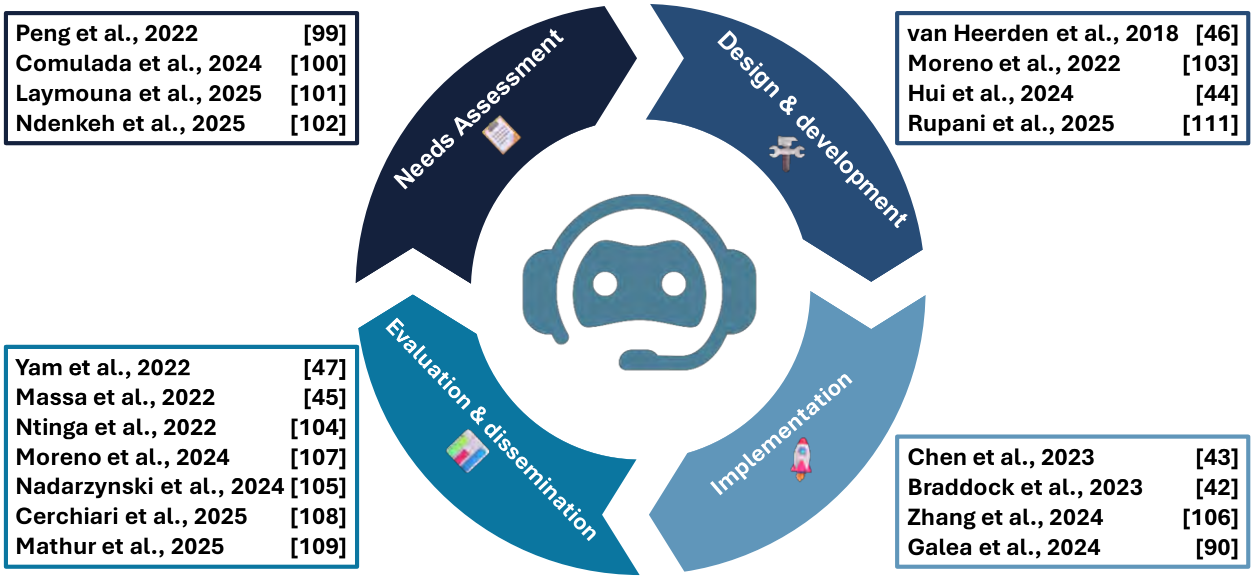

Peng et al., 2022 [99]
Comulada et al., 2024 [100]
Laymouna et al., 2025 [101]
Ndenkeh et al., 2025 [102]
Needs Assessment
Design & development
van Heerden et al., 2018 [46]
Moreno et al., 2022 [103]
Hui et al., 2024 [44]
Rupani et al., 2025 [111]
Yam et al., 2022 [47]
Massa et al., 2022 [45]
Ntinga et al., 2022 [104]
Moreno et al., 2024 [107]
Nadarzynski et al., 2024 [105]
Cerchiari et al., 2025 [108]
Mathur et al., 2025 [109]
Evaluation & dissemination
Implementation
Chen et al., 2023 [43]
Braddock et al., 2023 [42]
Zhang et al., 2024 [106]
Galea et al., 2024 [90]

Figure 2.3 Mapping of reviewed articles across the digital health intervention lifecycle.

Beyond implementation sustainability, the economic and environmental impacts of chatbot development must also be considered. For example, *Amanda Selfie* showed limited economic benefits [45], while Braddock et al. [42] highlighted financial and staffing constraints that hindered the development of more advanced algorithms. Moreover, advanced models require significant computational resources, raising environmental concerns [41]. Training large models for chatbot development can generate considerable carbon emissions, yet few studies report energy use or carbon footprint [125, 126]. Greater transparency in reporting the development process, including the environmental footprint, would help to clarify how responsible, resource-efficient AI development can be carried out and is critical to promoting sustainable AI innovation in healthcare.

Finally, fairness and transparency remain under-addressed. Multiple studies revealed that potential biases in training data or system behavior may inadvertently reinforce inequities [71, 127]. As chatbots move toward more adaptive, AI-driven systems, integrating fairness-aware design and evaluation practices would be essential to avoid reinforcing existing inequities. Furthermore, the absence of model explainability assessment and inconsistent data privacy and security practices limit transparency and trustworthiness. Collectively, these gaps highlight the need for stronger safeguards around fairness, interpretability, and transparency, particularly in real-world deployments.

2.3 LLM application in chronic disease self-management

The rapid advancement of LLMs has significantly transformed the development and capabilities of traditional chatbots. To further expand their utility, more sophisticated algorithms are needed to enable greater personalization, contextual understanding, and clinical relevance. This section reviews recent applications of LLMs in chronic disease self-management, including the more specific context of HIV care, published between January 1, 2021, and June 30, 2025, with the aim of identifying how these tools can support self-management and inform their integration into AI-driven chatbots.

In HIV care, the use of LLMs remains limited. A recent scoping review highlighted gaps in their integration with digital health tools and a lack of engineering-oriented development guidance [128]. Nonetheless, several early efforts have tested their applicability. BERTina is a BERT-based proof-of-concept model capable of answering basic HIV-related questions [129]. Koh et al. evaluated ChatGPT-3.5 for ART counseling and found it delivered accurate, guideline-consistent information but lacked contextualized recommendations, consistency of responses in complex scenarios, and reference to information sources [130]. Similarly, De vito et al. assessed ChatGPT's performance on HIV prevention topics, reporting high accuracy (88.4% of responses scored ≥ 5/6), but noted a lack of socio-political context and inclusivity, underscoring the need for AI tools that that are not only accurate but also equitable and context-aware [131].

Beyond HIV, LLMs have demonstrated broader utility in chronic disease self-management. They have been used to generate tailored health recommendations, simplify clinical language, and support patient education. For instance, GPT-4o significantly improved the readability, correctness, and relevance of cardiology discharge summaries compared to original versions ($p < 0.001$), although personalization remained limited [132]. In musculoskeletal care, GPT-4 generated self-management plans for knee osteoarthritis that outperformed clinician-generated materials in accuracy, personalization, and comprehensiveness, though vocabulary complexity posed challenges for patients with limited health literacy [133].

Furthermore, a recent review of twenty LLM applications in chronic disease management found an overall accuracy of 71% in generating relevant and understandable health recommendations, although performance dropped to 50% for certain conditions such as hepatocellular carcinoma [134]. Reported limitations included hallucinations, limited informational depth, and inability to support long-term behavior change. The review called for domain-specific model fine-tuning, integration of high-quality multimodal data, and rigorous clinical trials to enable the safe and effective deployment of LLMs in chronic care contexts.

Hybrid models are a potential approach that could further improve LLM utility. The CaRiFaM system combined a deep-learning long short–term memory (LSTM) model for cardiovascular risk prediction with GPT-4o to generate personalized recommendations [135]. This hybrid approach achieved 91% accuracy and 92% clinical guideline consistency, with most users reporting that the explanations were clear and useful. Similarly, DeepDR-LLM integrated a transformer-based model for diabetic retinopathy image analysis with an LLM to provide tailored self-management recommendations [136]. In a prospective study, 64% of patients were more likely to follow DeepDR-LLM’s guidance over that of primary care physicians, indicating its potential to enhance patient engagement and promote proactive self-management behaviors.

To summarize, despite the limited body of work to date, these early applications show promising potential for LLMs in chronic disease self-management. Emerging evidence suggests that hybrid approaches – combining chatbots with domain-specific model fine-tuning for targeted tasks – may offer the most effective path forward.

CHAPTER 3 RESEARCH OBJECTIVES AND STRATEGY

3.1 Research objectives

AI-based chatbots are evolving rapidly and there is growing interest in their applications in healthcare. However, their integration into HIV care to support long-term self-management remains limited. These tools offer significant potential to provide accessible, scalable, and personalized support for PWH as they navigate complex self-management tasks. Realizing this potential requires moving beyond short-term, narrowly scoped tools. It means pursuing solutions that are co-developed with PWH, enhanced by LLMs to enable more intelligent and personalized support, and grounded in implementation science to ensure they are deployed in ways that are usable, equitable, and clinically relevant.

To address these opportunities and challenges, this thesis seeks to answer the following overarching research question:

What are the key considerations in designing, developing, and implementing an AI-based chatbot to support the self-management of people with HIV in real-world care settings?

This investigation is guided by the following research objectives:

Primary objective [PO]: Design and Develop a chatbot-based personalized healthcare tool to support self-management among PWH using participatory design approach.

- **Sub objective 1 [SO1]:** Design and develop a methodological framework to guide the co-development, implementation, and evaluation of the chatbot.
- **Sub objective 2 [SO2]:** Co-design and co-develop a chatbot that supports the daily self-management needs of PWH.
- **Sub objective 3 [SO3]:** Assess the real-world feasibility and usability of the developed chatbot among PWH.
- **Sub objective 4 [SO4]:** Design, develop, and validate an LLM-based triage tool to personalize the chatbot’s support.

3.2 Strategy and methodology

3.2.1 Participatory design approach

This patient-oriented project combines expertise in medicine, pharmacy, lived experience of HIV, design engineering, and software development. Given its multidisciplinary scope and patient-centered focus, the project adopted a participatory design approach throughout all phases [137]. A design committee – comprising physicians, pharmacists, social workers, community organizations, researchers, and, most importantly, PWH – was established to guide the participatory design process.

The committee was actively involved throughout the project, from the initial identification of user needs and the evaluation of design parameters to the final assessment of the chatbot. Meetings were held every two to three months, typically lasting 60 minutes, with additional email communication as needed to address specific questions between the research team and committee members. This collaborative structure not only ensured that patient needs were consistently prioritized but also contributed to addressing equity and diversity considerations, ultimately with the goal of supporting the development of a more inclusive and effective chatbot intervention.

3.2.2 Digital health intervention lifecycle

The design and development of the chatbot followed the digital health intervention lifecycle introduced in Section 2.2.1.1, comprising four iterative phases: needs assessment, design and development, implementation, and evaluation [89, 94]. While the initial needs assessment was conducted prior to this PhD project [138, 139], the three articles forming this thesis are each aligned with one or several phases of this lifecycle. This framework provided a structured foundation for the chatbot development strategy and ensured that the work adhered to best practices in digital health intervention development.

3.2.3 Responsible AI – SHIFT framework

The chatbot development was also informed by the SHIFT framework for responsible AI [92], described in Section 2.2.1.3. This framework outlines five guiding principles for responsible health AI: sustainable, human-centered, inclusive, fair, and transparent. These principles served as a values-based lens for guiding chatbot development choices and supported the creation of an AI intervention that prioritized social responsibility alongside technical performance.

3.3 Summary of articles

The following section provides a brief overview of the three articles that form the core of this thesis. Each article addresses a different aspect of developing an AI-based chatbot to support HIV self-management. Together, they span the entire development lifecycle of a digital health intervention.

Article 1: Adapting and Evaluating an AI-Based Chatbot Through Patient and Stakeholder Engagement to Provide Information for Different Health Conditions: Master Protocol for an Adaptive Platform Trial (the MARVIN Chatbots Study) – SO1 and SO2

To address SO1, the first manuscript introduces a master protocol based on a three-phase adaptive platform trial design to guide the co-development, implementation and evaluation of MARVIN across different clinical contexts. Phase 1 involves co-construction through needs assessments, interdisciplinary workshops, and iterative development. Phase 2 employs mixed methods to evaluate usability and acceptability using validated instruments and group interviews informed by the Technology Acceptance Model (TAM) and the Nonadoption, Abandonment, and challenges to the Scale-up, Spread, and Sustainability (NASSS) framework. Phase 3 assesses real-world implementation by measuring usability, acceptability, appropriateness, adoption, and fidelity, and triangulating these findings with qualitative user feedback analyzed through the NASSS framework. Longitudinal stakeholder interviews are conducted throughout to assess the impact of their involvement. From July 2022 to October 2023, this approach supported four substudies (HIV, pharmacy-based HIV care, breast cancer, and pediatric infectious diseases). As the first digital health master protocol of its kind, it supports scalable chatbot development and real-world evidence generation, while accelerating patient access to advanced chatbot interventions and remaining responsive to evolving AI and regulatory landscapes through centralized management.

This manuscript also addresses SO2 by presenting the methodology used to design and develop MARVIN, an AI-based chatbot designed to support self-management in PWH. Co-developed with PWH and other stakeholders, MARVIN provides self-management support across over 30 topics. Its NLP system includes a transformer-based model for entity recognition and intent classification, a *FallbackClassifier* to manage incomprehensible inputs, and a hybrid dialogue management module combining memory-based and rule-based strategies. Responses are drawn from a knowledge base co-created with PWH and stakeholders. MARVIN has been deployed on Facebook Messenger, with its data confidentiality and security protocols detailed in the master protocol. In addition, an external regulatory committee has been established to support ethical oversight and respond to emerging challenges in health AI.

Article 2: The First AI-based Chatbot to Promote HIV Self-Management: A Mixed Methods Usability Study – SO3

The second manuscript addresses SO3 through a mixed-methods study assessing MARVIN's real-world feasibility, usability, and acceptability among PWH. Over a 3-week pilot, 28 participants were asked to engage in 20 conversations with the chatbot on predefined self-management topics, yielding a 70% retention rate 28/40). Mean usability scores exceeded the threshold of success (69.9/68), and mean acceptability was very close to target (23.8/24). Participants gave positive ratings across TAM subconstructs, including perceived ease of use, perceived usefulness, attitude towards use, and behavioral intention to use. Qualitative findings indicated that MARVIN was perceived as easy to use, emotionally safe, confidential, and a trusted source of real-time expert-validated advice. Noted limitations included challenges in chatbot comprehension, limited topic coverage, lack of memory features, alongside concerns about reliance of Facebook Messenger. Overall, the study confirms MARVIN's feasibility and usability in real-word settings and highlights key areas for improvement.

Article 3: Large Language Model-Based Triage to Identify Antiretroviral Therapy Adherence Barriers and Risks in Patient Messages – SO4

To address SO4, the third manuscript focused on ART adherence barriers, a crucial self-management challenges for PWH. The study investigated optimal strategies for designing and developing an LLM-based triage solution to identify ART adherence barriers (e.g., thoughts and feelings, habits, social or economic situations) and associated risk levels (high, medium, low, or none) from unstructured patient messages. Fine-tuned Flan-T5-xl achieved the best performance for barrier detection (Macro-F1=0.83), while Flan-T5-large performed best for risk level detection (Macro-F1=0.80). Fine-tuned general-domain models significantly outperformed clinical foundation models (*P*<0.001; Δ Macro-F1: -0.02 to -0.11) and GPT-4.1 and other large-scale open-source LLMs in zero-/five-shot settings (P<0.001; Δ Macro-F1: -0.07 to -0.38). To promote equitable and sustainable healthcare AI, the study conducted further fairness analyses and measured energy consumption and carbon footprints. Results showed that fine-tuned models demonstrated better reliability in minimizing disparities across sociodemographic descriptors and consumed approximately 30 times less energy than GPT-4.1. These findings highlighted the potential of LLM-based tools for adherence monitoring and the need for careful model selection, bias mitigation, and environmental sustainability in healthcare AI development.

CHAPTER 4 ARTICLE 1: ADAPTING AND EVALUATING AN AI-BASED CHATBOT THROUGH PATIENT AND STAKEHODLER ENGAGEMENT TO PROVIDE INFORMATION FOR DIFFERENT HEALTH CONDITIONS: MASTER PROTOCOL FOR AN ADAPTIVE PLATFORM TRIAL (THE MARVIN CHATBOTS STUDY)

Yuanchao Ma[1,2,3,4], Sofiane Achiche[1], Marie-Pascale Pomey[5,6,7], Jesseca Paquette[5], Nesrine Adjtoutah[5,6], Serge Vicente[2,3,8,9], Kim Engler[2,3], MARVIN chatbots Patient Expert Committee[2,10]; Moustafa Laymouna[2,3,8], David Lessard[2,3,4], Benoît Lemire[4], Jamil Asselah[11], Rachel Therrien[5], Esli Osmanlliu[2,12], Ma'n H Zawati[13], Yann Joly[13], Bertrand Lebouché[2,3,4,8]

1. Department of Biomedical Engineering, Polytechnique Montréal, Montreal, QC, Canada
2. Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre, Montreal, QC, Canada
3. Infectious Diseases and Immunity in Global Health Program, Research Institute of McGill University Health Centre, Montreal, QC, Canada
4. Chronic Viral Illness Service, Division of Infectious Disease, Department of Medicine, McGill University Health Centre, Montreal, QC, Canada
5. Research Centre of the University of Montreal Hospital Centre, Montreal, QC, Canada
6. Department of Health Policy, Management and Evaluation, School of Public Health, University of Montreal, Montreal, QC, Canada
7. Centre of Excellence on Partnership with Patients and the Public, Montreal, QC, Canada
8. Department of Family Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
9. Department of Mathematics and Statistics, University of Montreal, Montreal, QC, Canada
10. See Acknowledgements
11. Department of Medicine, Division of Medical Oncology, McGill University Health Centre, Montreal, QC, Canada
12. Department of Pediatrics, Montreal Children's Hospital, McGill University Health Centre, Montreal, QC, Canada
13. Centre of Genomics and Policy, McGill University, Montreal, QC, Canada

Published in *JMIR Research Protocols*, Volume 13, e54668, February 2024

4.1 Abstract

Background: Artificial intelligence (AI)–based chatbots could help address some of the challenges patients face in acquiring information essential to their self-health management, including unreliable sources and overburdened health care professionals. Research to ensure the proper design, implementation, and uptake of chatbots is imperative. Inclusive digital health research and responsible AI integration into health care require active and sustained patient and stakeholder engagement, yet corresponding activities and guidance are limited for this purpose.

Objective: In response, this manuscript presents a master protocol for the development, testing, and implementation of a chatbot family in partnership with stakeholders. This protocol aims to help efficiently translate an initial chatbot intervention (MARVIN) to multiple health domains and populations.

Methods: The MARVIN chatbots study has an adaptive platform trial design consisting of multiple parallel individual chatbot substudies with four common objectives to: (1) co-construct a tailored AI chatbot for a specific health care setting, (2) assess its usability with a small sample of participants, (3) measure implementation outcomes (usability, acceptability, appropriateness, adoption, and fidelity) within a large sample, and (4) evaluate the impact of patient and stakeholder partnerships on chatbot development. For objective 1, a needs assessment will be conducted within the setting involving four 2-hour focus groups with 5 participants each. Then, a co-construction design committee will be formed with patient partners, health care professionals, and researchers who will participate in 6 workshops for chatbot development, testing, and improvement. For objective 2, 30 participants will interact with the prototype for 3 weeks and assess its usability through a survey and 3 focus groups. Positive usability outcomes will lead to the initiation of objective 3, whereby the public will be able to access the chatbot for a 12-month real-world implementation study using web-based questionnaires to measure usability, acceptability, and appropriateness for 150 participants and meta-use data to inform adoption and fidelity. After each objective, for objective 4, focus groups will be conducted with the design committee to better understand their perspectives on the engagement process.

Results: From July 2022 to October 2023, this master protocol led to four substudies conducted at the McGill University Health Centre or the Centre hospitalier de l’Université de Montréal (both in Montreal, Quebec, Canada): (1) MARVIN for HIV (large-scale implementation expected in mid-2024), (2) MARVIN-Pharma for community pharmacists providing HIV care (usability study planned for mid-2024), (3) MARVINA for breast cancer, and (4) MARVIN-CHAMP for pediatric infectious conditions (both in preparation, with development to begin in early 2024).

Conclusions: This master protocol offers an approach to chatbot development in partnership with patients and health care professionals that includes a comprehensive assessment of implementation outcomes. It also contributes to best practice recommendations for patient and stakeholder engagement in digital health research.

Trial Registration: ClinicalTrials.gov NCT05789901
https://classic.clinicaltrials.gov/ct2/show/NCT05789901

4.2 Introduction

4.2.1 Background

Self-management is key to patient health, and interventions to promote it are being increasingly implemented in the delivery of health care. Effective self-management involves multiple aspects, including problem-solving, decision-making, resource use, patient–health care professional partnership building, and taking action [1,2]. To improve self-management, patients often seek guidance from health care professionals, staff and volunteers from community organizations or peers, and a variety of internet resources [3,4]. However, the availability of these actors and resources can be inconsistent. Moreover, information encountered on the internet varies in quality [5-7]. From reliable but complex scientific articles to opinion blogs or outdated web pages, these sources do not always provide the clearest, most accurate or reassuring guidance. The importance of self-management has been further highlighted by the COVID-19 pandemic, which introduced additional barriers to access regular follow-up [8] and exposed individuals worldwide to an infodemic [9,10].

In addition to these challenges, frontline professionals are often confronted with complex questions from patients regarding treatment instructions, comorbidity management, side effects, and drug interactions. However, in the context of swiftly evolving knowledge and overwhelming workloads, health care professionals may not have sufficient expertise and time for certain questions [11,12]. Their responses may vary depending on their individual training, proficiency, and clinical experience. Rapidly obtaining accurate and clear health information and relaying it to patients can be a daunting challenge for professionals.

Safe and effective digital health interventions could help address the limitations of existing self-management or care support. A particularly promising avenue is the emergence of artificial intelligence (AI)–based chatbots—software applications that interact with users through simulated human text or voice conversations via smartphones or computers. Often harnessing AI to enable natural language interpretation and assist in decision-making, such tools possess the potential to revolutionize patient self-management and support [13]. They can be applied to diverse platforms to foster mutually beneficial outcomes for health systems and patients, including less time spent in hospitals, outpatient efficiency, and personalized treatment [9].

Successful implementation of an intervention (ie, positive implementation outcomes) is necessary to achieve desired changes in clinical or service outcomes [14]. Multiple studies have investigated the implementation of health-oriented chatbots, including in the areas of mental health support [15-17], problematic substance use treatment [18], cirrhosis patient education [19], and asthma self-management [20]. Nevertheless, such studies have typically focused on feasibility and usability in small-scale prototype implementations [22-24]. Digital health products are often dealt with through an "implement now, clinically validate later" ethos [25] as they encompass a large number of different technologies. Thus, there is no clear consensus on methods for assessing the clinical effectiveness of digital health interventions [25], and data on the impact of chatbot interventions on clinical outcomes are scarce [26]. Large user samples are necessary to gain more robust insights into chatbot implementation. In-depth studies including other valuable implementation (eg, fidelity, appropriateness, sustainability, and cost-effectiveness) and clinical (eg, safety, effectiveness, and efficacy) outcomes will also be critical for scaling up and long-term adoption of chatbot interventions.

Finally, very little engagement of stakeholders, including patients and health care professionals, has led to poor usability and low adoption of many digital health interventions [27]. In the context of chatbot development, a scoping review investigating patient engagement revealed limited involvement of patients and insufficient reporting of the relevant activities [28]. Among the 16 studies included, only 8 mentioned patient engagement, with just 3 offering adequate details on the methods and approach used. The authors also pointed out that future chatbot development would need to integrate multifaceted means of patient participation and document them thoroughly. Stakeholders should be engaged in defining research objectives and designing interventions tailored to their needs [29]. According to the patient-public partnership continuum proposed by the Montreal model [30], this inclusivity can be extended to “co-construction,” where patients and stakeholders are involved throughout the process. Meanwhile, to answer the many questions raised by the use of intelligent machines and ensure that AI develops in harmony with democracy, the responsible integration of AI necessitates a co-construction process [31]. Engaging end users in discussions about the challenges posed by AI and drawing on their lived experiences can reveal key aspects of digital health research that might otherwise be overlooked [32], thus better paving the way for success.

Since 2020, led by YM, SA, and B Lebouché, an innovative chatbot named *Minimal AntiRetroViral INterference* (MARVIN) has been in development for people with HIV. Through a co-design approach involving patients and stakeholders, MARVIN aims to facilitate antiretroviral therapy self-management. The authors’ team subsequently trialed the MARVIN chatbot among people with HIV and validated its usability and acceptability [21]. Given the initial success of MARVIN among people with HIV, we intend to build on this pilot study to increase MARVIN’s areas of specialization and continue to develop our algorithms to improve MARVIN’s intelligence, thereby expanding its reach and potential benefits to a broader audience.

4.2.2 Aim and objectives

Grounded in patient and stakeholder engagement strategies and implementation science, this protocol aims to describe the methods and tools necessary to efficiently develop the innovative MARVIN chatbot interventions across multiple health domains and populations and assess their implementation for robust and widespread use. The study’s primary objectives are to co-construct versions of MARVIN adapted for different health conditions and target populations (objective 1, development), assess their global usability (usability and acceptability) in a small participant sample context (objective 2, usability), and measure implementation outcomes (usability, acceptability, appropriateness, adoption, and fidelity) in the context of a large sample (objective 3, implementation) through a mixed methods approach in the respective setting of each chatbot. The secondary objective is to evaluate the impact of different stakeholder partnerships established for the aforementioned objectives on the development of the AI health care chatbots (objective 4, partnership evaluation).

4.3 Methods

4.3.1 Study design

This multicenter study follows the CONSORT (Consolidated Standards of Reporting Trials) extension for pilot and feasibility trials [33], CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) [34], and CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) [35] guidelines (Multimedia Appendices 1-3 [33]), as well as the Montréal Declaration for a Responsible Development of Artificial Intelligence [31] regarding the ethical conduct of research involving humans and AI.

The study is presented in the form of a master protocol, defined as a single overarching design developed to evaluate multiple hypotheses with the overall goal of increasing efficiency and establishing uniformity through the standardization of procedures in the development and evaluation of different interventions [36]. This master protocol will encompass multiple parallel individual chatbot projects sharing the MARVIN chatbot technology, subsequently referred to as "substudies." The chatbot of each substudy is considered as a distinct intervention as it will integrate specific features or content tailored to the corresponding health care setting (health condition and target populations). The chatbots will be implemented in these different populations without control groups.

To accommodate this, we used an adaptive platform trial design, which combines features of both basket trials (designed to test a single intervention in different populations) and platform trials (designed to test multiple interventions in the context of a single disease) [37]. As defined by the US Food and Drug Administration, an adaptive platform design is appropriate for our trial as it allows for flexibility in managing multiple interventions adapted to different populations while enabling the early removal of ineffective interventions and introduction of new interventions based on interim data [38].

As shown in the example in Figure 4.1, substudy arms targeting different conditions can be initiated at different time points. The substudies are independent of each other, and their processes are illustrated in Figure 4.2. Objectives 1 to 3 will be completed in a sequential manner, whereas objective 4 will be assessed throughout the process. A decision will be made on whether to continue with the same chatbot intervention version in objective 3 (implementation) based on the results of objective 2 (usability). In addition, there is no initial fixed duration or sample size for each substudy.

This master protocol outlines the common elements of the individual chatbot substudies in terms of objectives and processes. Concurrently, each substudy will have its own subprotocol describing more specific standardized operational structures; additional inclusion and exclusion criteria; recruitment and selection methods; and data collection, analysis, and management. All subprotocols will be managed as appendices to this master protocol and will be subject to further review by the research ethics board (REB) as they are ready. In addition, certain criteria of the master protocol may be redefined by the research team based on progress in each substudy and submitted as amendments to the REB for review for subsequent application to all substudies.

4.3.2 Ethical considerations

This study received approval from the McGill University Health Centre (MUHC) REB on August 9, 2023 (approval MP-37-2023-9333); the Centre hospitalier de l’Université de Montréal REB (approval MEO-37-2024-11732) on October 13, 2023; and the Polytechnique Montréal REB (approval CER-2324-29-D) on October 10, 2023, and is registered in the ClinicalTrials.gov database (NCT05789901).

4.3.3 Settings and participants

This study will be conducted at the MUHC and the Centre hospitalier de l’Université de Montréal, both located in Montreal, Quebec, Canada. The study participants include patients and health care professionals, who will be the end users of the chatbots.

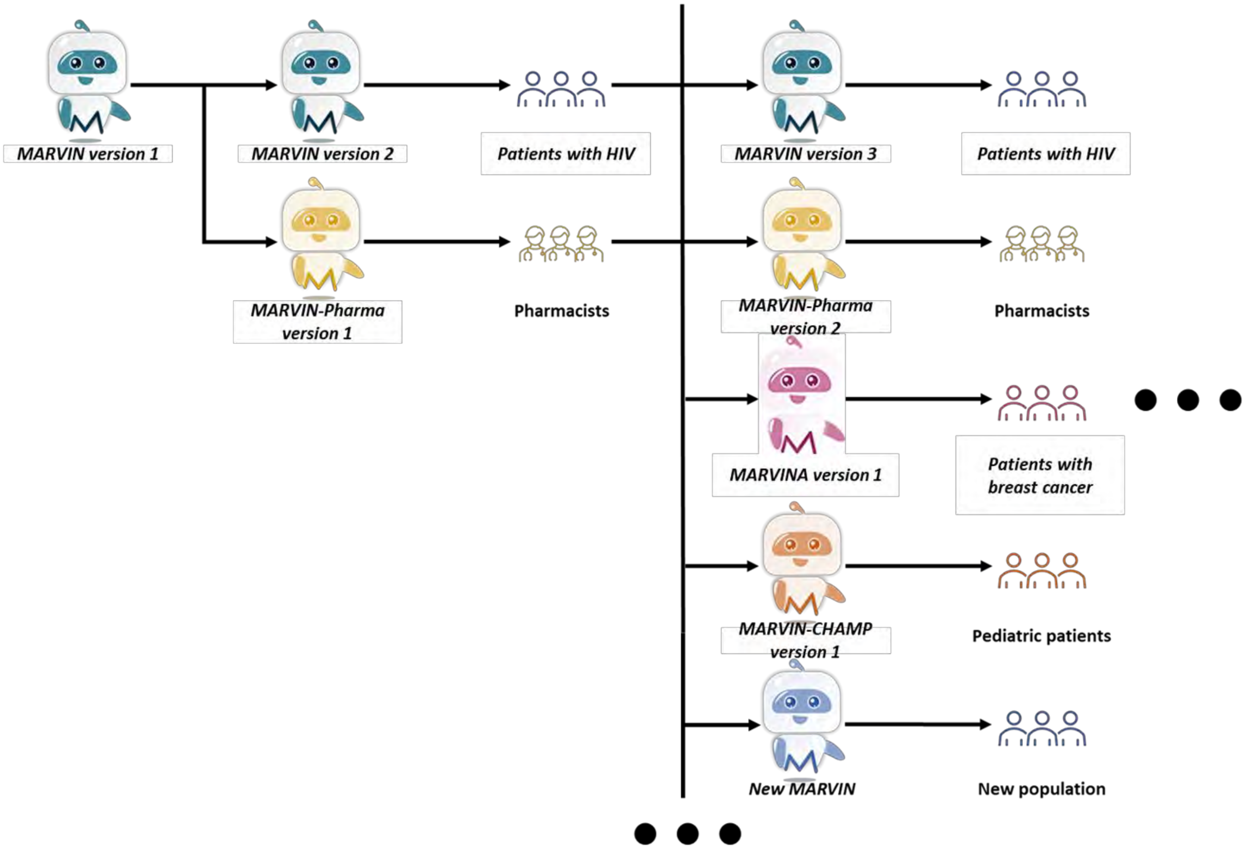


Figure 4.1 Adaptive platform trial design without control group.

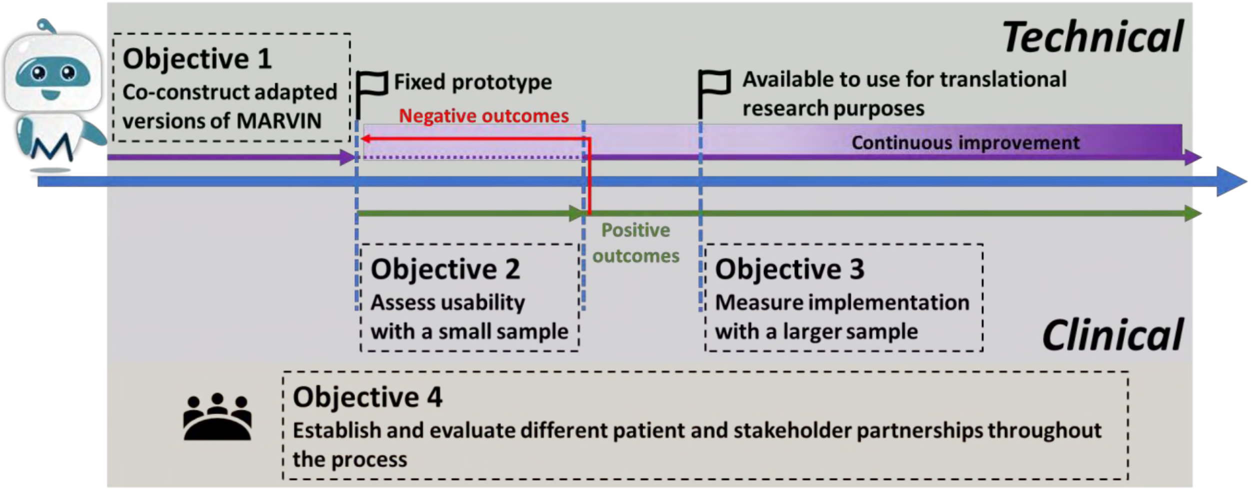

Technical
Objective 1
Co-construct adapted versions of MARVIN
Fixed prototype
Negative outcomes
Available to use for translational research purposes
Continuous improvement
Positive outcomes
Objective 2
Assess usability with a small sample
Objective 3
Measure implementation with a larger sample
Clinical
Objective 4
Establish and evaluate different patient and stakeholder partnerships throughout the process

Figure 4.2 Study steps for each substudy.

4.3.4 Eligibility criteria

The primary inclusion criteria for participants are as follows: (1) age of ≥18 years; (2) fluency in English or French; (3) ability to understand the requirements of study participation and provide oral and written informed consent before and during the implementation of the study; (4) access to a smartphone, tablet, or computer in a private environment; and (5) access to an internet connection or data plan on their device. Specific to objectives 2 (usability) and 3 (implementation), additional inclusion criteria are (6) acceptance of using or creating a personal Facebook (Meta Platforms) account, (7) acceptance of using a Facebook Messenger-based chatbot, and (8) acceptance of Facebook's privacy and data security policies.

Participants may not take part if they (1) are affected by a cognitive deficit or medical instability that prevents them from participating in any aspect of the study and (2) self-report being insufficiently able to use the chatbot with the technical support provided. For objectives 2 (usability) and 3 (implementation), the patient partners involved in the co-construction design committee (defined in the following section) will not be able to participate in either phase.

4.3.5 Recruitment and sample size justification

Objectives 1, 2, and 3 will use convenience sampling to recruit participants through different channels, including clinics, patient foundations, community-based organizations, and professional associations [39]. For example, patient participants will be introduced to the study by their health care service provider (eg, physician, nurse, or social and community worker) during their visits. Informational materials such as study flyers and a video (Multimedia Appendix 4) showcasing the MARVIN chatbot will be disseminated via email, newsletters, or a website.

Participants will be required to consent before engaging in each objective. For objectives 1 (development) and 2 (usability), interested individuals can reach out to the study coordinator. Detailed study procedures, eligibility checks, and consent collection will be facilitated by the study coordinator for those expressing interest. Verbal consent may be adopted for participants who continue from objective 1 to objective 2. This will be obtained remotely through teleconference with an impartial witness to ensure compliance with all aspects of free and informed consent. Following the co-construction approach [31], a co-construction design committee consisting of researchers, health care professionals, and volunteer patient partners will be formed starting from objective 1 to carry out the subsequent development. The sample size for objective 1 is 20 participants to ensure saturation for the needs assessment focus groups [40], whereas 30 participants will be recruited to complete the corresponding usability test for objective 2. This sample size is common for pilot studies and satisfies the minimum size recommended in the literature [41,42].

For objective 3 (implementation), an optimized version of the chatbot will be accessible from the web around the clock, 7 days a week for translational research. As currently the MARVIN chatbots will be only released in Canada, recruitment will be exclusive to the Canadian population, and participants will engage in an electronic consent process through the MARVIN chatbots. Before using the chatbot, individuals will be required to review and accept MARVIN’s privacy policy (Multimedia Appendix 5). Subsequently, they will be asked to review an electronic version of the information and consent form and then answer the following verification questions for eligibility criteria 1 to 3 via the chatbot: (1) *Are you at least 18 years old?* (2) *Are you comfortable using English or French while communicating with the chatbot?* (3) *Do you agree to participate in the study as described above?* The remaining criteria (4 to 8) will be met once users connect to the chatbot. Should participants agree to participate, their responses will be securely recorded in a separate encrypted database on the MARVIN cloud servers and synchronized to the electronic enrollment log. If users opt not to participate, no records will be kept. As a Facebook account will be a prerequisite for using the chatbot, no further measures will be taken to detect or prevent the possibility of multiple identities. The target sample size of 30 to 150 participants was obtained based on the usability, acceptability, and appropriateness outcomes, which are considered key outcomes for this objective. The analysis involves a 1-sided Student *t* test (1-tailed) evaluating whether the corresponding average attains a predetermined threshold. A power analysis for a 1-sample *t* test is then performed, with 80% power and a 5% significance level. Within the targeted sample size ($30 \leq n \leq 150$), small to moderate standardized effect sizes ($0.2 \leq$ Cohen $d \leq 0.5$) are detectable in the total sample with the aforementioned statistical power.

Regarding objective 4 (partnership evaluation), upon completion of each objective, an email invitation will be sent to organize focus groups with stakeholders involved in the chatbot co-construction design committee.

The information and consent forms outline detailed study procedures, anticipated benefits, and potential risks. All template versions are available in Multimedia Appendix 6. Participants may withdraw from the study at any time after providing informed consent. This information will be recorded in the electronic enrollment log, and the related privacy management and protection measures will be detailed later. To protect the participants' personal data and identity, no identifying information will appear in any manuscript or report from this study.

4.3.6 Intervention: Status Quo of the MARVIN Chatbot

4.3.6.1 Overview

Running on Messenger 24/7 for free [43], MARVIN was created as a bilingual chatbot in both English and French trained to converse with people with HIV on the following self-management aspects: (1) guidance for antiretroviral therapy medication (with regard to time management, dosing, drug interactions, and medication reminders, among other things), (2) antiretroviral therapy management when traveling, and (3) common HIV-related knowledge (eg, symptoms, modes of transmission and prevention, and vaccination recommendations).

The team conducted the first pilot project (MUHC REB 2021-7191) in April 2021 with 28 people with HIV receiving treatment at the MUHC. The study results showed that MARVIN was tailored to patient requirements and was easy to use and approachable but that the chatbot's comprehension had limits [21]. For example, if a patient asked a question that was outside the range of topics in the question bank or was worded differently, the chatbot did not always understand it. Nonetheless, considering the development phase, participants reported being satisfied with MARVIN and mentioned that they intended to use it. Thus, by talking to the on-call chatbot, people with HIV could obtain the information they needed for self-management.

4.3.6.2 AI algorithms and strategies

MARVIN is currently being developed using the Rasa platform (Rasa Technologies Inc), an open-source machine learning framework for automating text- and voice-based virtual assistants [44]. As shown in Figure 4.3, MARVIN's architecture comprises 3 distinct modules: natural language understanding, dialogue management, and response selection. Together with a self-built knowledge database, these modules are used to process the message input and generate the message output.

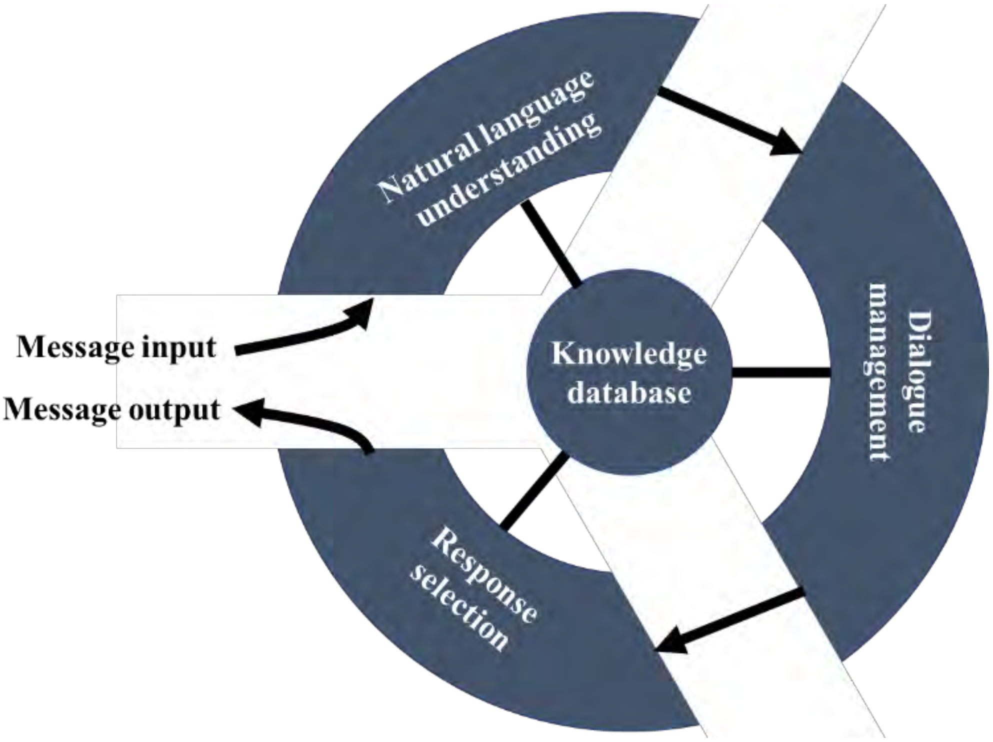
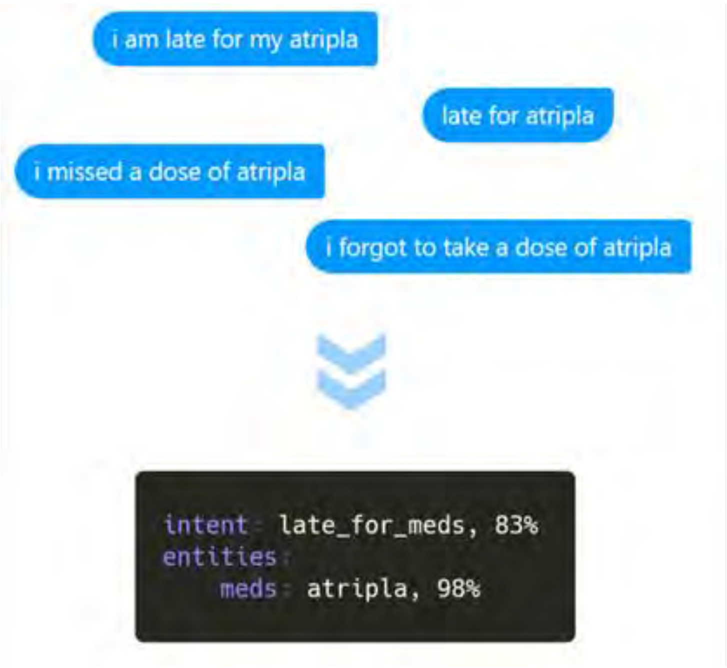


Figure 4.3 Operation process of MARVIN.

The acceptable input data include natural language in text form as well as some auxiliary expressions commonly used for chatting (eg, emojis such as the thumbs up, smiley face, and sad face). The initial natural language understanding module allows MARVIN to semantically process the input messages through different algorithms such as text preprocessing, entity extraction, and intent classification. The current training data set encompasses a corpus of >3000 questions covering >30 topics. Figure 4.4 shows an example of text processing. The entity extraction mechanism enables the chatbot to obtain structured information (eg, date, time, and medication name) from the input messages, whereas intent classification helps MARVIN discern the purpose of the information received.

In particular, MARVIN adopts the *FallbackClassifier* algorithm as a fail-safe measure for unprocessable input forms such as images, videos, and sounds. It is also activated when the confidence level of the anticipated intent falls below an established threshold. If the input remains incomprehensible after 2 attempts, the chatbot responds in a uniform manner: “I’m sorry, I can’t understand your question right now. Please contact a healthcare professional if you need to.” As illustrated in Figure 4.5, this approach reduces the probability of the chatbot taking incorrect actions when confronted with ambiguity.



i am late for my atripla

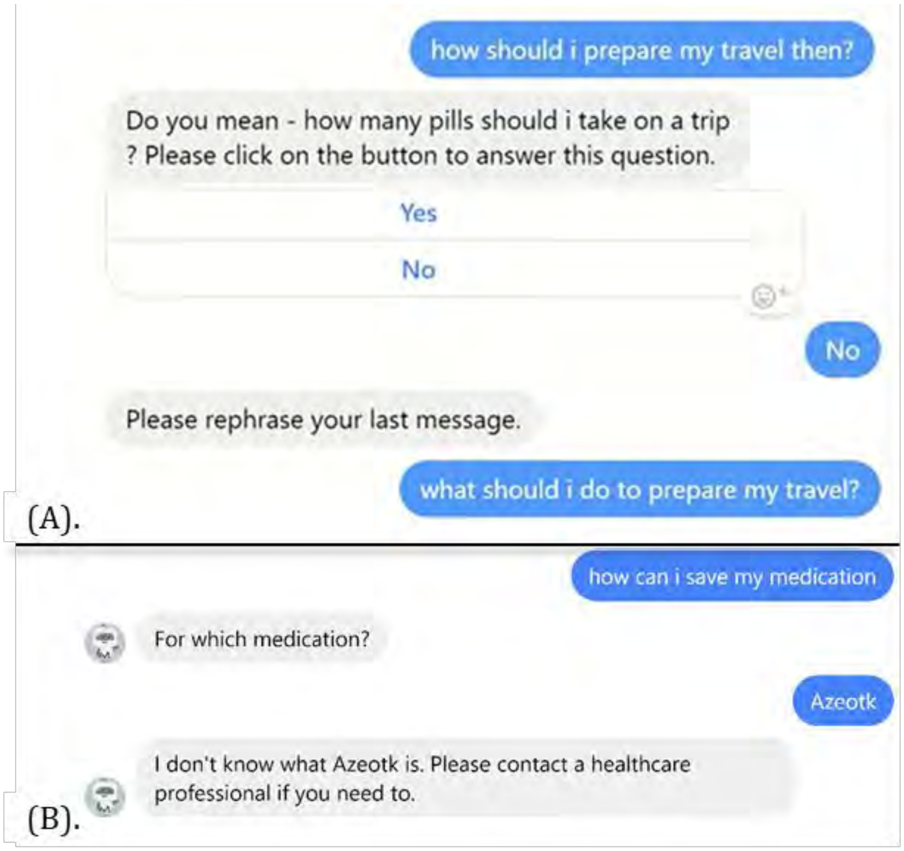
late for atripla

i missed a dose of atripla

i forgot to take a dose of atripla

```
intent: late_for_meds, 83%
entities:
    meds: atripla, 98%
```

Figure 4.4 Example of text processing in MARVIN.



how should i prepare my travel then?

Do you mean - how many pills should i take on a trip ? Please click on the button to answer this question.

No

Please rephrase your last message.

what should i do to prepare my travel?

how can i save my medication

For which medication?

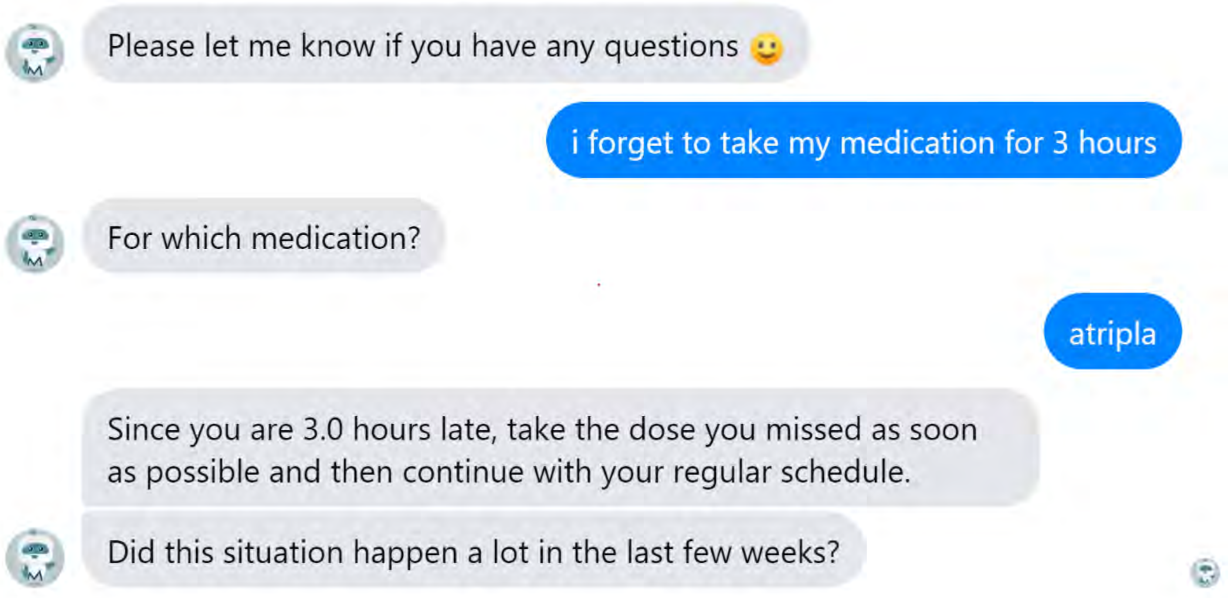
Azeotk

I don't know what Azeotk is. Please contact a healthcare professional if you need to.

Figure 4.5 Examples of MARVIN handling bad inputs: (A) fallback policy; (B) redirection after 2 failed attempts.

The machine learning–driven conversation management determines MARVIN’s subsequent actions based on the input message and context in the conversation. A hybrid strategy of the *memoization policy* and *rule policy* was adopted [45]. The *memoization policy* remembers the decision trees from the training data and predicts the next action in conjunction with the derived intention: asking clarifying questions, providing tailored answers, and taking a fallback action. MARVIN can also remember and analyze a certain number of rounds of dialogue to support the prediction. Meanwhile, the *rule policy* handles pieces of conversation that should always follow the fixed behavior defined in the training data (eg, question: *What is ART?* answer: *It means AntiRetroviral Therapies*).

The final response selection module enables MARVIN to select messages predefined by the health care team as output messages. All the data required for processing by the first 3 modules are stored in the knowledge database—after each decision is made in the processing module, MARVIN communicates with the database to compare, extract, or save the data. All data are selected, edited, and validated by the medical team and then processed and added to the database by the development team. Data types include the previously mentioned antiretroviral therapy management–related information, decision trees for different problem scenarios, and predefined answers. An example of response selection from predefined answers in MARVIN is shown in Figure 4.6.



Please let me know if you have any questions 🙂

i forget to take my medication for 3 hours

For which medication?

atripla

Since you are 3.0 hours late, take the dose you missed as soon as possible and then continue with your regular schedule.

Did this situation happen a lot in the last few weeks?

Figure 4.6 Conversation example with MARVIN.

4.3.6.3 Content development and improvement process

To train MARVIN to understand different types of questions and provide relevant and accurate answers, we assembled a multidisciplinary team of 2 clinicians (ML and B Lebouché), 2 pharmacists (B Lemire and RT), 4 patients, 2 engineers (YM and SA), and multiple developers. Each group's expertise played a crucial role in chatbot content development and continuous improvement: (1) health care professionals validated the medical accuracy of the answers and information given by the chatbot; (2) together with patient partners, they collaborated to identify common self-care challenges; (3) engineers then assisted in building decision models and preparing chatbot training data based on identified problems; and (4) finally, engineers and patients collaborated on testing and refining the chatbot based on feedback, completing the interdisciplinary cycle. This multidisciplinary team allowed us to ensure the accuracy and quality of the information provided as well as conduct holistic research and ongoing evaluation of the chatbot to promote its long-term usability.

4.3.6.4 The interface

Messenger has 1.3 billion monthly active users worldwide exchanging 8 billion messages per day [46]. This indicates a robust familiarity with the Messenger interface among the population, simplifying their grasp of the platform's specifications and interaction capabilities, ultimately favoring the uptake of the MARVIN chatbots. In addition, the versatility of Messenger, compatible across smartphones, tablets, and desktops, reduces the potential risk of participant exclusion because of hardware limitations. Nevertheless, the privacy concerns brought up with Facebook in the past may make people hesitant, which could be one of the potential barriers to the implementation of Messenger for health purposes. Indeed, other user interface options such as a stand-alone mobile app, a web page, or other third-party applications (eg, Telegram or Instagram) are under consideration. As the study proceeds and evolves, relevant changes will be implemented as necessary.

4.3.7 Main study process

4.3.7.1 Overview

Multimedia Appendix 7 illustrates the entire study process.

4.3.7.2 Objective 1: Co-construction of the MARVIN Chatbots

4.3.7.2.1 Overview

Objective 1 is intended to obtain a stabilized version adapted to the health care context for the subsequent objectives. Similar to the original MARVIN's development, it will follow 3 steps: user needs assessment, knowledge database creation, and continuous improvement of the prototype.

The expected duration of participant involvement in objective 1 is 4 months.

4.3.7.2.2 Step 1: Needs assessment

Four 2-hour focus groups will be organized with 5 participants each led by a trained interviewer. Participants will first be shown a demonstration of the current MARVIN describing the interface, dialogue process, and other instructions for use. Semistructured focus groups will then be conducted to identify use scenarios, topics related to integration, and user expectations and preferences for chatbots.

The co-construction design committee will then be formed and meet every 2 weeks to participate in design tasks. The team may also contact them via email with specific questions.

4.3.7.2.3 Step 2: Knowledge database creation

A knowledge database will be built based on the results of the needs assessment, which is the core of each new chatbot. This database will include a bank of different questions, plausible answers, and necessary conversation templates. Among the most important components of this knowledge database is a corpus of qualified and trusted answers. The challenge is to generate medically accurate information that is easy to understand and sufficiently colloquial while maintaining professionalism. These data will be collected from different sources and validated by the co-construction design committee for adoption to ensure that the chatbot can respond with appropriate output [30,31]. A total of three 2-hour co-construction workshops will be conducted.

4.3.7.2.4 Step 3: Testing, validation, and continuous improvement

The team of engineers will work closely with the co-construction design committee through development workshops to test the prototype's performance, especially its quality and safety. Participants will be invited to converse with the test prototype and complete an assessment. Such an approach will facilitate the continuous improvement of the prototype and the addition of new features (eg, user interface, questioning methods, and confidentiality measures) as necessary. Thus, each chatbot will evolve in real time during this step. A total of three 2-hour development workshops will be conducted.

It is important to note that, given the nature of software development engineering, steps 2 and 3 will be a continuous cyclic process. The research team will need to continually update the MARVIN chatbots based on feedback from the co-construction design committee on new requirements, improvement ideas, and technology updates.

4.3.7.2.5 Quantitative data collection and analysis

During each development workshop, co-construction design committee members will perform a quick descriptive assessment of the tested prototype using the adapted Mobile App Rating Scale for health-related apps [47] (Multimedia Appendix 8).

The scale has 28 items (range 1-5) in 6 sections covering subjects including engagement, functionality, esthetics, information, subjective quality, and health-related quality. Item scores will be averaged to obtain a score for each section and an overall score. It has shown an excellent internal consistency (Cronbach α=.938) and interrater reliability (2-way mixed intraclass correlation coefficient=0.920, 95% CI 0.797-0.987) for the independent overall score ratings of 37 different digital health tools [47]. On the basis of developer recommendations, we will consider a successful prototype for testing to have an average score of at least 4 for the sections on information and health-related quality, as well as an overall average score of 4, as they are identified as key indicators of prototype safety.

4.3.7.2.6 Qualitative data collection and analysis

Throughout the study, participants will have the option to take part in either English or French focus groups and workshops. All activities will be moderated by an experienced researcher with an assistant, digitally audio recorded, and manually transcribed verbatim for analysis while removing any nominal information provided to protect the participants’ identity. NVivo R1 (QSR International) will be used for qualitative data management.

Participants will also receive transcripts of the sessions in which they participated to ensure the trustworthiness of these data. Therefore, they will have the opportunity to challenge information that is perceived as incorrect. This will also allow them to verify that no information that could potentially identify them was inadvertently retained [48].

A focus group guide will be developed for each substudy considering its specific setting. Focus groups will be analyzed using an inductive thematic analysis approach to gather user recommendations (ie, topics to be addressed, expected conversational style, and desired features). Qualitative workshop data will contribute to thematic analysis in pursuit of objective 4.

4.3.7.3 Objective 2: Usability assessment

4.3.7.3.1 Overview

Although a usability study among people with HIV has already been conducted [21], usability will be evaluated for other individual chatbot substudies following the same methodology. During this stage, no updates will be made to the chatbots unless (1) the chatbots are not available because of force majeure (eg, host server failure or algorithm dependency update), in which case the team will implement updates to ensure the proper conduct of the study; (2) the results indicate suboptimal usability, in which case we will update the version and repeat the usability study; or (3) new medical information emerges that could benefit participants or prevent harm. Any such event will be documented in detail.

The expected duration of participation for objective 2 is 1 month. At enrollment, 30 participants will be required to complete an initial sociodemographic questionnaire. They will be given a training session and a user guide with instructions on how to access MARVIN via Messenger and the topics of questions they can ask MARVIN. Participants will be enrolled for a 3-week period of interacting with MARVIN by having at least 20 conversations, the topics of which will be specified in each subprotocol. Quantitative data will be collected using a usability survey once their 20 conversations are recorded.

Participants will be free to complete the tasks at any time during this 3-week period. If the chatbot does not receive a message from them within a week of their most recent conversation, it will proactively send a reminder. There will be no active third-party human involvement in the entire testing process between the chatbot and participant user except in the case of user-initiated requests (eg, to solve unexpected bugs).

In week 4, a total of 3 focus groups with 5 randomly selected participants each will be conducted to explore MARVIN's usability in greater depth and in the participants' own words. In total, 3 focus groups can capture at least 80% of the themes, which is sufficient saturation for a usability study [49].

4.3.7.3.2 Quantitative data collection and analysis

The initial questionnaire for objective 2 will gather fundamental sociodemographic information (eg, year of birth, preferred language, gender, and ethnic group identity) and digital technology use (Multimedia Appendix 8).

Descriptive statistics will be used to depict the sociodemographic characteristics and digital technology use of the participants. Continuous variables will be reported using measures such as minimum, maximum, mean, and SD. In the case of ordinal and nominal qualitative variables, we will report both counts and proportions.

Global usability will be collected using 2 validated scales: the *shorter version of the Usability Metric for User Experience (UMUX-Lite)* [50] and the *Acceptability E-scale (AES)* [51] (Multimedia Appendix 8).

Usability is defined in part as "the extent to which a product can be used to be effective, efficient, and provide users' satisfaction within its defined goal" [52]. The UMUX-Lite is a 2-item questionnaire answered on a 7-point Likert scale that is deemed appropriate for use in the evaluation of health technology [53]. The items ask whether the chatbot meets user needs and about perceived ease of use.

Acceptability is related to how agreeable, palatable, or satisfactory an intervention is perceived to be by stakeholders and is also considered part of global usability [14]. The AES contains 6 items rated on different 5-point Likert scales. It is a validated measure of the acceptability and usability of computer-based interventions for health care populations. Items evaluate, for example, how easy and enjoyable the innovation is to use, how helpful it is, and whether the amount of time to engage with it is acceptable.

As continuous outcomes, both usability and acceptability will be summarized using the minimum, maximum, mean, and SD. The sample mean of each global usability outcome will be compared with its recommended usability thresholds—68/100 for the UMUX-Lite score and 24/30 for the

AES score—using a Student *t* test. It will test the null hypothesis that the average UMUX-Lite score is ≤68 and that the average AES score is ≤24. A significance level of 5% will be adopted.

Four central subconstructs of the technology acceptance model (TAM) framework will also be assessed as secondary end points via validated instruments to complement the data: (1) perceived ease of use, (2) perceived usefulness, (3) attitude toward use, and (4) behavioral intention to use the chatbots.

Perceived ease of use will be measured using the Single Ease Question [54] responding on a 7-point Likert scale.

Drawing on instruments by Chau and Hu [55] and Davis [56], perceived usefulness will be measured on a 7-point Likert scale with 4 items slightly adapted for relevance to chatbots. The final score will be the average of these items.

Attitude toward using chatbots will be measured using the net promoter score (NPS) [57], which is used as a measure of user satisfaction. A single question will be asked on an 11-point Likert scale. To calculate the NPS, 3 groups are created: promoters (score of 9-10), passives (score of 7-8), and detractors (score of 0-6). Subtracting the percentage of detractors from that of promoters provides the NPS (range −100 to 100). Positive scores, and especially those of >50%, are judged positively.

Finally, behavioral intention to use the chatbots will be assessed using a validated 2-item questionnaire [58] rated on a 7-point Likert scale averaged to produce a final score.

Predicted positive associations between subconstructs within the TAM framework will be evaluated using simple linear regressions considering the slope coefficient. Their significance will be tested using a Student *t* test.

All statistical analyses will be conducted using the R software (R Foundation for Statistical Computing) [59].

4.3.7.3.3 Qualitative data collection and analysis

Participants’ experiences with MARVIN will be explored through a semistructured focus group on the main constructs of the TAM framework (ie, usefulness, ease of use, attitude, and behavioral intention to use), the analysis of which will help further explain the quantitative analysis. A focus group interview guide can be found in Multimedia Appendix 8.

A composite coding matrix based on the TAM and the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework will be favored for deductive content analysis [60]. The NASSS framework [61] was developed to support the implementation and scale-up of technological innovations in health care. It includes seven relevant domains: (1) the illness or condition, (2) the technology, (3) the value proposition, (4) the adopter system, (5) the organization, (6) the wider context, and (7) embedding and adaptation over time. Given that 5 to 7 are more focused on scaling up implementation, objective 2 will likely focus on the first 4 subdomains for analysis to illustrate associated barriers and facilitators of the early phases of implementation. All facilitators and barriers identified will be then subsequently matched to the subconstructs of the TAM when possible to understand their impact on global usability.

The content analysis involves 3 phases. In phase 1, preparation, the analyst will attempt to understand the entire data set through immersion in the data. In phase 2, organization, a composite coding matrix will be devised using the NASSS domains and the TAM subconstructs. The data will be coded and categorized using NVivo R1. Finally, in phase 3, reporting, the descriptive content of the categorization will be presented, addressing trustworthiness.

The saved transcripts of users' conversations with the chatbots may also be submitted for content analysis to describe and better understand the nature of chatbot-participant interactions, such as conversation topic trends.

4.3.7.4 Objective 3: Implementation assessment

4.3.7.4.1 Overview

For this objective, we will further assess the implementation outcomes of the MARVIN chatbots after they are deployed to the general population via Messenger. During this stage, the chatbots will be regularly updated to introduce new features or content, and their impact on implementation will be assessed. If the outcomes are negative, the corresponding version of the chatbot will be submitted for continuous improvement and evaluation.

We anticipate that the participation period will be 12 months and could be adjusted according to the health care context. Participants will receive a link to the same sociodemographic questionnaire as before (ie, via REDCap [Research Electronic Data Capture; Vanderbilt University] or Google Forms) directly through the chatbot at entry into the study (Multimedia Appendix 8). They will then be able to send messages to MARVIN whenever they wish. If the chatbot does not receive a message from the participant within a month of its most recent conversation with them, it will proactively ask the participant the following: “It’s been a month since our last conversation. How have you been?” The participant will be able to turn off this inquiry if they so wish.

Every 2 months, participants will receive a link to a questionnaire on implementation outcomes (Appendix 7). The chatbot will also ask 3 open-ended questions to collect information on the participant’s overall experience and suggestions for continuous improvement of the chatbot.

4.3.7.4.2 Quantitative data collection and analysis

The implementation outcome questionnaire will assess usability, acceptability, and appropriateness using validated instruments. Fidelity and adoption will be summarized using descriptive statistics on chatbot use metadata.

Usability and acceptability will be measured for objective 3 (implementation) as they will be for objective 2 (usability) using the UMUX-Lite and AES questionnaires as both measures are dynamic and vary with experience. Thus, usability and acceptability ratings may be different for each stage of implementation [14,62].

Appropriateness relates to the relevance or compatibility of the innovation to address a particular issue or problem [14]. The compatibility of an IT innovation is the extent to which it is considered consistent with users’ values, needs, and past experiences [63]. The *Compatibility Subscale* is a validated tool that contains 3 items rated on a 7-point Likert scale (range 1-7) to assess how an IT innovation “fits” with the user’s work style [64]. An adapted version will be administered to health care professional participants. A minimum average score of 5.5 is set as the threshold for adequate compatibility. For patient participants, the *Intervention Appropriateness Measure* will be used [65]. It contains 4 items scored on a 5-point scale to assess an innovation’s suitability for a user. A mean score of at least 4 will indicate the appropriateness of the chatbot intervention for the patient population.

Fidelity is the degree to which the intervention is implemented as intended [14]. On the basis of the fidelity measures of digital health intervention implementation identified by Coorey et al [66], we will analyze the following metadata to comprehensively assess the chatbot fidelity: (1) intervention fidelity (the proportion of participants who continue to use the chatbot after a 1-month period), (2) frequency and duration (the monthly frequency with which participants use the chatbot and the average total duration of using the chatbot), (3) messages delivered (the total number of messages that participants interacted with in the chatbot over the course of the study period as well as the average number per participant), and (4) range of messages received (the frequency distribution of different conversational topics triggered by all participants).

Adoption, or uptake, is “the intention, initial decision, or action to try or employ an innovation or evidence-based practice” [14]. It will be measured using the proportion of monthly new users enrolled to all users in the study. The target will be 5% per month given that the median user growth rate for small-scale software services is 4.4% [67,68].

Statistically, a strategy similar to that of objective 2 will be adopted to describe the data. For usability, acceptability, and appropriateness, a Student *t* test will be used to test the null hypothesis that the average score is inferior or equal to the predetermined thresholds.

4.3.7.4.3 Qualitative data collection and analysis

To better understand the implementation outcomes, participants will be invited to answer three open-ended questions: (1) What did you like most about using MARVIN? (2) What did you dislike about using MARVIN? (3) How would you improve MARVIN?

The content analysis of this material using the same coding matrix as in objective 2 will likely include the last 3 subdomains of the NASSS framework given the implementation stage at objective 3. This will help document and detail barriers to and facilitators of using the chatbots and their associated implementation outcomes as well as identify targets for continuous improvement.

As with objective 2 (usability), the saved transcripts of chatbot conversations may also be submitted for content analysis to characterize and better understand the nature of chatbot-participant interactions.

4.3.7.5 Objective 4: Evaluate the impact of different stakeholder partnerships

4.3.7.5.1 Overview

The reporting of patient and stakeholder engagement roles will follow the revised Guidance for Reporting Involvement of Patients and the Public [69].

4.3.7.5.2 Quantitative data collection and analysis

Quantitative data on patient and stakeholder engagement activities, such as the number and length of actual workshops or focus groups conducted; the number of attendees; and the user requirements, design parameters, and improvement recommendations made by stakeholders during the workshops or focus groups, will be reported to illustrate the impact of patient and stakeholder engagement on the adaptation or development of the MARVIN chatbots.

4.3.7.5.3 Qualitative data collection and analysis

Upon completion of each objective, focus groups will be conducted with stakeholders involved in the co-construction of the chatbots to identify the participants' perspectives on target users' involvement in research. A focus group guide will be developed for each substudy considering its specific setting. These data will be analyzed thematically along with workshop data from objective 1 to determine how potential end users (patient partners and health care professionals) were integrated into the research team, participated in the work, influenced decision-making, and contributed to chatbot adaptation or development.

4.3.8 Data management, confidentiality, and security

Only data relevant to this study outlined in this protocol will be collected by the research team. A comprehensive overview of the study's data management strategy is presented in Table 4.1, including the collection of recruitment and study data. For objectives 1 (development), 2 (usability), and 4 (partnership evaluation), participants' basic sociodemographic data and contact information will be recorded. For objective 3 (implementation), only participants' Facebook account names will be gathered alongside their eligibility responses. Participants will be identified using alphanumeric codes. The link between these codes and the participants' identities will be kept by the research team within a password-protected digital file safeguarded by the MUHC firewall. Access to these records will be exclusive to the research team.

Table 4.1 Study data management strategy.

	Objective 1	Objective 2	Objective 3	Objective 4
Recruitment data				
Data type	• Basic sociodemographic data and contact information	• Basic sociodemographic data and contact information	• Facebook account name, answers to eligibility questions, and web-based consent records	• Basic sociodemographic data and contact information
Protection method	• Password-protected digital files	• Password-protected digital files	• Encrypted database on Amazon Web Services cloud server and synchronized to password-protected digital files	• Password-protected digital files
Storage location	• MUHC[a] internal storage	• MUHC internal storage	• Amazon Web Services and	• MUHC internal storage

				MUHC internal storage	
Study data					
	Data type	• Conversation histories	• Conversation histories • Study questionnaire data • Qualitative data	• Conversation histories • Study questionnaire data	• Qualitative data
	Protection method	• Encrypted database on Amazon Web Services cloud server and 2-factor–authenticated Facebook MARVIN account	• Encrypted database on Amazon Web Services cloud server and 2-factor–authenticated Facebook MARVIN account • 2-factor–authenticated Google Forms or REDCap[b] (Vanderbilt University) accounts and password-protected digital files	• Encrypted database on Amazon Web Services cloud server and 2-factor–authenticated Facebook MARVIN account • 2-factor–authenticated Google Forms or REDCap accounts and	• Password-protected digital files

			- Password-protected digital files	password-protected digital files	
	Storage location	- Amazon Web Services and Facebook	- Amazon Web Services and Facebook - Google Forms or REDCap and MUHC internal storage - MUHC internal storage	- Amazon Web Services and Facebook - Google Forms or REDCap and MUHC internal storage	- MUHC internal storage

[a]MUHC: McGill University Health Centre.

[b]REDCap: Research Electronic Data Capture.

The scope of the study data will include information from the web-based research questionnaire, qualitative data, transcripts of participant conversations with MARVIN, and MARVIN-related metadata. Data required for distinct study objectives through questionnaires will be collected using an appropriate collection tool (eg, Google Forms or REDCap) and subsequently extracted into a password-protected Microsoft Excel (Microsoft Corp) spreadsheet for analysis. Qualitative data, once collected and transcribed, will be password protected and deidentified during analysis. All study data will only be accessible to the research team.

Regarding the technical cybersecurity aspects of the MARVIN chatbots, Figure 4.7 offers a visual representation of the data flow as participants engage with the chatbots. Throughout the study, participants will send messages to MARVIN chatbots via their personal devices. These messages will be relayed through Messenger’s application programming interface to the Amazon Web Services (AWS) cloud server, where the research team deployed the MARVIN service. Response messages from MARVIN chatbots will then be sent back to participants’ devices via the

Messenger application programming interface. The entire communication process will be encrypted. The research team will be responsible for all of Facebook's MARVIN chatbot accounts and MARVIN servers deployed on the AWS cloud server. Records of chatbot conversations will be stored on an encrypted AWS cloud server. These data will be anonymized by the research team and used exclusively for future model training to enhance chatbot performance for research and quality assurance purposes.

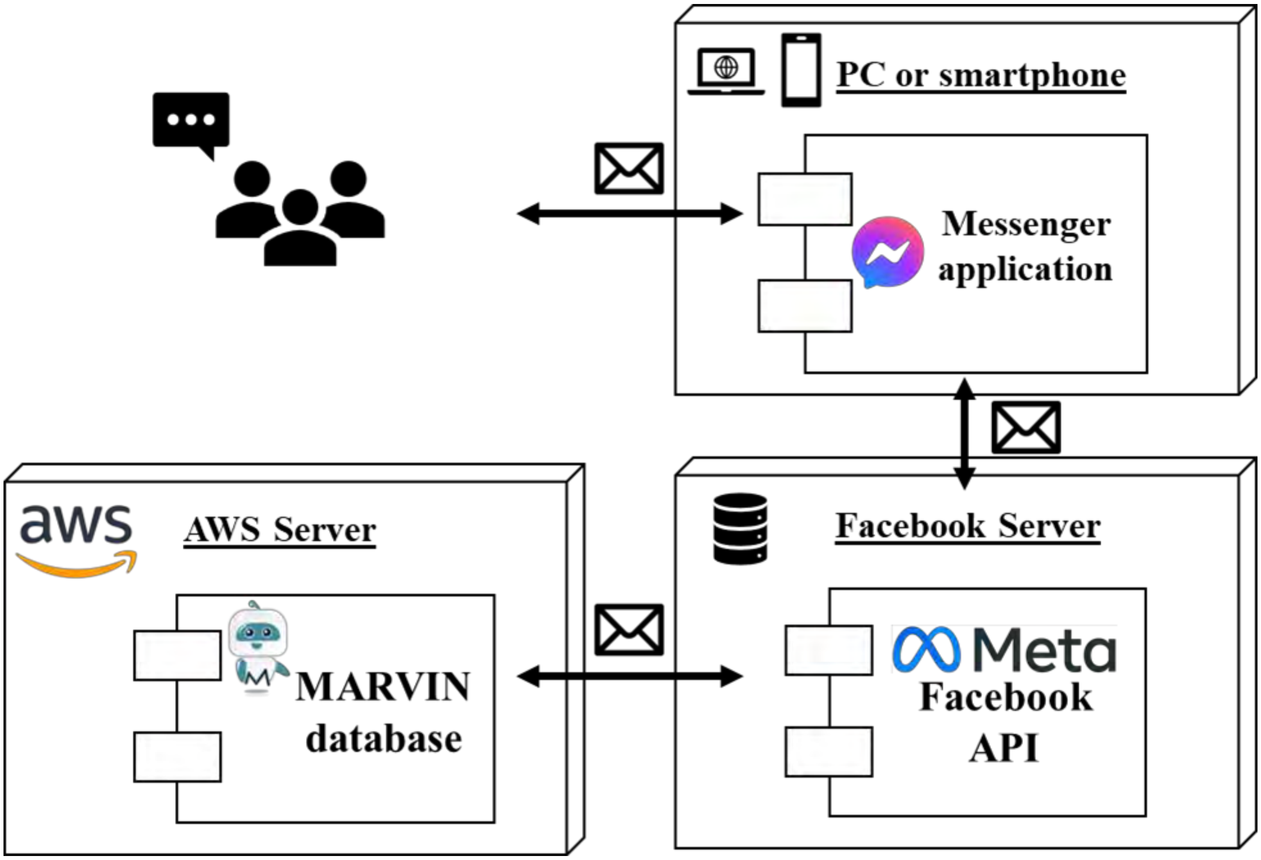


Figure 4.7 MARVIN chatbot—message data flow diagram. API: application programming interface; AWS: Amazon Web Services.

Conversations on Messenger will also be logged and stored on Messenger's server, which is necessary for display on both the participant and chatbot interfaces. The data handling in this context adheres to Facebook's policies [70-72]. If a participant withdraws from the study, the collected study data will be removed as well if the participant so wishes. In particular, in accordance with Messenger's privacy policy, participants will be asked to delete the conversations with MARVIN from their personal accounts, and the research team will delete the conversations from MARVIN’s account. Thus, Facebook will stop storing these data as they are no longer required to provide their services and Meta Platforms products.

Note that all platforms involved, including Messenger, Google Forms, and REDCap, comply with the General Data Protection Regulation implemented by the European Union. It is recognized as the highest standard available, especially in terms of AI applications, equivalent to or surpassing the Personal Information Protection and Electronic Documents Act in Canada [73], where MARVIN is deployed. Access to each platform’s accounts will be exclusive to the research team and secured through 2-factor authentication.

4.3.9 Governance board

In view of the current ever shifting regulatory landscape surrounding health care AI applications, a governance board will be formed to help the research team obtain external perspectives; assess the ethical and technological issues that may arise from the MARVIN chatbots; and ensure prudent development, testing, and mitigation of associated risks. The board will also summarize best practices from the substudies to enhance future iterations.

Candidates for the governance board include expert patients as well as experts in the fields of AI, clinical science, ethics, legal affairs, and communications. Invitations will be sent by the research team via email. Annual assemblies will be held for reviewing study progress and exchanging insights. For specific inquiries, members of the governance board will remain reachable by the MARVIN research team via email summons.

4.3.10 Anticipated risks and benefits

Participants in this study will not be exposed to direct physical risk while partaking in focus groups or interviews, conversing with chatbots via messaging, or completing web-based surveys as they will not receive any pharmaceutical or invasive medical interventions. Furthermore, the preceding pilot study did not reveal any known risks of participation.

However, potential indirect risks are worth considering. In the case of web-based recruitment and verbal consent acquisition, there is a risk of confidentiality breach. This risk can be exacerbated if participants use a personal email address to communicate with the research team. In response, researchers will only use institutional email addresses for correspondence purposes. Participants will also be advised to protect their pertinent personal electronic data.

When using the web-based chatbot, participants will use their personal Facebook account and may share details of their participation in the study. Vulnerability to security breaches (eg, device loss, inadvertent device exposure, phishing, and malware) may arise concerning participants' Facebook accounts. To mitigate this risk, participants will be explicitly reminded during recruitment to secure relevant personal information. The MARVIN chatbots will similarly provide appropriate reminders in the electronic information and consent forms, such as "I recommend that you do not share study-related information with others unnecessarily."

The time required to complete the questionnaires and participate in interviews or focus groups may be inconvenient and distressing for certain individuals. Others may also be uncomfortable answering specific questions or interacting with the chatbot. In situations in which questions are deemed sensitive, private, or distressing, participants are not obliged to respond. The research team will always be available to discuss participant concerns and refer them to appropriate resources, including teleconsultation with a mental health professional or other support service.

It is possible that the MARVIN chatbots will have difficulty understanding messages from participants during the study. In such instances, the chatbot will indicate that it cannot understand, as described previously in Figure 4.5, and suggest seeking help from a health care professional. There is also a small chance that the chatbot will provide erroneous advice. As an example, the chatbot might inform a patient who missed a 2-hour dose to stop taking the medication completely and consult a health care professional immediately. To minimize this potential risk, participants will be informed of the limitations of the chatbots during the consent process. In addition, participants will be prompted to report any perceived inaccuracies and their consequences to the research team in a timely manner. Weekly revisions of user chat logs will be conducted by the research team to ensure timely human intervention in the event of errors. The governance board for this study will consistently monitor and discuss these reports.

Finally, any other service-related risks (eg, chatbot or Facebook network service disruptions) will be communicated to each participant in a timely manner and properly documented by the study coordinator.

Participation in the study presents benefits, including early access to chatbot interventions with validated health care information. Participants will also be compensated appropriately upon completion of each study objective [74] in the form of a gift card or, exceptionally, a money transfer.

4.4 Results

From July 2022 to October 2023, four substudies were established in conjunction with the completion of this master protocol:

1. The first study is a continuation of the original MARVIN for HIV self-management. This project has secured funding from the Fonds de recherche du Québec – Santé Réseau SIDA/Maladies Infectieuses (AIDS and Infectious Disease Network). A subprotocol targeting objective 3 (implementation) is currently being prepared and scheduled for REB submission in early 2024. Recruitment is scheduled to begin in mid 2024, and the related data analysis will begin in early 2025.

2. The second study is MARVIN-Pharma, a project to promote community pharmacists' knowledge of HIV treatment, with its prototype to be completed by the end of 2023. A related manuscript based on a pharmacist needs assessment is in preparation. A subprotocol for objective 2 (usability) is being developed and is scheduled to be submitted for REB approval in early 2024, and recruitment is scheduled for mid 2024.

3. The third project is to develop the MARVINA chatbot for self-management of patients with breast cancer. This study is supported by funding from the MUHC Cedars Cancer Foundation. The subprotocol was approved by the MUHC REB on September 26, 2023 (approval MP-37-2024-9633). Recruitment for objective 1 (development) is expected to begin in early 2024.

4. The fourth study is to develop MARVIN-CHAMP, an accessible chatbot to assist in the management of pediatric patients with infectious conditions. A funding proposal for this project will be submitted to the Canadian Institutes of Health Research Spring 2024 Project Grant program. A subprotocol for objective 1 (development) is under development and scheduled to be submitted for REB approval in mid October 2023. Recruitment for objective 1 (development) is expected to start in early 2024.

None of the funding sources had any role in the design of this study and will not be involved in the interpretation of the results or the decision to submit them for publication.

4.5 Discussion

4.5.1 Expected findings

AI technologies have made phenomenal advances, but relevant clinical translation in key areas remains slow. To the best of our knowledge, this is the first known master protocol in digital health dedicated to implementing chatbot interventions across diverse health conditions and clinical settings. Before this, master protocols had been structured primarily for pharmaceutical intervention studies [75]. This master protocol shares key experimental components and operational processes while capitalizing on the similarities of the underlying IT infrastructures already in place. Coupled subsequently with thorough discussion and deliberation among intended users, developers, administrators, and regulators, the efficiency of creating and coordinating multiple studies of the same type has greatly improved. Although the upfront costs and planning time were significant, with this master protocol taking a year to complete from inception to REB approval, it will facilitate the generation of high-quality evidence essential for guiding medical practice. Through centralized management and shared governance, it also reduces development costs, enables broad decision-making, and allows patients to benefit earlier from advanced interventions.

The selected adaptive platform trial format, although designed initially for oncology and infectious disease drug development, has also been identified as being applicable to digital health interventions [76]. Its flexibility is a noteworthy advantage. Digital health interventions are now increasingly being applied to a wide range of conditions. The flexibility of the infrastructure facilitates the addition of substudies or necessary adjustments to each substudy, and health authorities, institutional review boards, and ethics committees will have a clear understanding of what changes are occurring across substudies [77]. Second, design features such as early termination of the trial or re-estimation of the sample size can avoid wasting resources, thus allowing for faster dissemination of research results to the communities that will benefit the most [38].

Patient and stakeholder contributions are integral to shaping this master protocol and associated materials. Their co-constructive engagement allows them to take a leading role in the ongoing digitization of health care and can help mitigate or even address the risks that chatbots face during implementation, such as those tied to trustworthiness, data privacy, and exacerbating inequalities in access to health care. Incorporating patient and stakeholder involvement strategies in developing master protocols, as suggested by Huml et al [78], can make them more successful for patients, providers, and sponsors. Patients and stakeholders will be invited to contribute on an ongoing basis to subsequent chatbot development, reporting of trial results, product marketing, and knowledge translation. Systematically documenting, investigating, and reporting this entire process will be beneficial in providing the scientific community with a clear and replicable model for responsible AI medical research.

Certain limitations need to be recognized. This master protocol focuses solely on the assessment of chatbot implementation outcomes, similar to master protocols typically prepared for pharmaceutical investigations that focus on “early exploratory development phases” [78]. Clinical and service-related outcomes are not included in this master protocol as there is no consensus on their assessment methods. Notably, <25% of AI-based digital health trials include patient-reported outcome measures as end points despite their widespread use in other health care trials [79]. Therefore, corresponding research efforts are necessary to develop relevant high-quality assessment metrics to foster the development and validation of user-centered chatbots. If the substudy decides to assess their clinical effectiveness, appropriate amendments will be made.

Large language model (LLM) technology is also not considered in this master protocol for the time being. The release of LLMs, represented by GPT-4, has been indeed impressive. Unfortunately, LLM-based chatbots exhibit limitations such as a lack of transparency, sharing of unverified health information, and poor interpretability [80]. These factors threaten the trustworthiness and security of chatbots, which are key to their successful implementation in health care. Although current state-of-the-art LLMs are, of course, embraced to help generate more diverse and personalized responses, trialing these models in clinical settings also remains a challenge owing to the lack of relevant regulatory compliance and the fact that the existing software-as-a-medical-device framework is not suited to such models [81]. Chatbots can be a safe tool to seek information if they are validated and approved by health care professionals and all personal data are properly secured through robust up-to-date privacy and security safeguards [82,83]. In line with this protocol, the team strongly believes in and adheres to a careful content validation and model fine-tuning strategy so as to maximize the accuracy, trustworthiness, and safety of the solutions being delivered for responsible AI innovation.

Finally, another limitation of the protocol is the use of convenience sampling, potentially introducing bias toward individuals more inclined to use digital health technologies. Furthermore, in the web-based direct recruitment process for objective 3 (implementation), all participants will be self-referred, which will make it challenging to assess whether the participants genuinely meet the inclusion criteria. In addition, the self-reported quantitative questionnaire may be susceptible to random participant responses. Outliers in the data will be checked, and appropriate statistical methods will be used to improve the quality of the final data.

4.6 Conclusions

Overall, the development and adaptation of the MARVIN chatbots, co-constructed with those directly involved, holds the promise of fostering patient self-management and enhancing health care efficiency. This study shall provide a comprehensive examination of the implementation outcomes of innovative chatbot interventions tailored for patients and health care professionals. Moreover, it will contribute to the formulation of best practice recommendations for the co-constructive engagement of patients and stakeholders in digital health research. If properly applied, this master protocol has the potential to be sustained for years or even decades and allow innovations to be rapidly translated to clinical practice. Advances in methodology combined with the surge in AI will provide deeper evidence to achieve the goal of patient-partnered personalized medicine and, ultimately, help deliver the right interventions for the right patient at the right time.

4.7 Acknowledgments

This study secured funding from the Fonds de recherche du Québec – Santé (FRQS) Réseau SIDA-Maladies Infectieuses (AIDS and Infectious Disease Network; project title: *Developing an AI chatbot-based triage solution to support people with HIV*; principal investigators [PIs]: YM, SA, and B Lebouché). This study is also supported by the Canadian Institutes of Health Research Strategy for Patient-Oriented Research Québec Support Unit–Methodological Developments (grant M006; PI: B Lebouché), the Partnerships to Improve HIV Outcomes and Treatments program from ViiV Healthcare (grant 2021-2186-PIHVOT; PI: RT), and the Cedars Cancer Foundation (project title: *Developing MARVINA, an Intelligent Conversational Agent for Breast Cancer Self-Management through Patient Stakeholder Engagement*; PIs: JA and B Lebouché). YM is supported by the Postgraduate Scholarships – Doctoral program; the Optimizing Power Skills in Interdisciplinary, Diverse, and Innovative Academic Networks (OPSIDIAN) program from the Natural Sciences and Engineering Research Council; and a doctoral research award from the Fonds de recherche du Québec – Nature et technologies in partnership with the Unité de soutien au système de santé apprenant Québec. MPP is supported by a Senior Career Award financed by the FRQS (322746 and 322747), the Centre de Recherche du Centre hospitalier de l'Université de Montréal, and the Québec Ministry of Health and Social Services. B Lebouché is supported by 2 career awards—a Senior Salary Award from the FRQS (311200) and the Lettre d'entente 250 from the Quebec Ministry of Health and Social Services for researchers in family medicine—and holds a Canadian Institutes for Health Research Strategy for Patient-Oriented Research Mentorship Chair in Innovative Clinical Trials in HIV. EO is funded as a Clinical Research Scholar–Junior 1 (2023-2027) by the FRQS (331771) awarded in partnership with the Unité de soutien au système de santé apprenant Québec. MHZ is supported by a Junior 1 Career Award from the FRQS (309777). The funding sources had no role in the design of this study and were not involved in the interpretation of the results or the decision to submit them for publication. The authors acknowledge the technical support provided by students and interns from Polytechnique Montreal for the development of the MARVIN chatbots and Lévis Thériault for referring the students. The authors also acknowledge Imane Frih, Maria Nait El Haj, and Gavin Tu for the creation of the clinical content. The names of the members of the MARVIN chatbot patient expert committee cannot be provided as they are patients and their names must remain confidential. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

4.8 Data availability

The data sets generated and analyzed for this study are available from the corresponding author (B Lebouché) upon reasonable request.

4.9 Authors' contributions

In no order of contribution, YM, MPP, JP, NA, SV, KE, and B Lebouché helped conceptualize the study and data collection tools. YM and SA on the software side and ML, B Lemire, RT, B Lebouché, and the MARVIN chatbot patient expert committee on the clinical side collaborated to develop the MARVIN chatbots. YM, JP, and B Lebouché wrote the manuscript. All authors critically reviewed the manuscript and approved the final version.

4.10 Conflicts of interest

B Lebouché has received research support, consulting fees, and speaker fees from ViiV Healthcare, Merck, and Gilead. The authors are developers of the MARVIN chatbot intervention.

4.11 References

[1] Lorig KR, Holman HR. Self-management education: history, definition, outcomes, and mechanisms. Ann Behav Med. Aug 2003;26(1):1-7.

[2] Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. J Med Internet Res. May 16, 2016;18(5):e97.

[3] Wong DK, Cheung MK. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res. Mar 27, 2019;21(3):e10831.

[4] Lee HY, Jin SW, Henning-Smith C, Lee J, Lee J. Role of health literacy in health-related information-seeking behavior online: cross-sectional study. J Med Internet Res. Jan 27, 2021;23(1):e14088.

[5] Daraz L, Morrow AS, Ponce OJ, Beuschel B, Farah MH, Katabi A, et al. Can Patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. J Gen Intern Med. Sep 2019;34(9):1884-1891.

[6] Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res. May 02, 2019;21(5):e12522.

[7] Thoër C, Millerand F, Myles D, Orange V, Gignac O. Enjeux éthiques de la recherche sur les forums Internet portant sur l’utilisation des médicaments à des fins non médicales. Communiquer. 2012(7):1-22.

[8] Miner H, Fatehi A, Ring D, Reichenberg JS. Clinician telemedicine perceptions during the COVID-19 pandemic. Telemed J E Health. May 01, 2021;27(5):508-512.

[9] Gentili S, Huang V, Mamo J, Cuschieri S. Chronic diseases in 2021. In: Proceedings of the 14th European Public Health Conference. Presented at: EPH '21; November 10-12, 2021, 2022; Virtual Event. URL: https://eupha.org/repository/Chronic %20Diseases%20in%202021%20-%20EUPHA%20track%20report.pdf/>

[10] Gisondi MA, Barber R, Faust JS, Raja A, Strehlow MC, Westafer LM, et al. A deadly infodemic: social media and the power of COVID-19 misinformation. J Med Internet Res. Feb 01, 2022;24(2):e35552.

[11] Dubin RE, Flannery J, Taenzer P, Smith A, Smith K, Fabico R, et al. ECHO Ontario chronic pain and opioid stewardship: providing access and building capacity for primary care providers in underserviced, rural, and remote communities. Stud Health Technol Inform. 2015;209:15-22.

[12] Henny KD, Duke CC, Geter A, Gaul Z, Frazier C, Peterson J, et al. HIV-related training and correlates of knowledge, HIV screening and prescribing of nPEP and PrEP among primary care providers in southeast United States, 2017. AIDS Behav. Nov 2019;23(11):2926-2935.

[13] Xing Z, Yu F, Qanir YA, Guan T, Walker J, Song L. Intelligent conversational agents in patient self-management: a systematic survey using multi data sources. Stud Health Technol Inform. Aug 21, 2019;264:1813-1814.

[14] Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health. Mar 2011;38(2):65-76.

[15] Boggiss A, Consedine N, Hopkins S, Silvester C, Jefferies C, Hofman P, et al. Improving the well-being of adolescents with type 1 diabetes during the COVID-19 pandemic: qualitative study exploring acceptability and clinical usability of a self-compassion chatbot. JMIR Diabetes. May 05, 2023;8:e40641.

[16] Anmella G, Sanabra M, Primé-Tous M, Segú X, Cavero M, Morilla I, et al. Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and health care professionals: development, feasibility, and potential effectiveness studies. J Med Internet Res. Apr 03, 2023;25:e43293.

[17] Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res. Aug 12, 2021;5(8):e20678.

[18] Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res. Mar 23, 2021;23(3):e24850.

[19] Au J, Falloon C, Ravi A, Ha P, Le S. A beta-prototype chatbot for increasing health literacy of patients with decompensated cirrhosis: usability study. JMIR Hum Factors. Aug 15, 2023;10:e42506.

[20] Kadariya D, Venkataramanan R, Yip HY, Kalra M, Thirunarayanan K, Sheth A. kBot: knowledge-enabled personalized chatbot for asthma self-managemen. In: Proceedings of the 2019 IEEE International Conference on Smart Computing. Presented at: SMARTCOMP '19; June 12-15, 2019, 2019;138-143; Washington, DC. URL: https://ieeexplore.ieee.org/document/8784055

[21] Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res. Oct 22, 2020;22(10):e20346.

[22] Griffin AC, Xing Z, Khairat S, Wang Y, Bailey S, Arguello J, et al. Conversational agents for chronic disease self-management: a systematic review. AMIA Annu Symp Proc. 2020;2020:504-513.

[23] Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer. Nov 29, 2021;7(4):e27850.

[24] Shah SS, Gvozdanovic A. Digital health; what do we mean by clinical validation? Expert Rev Med Devices. Dec 12, 2021;18(sup1):5-8.

[25] Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res. Sep 14, 2020;22(9):e20701.

[26] Solomon DH, Rudin RS. Digital health technologies: opportunities and challenges in rheumatology. Nat Rev Rheumatol. Sep 24, 2020;16(9):525-535.

[27] Sadasivan C, Cruz C, Dolgoy N, Hyde A, Campbell S, McNeely M, et al. Examining patient engagement in chatbot development approaches for healthy lifestyle and mental wellness interventions: scoping review. J Particip Med. May 22, 2023;15:e45772.

[28] Kildea J, Battista J, Cabral B, Hendren L, Herrera D, Hijal T, et al. Design and development of a person-centered patient portal using participatory stakeholder co-design. J Med Internet Res. Feb 11, 2019;21(2):e11371.

[29] Pomey MP, Flora L, Karazivan P, Dumez V, Lebel P, Vanier MC, et al. Le « Montreal model » : enjeux du partenariat relationnel entre patients et professionnels de la santé. Santé Publique. 2015;1:41-50.

[30] Dilhac MA, Abrassart C, Voarino N. Rapport de la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle. Raisons Politiques. 2020;77(1):67-81.

[31] Gagnon MP, Desmartis M, Lepage-Savary D, Gagnon J, St-Pierre M, Rhainds M, et al. Introducing patients' and the public's perspectives to health technology assessment: a systematic review of international experiences. Int J Technol Assess Health Care. Jan 2011;27(1):31-42.

[32] Ma Y, Tu G, Lessard D, Vicente S, Engler K, Achiche S, et al. An artificial intelligence-based chatbot to promote HIV primary care self-management: a mixed method usability study. Ann Fam Med. Nov 2023;21(Supplement 3):5267.

[33] Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot Feasibility Stud. 2016;2:64.

[34] Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res. Dec 31, 2011;13(4):e126.

[35] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AICONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. Sep 2020;26(9):1364-1374.

[36] Park JJ, Siden E, Zoratti MJ, Dron L, Harari O, Singer J, et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. Trials. Sep 18, 2019;20(1):572.

[37] Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. N Engl J Med. Jul 06, 2017;377(1):62-70.

[38] Kaizer AM, Belli HM, Ma Z, Nicklawsky AG, Roberts SC, Wild J, et al. Recent innovations in adaptive trial designs: a review of design opportunities in translational research. J Clin Transl Sci. 2023;7(1):e125.

[39] Etikan I, Musa SA, Alkassim RS. Comparison of convenience sampling and purposive sampling. Am J Theor Appl Stat. 2016;5(1):1.

[40] Lewis JR. Sample sizes for usability studies: additional considerations. Hum Factors. Jun 23, 1994;36(2):368-378.

[41] Lewis M, Bromley K, Sutton CJ, McCray G, Myers HL, Lancaster GA. Determining sample size for progression criteria for pragmatic pilot RCTs: the hypothesis test strikes back!. Pilot Feasibility Stud. Feb 03, 2021;7(1):40.

[42] Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract. May 2004;10(2):307-312.

[43] Ma Y, Achiche S, Lebouché B. Chatbot Marvin Meta page. Meta. URL: https://www.facebook.com/ChatbotMarvin [accessed 2023-10-23]

[44] Bocklisch T, Faulkner J, Pawlowski N, Nichol A. Rasa: Open source language understanding and dialogue management. arXiv. Preprint posted online Dec 14, 2017

[45] Vlasov V, Drissner-Schmid A, Nichol A. Few-shot generalization across dialogue tasks. arXiv. Preprint posted online November 28, 2018

[46] Smutny P, Schreiberova P. Chatbots for learning: a review of educational chatbots for the Facebook Messenger. Comput Educ. Jul 2020;151:103862.

[47] Roberts AE, Davenport TA, Wong T, Moon HW, Hickie IB, LaMonica HM. Evaluating the quality and safety of health-related apps and e-tools: adapting the Mobile App Rating Scale and developing a quality assurance protocol. Internet Interv. Apr 2021;24:100379.

[48] Grové C. Co-developing a mental health and wellbeing chatbot with and for young people. Front Psychiatry. 2020;11:606041.

[49] Guest G, Namey E, McKenna K. How many focus groups are enough? Building an evidence base for nonprobability sample sizes. Field Methods. Jul 24, 2016;29(1):3-22.

[50] Lewis JR, Utesch BS, Maher DE. UMUX-LITE: when there's no time for the SUS. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Presented at: CHI '13; April 27-May 2, 2013, 2013;2099-2102; Paris, France. URL: https://dl.acm.org/doi/10.1145/2470654.2481287

[51] Tariman JD, Berry DL, Halpenny B, Wolpin S, Schepp K. Validation and testing of the acceptability e-scale for web-based patient-reported outcomes in cancer care. Appl Nurs Res. Feb 2011;24(1):53-58.

[52] ISO/TS 20282-2(en) usability of consumer products and products for public use- part 2: ummative test method. International Organization for Standardization. 2013. URL: https://www.iso.org/obp/ui/#iso:std:iso:ts:20282:-2:ed-2:v1:en [accessed 2024-01-18]

[53] Borsci S, Buckle P, Walne S. Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? Appl Ergon. Apr 2020;84:103007.

[54] Sauro J, Dumas JS. Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Presented at: CHI '09; April 4-9, 2009, 2009;1599-1608; Boston, MA. URL: https://dl.acm.org/doi/10.1145/1518701.1518946

[55] Chau PY, Hu PJ. Investigating healthcare professionals' decisions to accept telemedicine technology: an empirical test of competing theories. Inf Manag. Jan 2002;39(4):297-311.

[56] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. Sep 1989;13(3):319.

[57] Adams C, Walpola R, Schembri AM, Harrison R. The ultimate question? Evaluating the use of net promoter score in healthcare: a systematic review. Health Expect. Oct 19, 2022;25(5):2328-2339.

[58] Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Manage Sci. Feb 2000;46(2):186-204.

[59] R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2018. URL: https://www.R-project.org/ [accessed 2024-01-12]

[60] Braun V, Clarke V. Thematic analysis. In: Cooper H, Coutanche MN, McMullen LM, Panter AT, Rindskopf D, Sher KJ, editors. APA Handbook of Research Methods in Psychology. Washington, DC. American Psychological Association; 2012;57-71.

[61] Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform. Jul 30, 2019;263:193-204.

[62] Hornbæk K. Current practice in measuring usability: challenges to usability studies and research. Int J Hum Comput Stud. Feb 2006;64(2):79-102.

[63] Engler K, Vicente S, Ma Y, Hijal T, Cox J, Ahmed S, et al. Implementation of an electronic patient-reported measure of barriers to antiretroviral therapy adherence with the Opal patient portal: protocol for a mixed method type 3 hybrid pilot study at a large Montreal HIV clinic. PLoS One. 2021;16(12):e0261006.

[64] Moore GC, Benbasat I. Development of an instrument to measure the perceptions of adopting an information technology innovation. Inf Syst Res. Sep 1991;2(3):192-222.

[65] Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci. Aug 29, 2017;12(1):108.

[66] Coorey G, Peiris D, Scaria A, Mulley J, Neubeck L, Hafiz N, et al. An internet-based intervention for cardiovascular disease management integrated with primary care electronic health records: mixed methods evaluation of implementation fidelity and user engagement. J Med Internet Res. Apr 26, 2021;23(4):e25333.

[67] Anastaselos T. What is a good SaaS growth rate? ChartMogul. URL: https://chartmogul.com/blog/good-monthly-growth-rate/ [accessed 2023-09-07]

[68] Standing out in the crowd: 7 causes of slow or low user growth. Specno. Feb 2023. URL: https://www.specno.com/blog/low-user-growth#:~:text=Daily%20Active%20Users%20(DAU)%20%E2%80%93,over%2025%25%20is%20excellent!) [accessed 2023-09-07]

[69] Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ. Aug 02, 2017;358:j3453.

[70] Meta privacy policy. Meta. URL: https://www.facebook.com/privacy/policy/ [accessed 2024-01-12]

[71] Meta data security terms. Meta. URL: https://www.facebook.com/legal/terms/data_security_terms [accessed 2024-01-12]

[72] Privacy and safety on Messenger. Meta. URL: https://www.facebook.com/help/messenger-app/1064701417063145/?helpref =hc_fnav [accessed 2024-01-12]

[73] Merrick R, Ryan S. Data privacy governance in the age of GDPR. Risk Manag. 2019;66(3):38-43.

[74] Geneviève D, Alexandre G, Antoine B, Vincent D. Principes directeurs – Dédommagement financier pour la recherche en partenariat avec les patients et le public. Unité de soutien SRAP du Québec. 2022. URL: https://ceppp.ca/wp-content/uploads /2021/01/USSQ_Principes-directeurs_Dedommagement_vAou%CC%82t2018.pdf [accessed 2024-01-18]

[75] Siden EG, Park JJ, Zoratti MJ, Dron L, Harari O, Thorlund K, et al. Reporting of master protocols towards a standardized approach: a systematic review. Contemp Clin Trials Commun. Sep 2019;15:100406.

[76] Subbiah V. The next generation of evidence-based medicine. Nat Med. Jan 2023;29(1):49-58.

[77] Meyer EL, Mesenbrink P, Dunger-Baldauf C, Fülle HJ, Glimm E, Li Y, et al. The evolution of master protocol clinical trial designs: a systematic literature review. Clin Ther. Jul 2020;42(7):1330-1360.

[78] Huml RA, Collyar D, Antonijevic Z, Beckman RA, Quek RG, Ye J. Aiding the adoption of master protocols by optimizing patient engagement. Ther Innov Regul Sci. Nov 2023;57(6):1136-1147.

[79] Pearce FJ, Cruz Rivera S, Liu X, Manna E, Denniston AK, Calvert MJ. The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic

evaluation of ClinicalTrials.gov records (1997-2022). Lancet Digit Health. Mar 2023;5(3):e160-e167.

[80] Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940.

[81] Lee P, Goldberg C, Kohane I. The AI Revolution in Medicine: GPT-4 and Beyond. New York, NY. Pearson; 2023.

[82] Greene A, Greene CC, Greene C. Artificial intelligence, chatbots, and the future of medicine. Lancet Oncol. Apr 2019;20(4):481-482.

[83] Surani A, Das S. Understanding privacy and security postures of healthcare chatbots. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Presented at: CHI '22; April 29-May 5, 2022, 2022;1-7; New Orleans, LA. URL: https://cui.acm.org/workshops/CHI2022/pdfs/das_Healthcare_Chatbot_Privacy_and_Security.pdf

CHAPTER 5 ARTICLE 2: THE FIRST AI-BASED CHATBOT TO PROMOTE HIV SELF-MANAGEMENT: A MIXED METHODS USABILITY STUDY

Yuanchao Ma[1,2,3,4], Sofiane Achiche[1], Gavin Tu[5], Serge Vicente[2,3,6,7], David Lessard[2,3,4], Kim Engler[2,3], Benoît Lemire[4,8], MARVIN chatbots Patient Expert Committee, Moustafa Laymouna[2,3,6], Alexandra de Pokomandy[2,3,4,6], Joseph Cox[2,3,4,9], Bertrand Lebouché[2,3,4,6]

1. Department of Biomedical Engineering, Polytechnique Montréal, Montreal, Quebec, Canada
2. Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
3. Infectious Diseases and Immunity in Global Health Program, Research Institute of McGill University Health Centre, Montreal, Quebec, Canada
4. Chronic Viral Illness Service, Division of Infectious Disease, Department of Medicine, McGill University Health Centre, Montreal, Quebec, Canada
5. Faculty of Medicine, Université Laval, Quebec, Canada
6. Department of Family Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada
7. Department of Mathematics and Statistics, University of Montreal, Montreal, Quebec, Canada
8. Department of Pharmacy, McGill University Health Centre, Montreal, Canada
9. Department of Epidemiology, Biostatistics, and Occupational Health, Faculty of Medicine and Health Sciences, McGill University, Montreal, Canada

Published in *HIV medicine*, Volume 26, Issue 2, Pages 184-206, September 2024

5.1 Abstract

Background: We developed MARVIN, an artificial intelligence (AI)-based chatbot that provides 24/7 expert-validated information on self-management related topics for people with HIV. This study assessed (1) the feasibility of using MARVIN, (2) its usability and acceptability, and (3) four usability subconstructs (perceived ease of use, perceived usefulness, attitude towards use, and behavioural intention to use).

Methods: In a mixed-methods study conducted at the McGill University Health Centre, enrolled participants were asked to have 20 conversations within 3 weeks with MARVIN on predetermined topics and to complete a usability questionnaire. Feasibility, usability, acceptability, and usability subconstructs were examined against predetermined success thresholds. Qualitatively, randomly selected participants were invited to semi-structured focus groups/interviews to discuss their experiences with MARVIN. Barriers and facilitators were identified according to the four usability subconstructs.

Results: From March 2021 to April 2022, 28 participants were surveyed after a 3-week testing period, and nine were interviewed. Study retention was 70% (28/40). Mean usability exceeded the threshold (69.9/68), whereas mean acceptability was very close to target (23.8/24). Ratings of attitude towards MARVIN's use were positive (+14%), with the remaining subconstructs exceeding the target (5/7). Facilitators included MARVIN's reliable and useful real-time information support, its easy accessibility, provision of convivial conversations, confidentiality, and perception as being emotionally safe. However, MARVIN's limited comprehension and the use of Facebook as an implementation platform were identified as barriers, along with the need for more conversation topics and new features (e.g., memorization).

Conclusions: The study demonstrated MARVIN's global usability. Our findings show its potential for HIV self-management and provide direction for further development.

Keywords: antiretroviral, artificial intelligence, Canada, chatbot, conversational agent, digital health, feasibility, HIV, implementation science, mixed methods, mobile phone, patient and stakeholder engagement, self-management, telehealth, usability

Trial Registration: ClinicalTrials.gov NCT05789901:

https://classic.clinicaltrials.gov/ct2/show/NCT05789901

5.2 Introduction

5.2.1 Background

In 2022, around 39 million people were living with HIV globally [1]. In Canada, the estimated number of new HIV diagnoses in 2022 stood at 1,833, a 25% increase compared to 2021 [2]. Four decades into the global HIV pandemic, effective antiretroviral therapy (ART) has significantly improved the life expectancy of people with HIV (PWH), closing the gap with HIV-negative individuals [3]. HIV is now a manageable chronic disease that [4], nevertheless, requires lifelong self-management, including engagement in beneficial behaviors such as attending regular healthcare appointments, acquiring self-management related knowledge, and developing decision-making skills [5]. Adherence to ART is especially important to maintain viral suppression and thereby avoid complications and forward transmission [6]. However, self-management and ART adherence are challenged by a variety of factors, including medication beliefs and concerns (lack of understanding of treatment, side effects), lifestyle (disrupted routine, substance use), interpersonal relationships (fear of disclosure, stigma), and healthcare-related factors (patient-provider communication) [7].

Digital health, the use of information technology (IT) to manage illnesses and promote wellness, can provide innovative solutions to help PWH address these self-management barriers. A systematic review revealed that telephone- and website-supported counseling and messaging facilitated remote access and timely information exchange with PWH, resulting in effective improvements in medication adherence, coping with HIV-related conditions and management of side effects [8]. A Canadian study also established that weekly text message follow-up improved medication adherence and viral suppression among PWH [9]. While minimizing travel costs, saving time and protecting privacy [8], IT-assisted interventions could enable swift access to reliable health information for PWH and optimize their self-management practices.

Within health-related IT, chatbots are among the most promising tools, with the potential to revolutionize patient self-management and support [10]. Chatbots are acceptable by patients [11, 12], and can establish a good collaborative connection with them [13-15], thus promoting active engagement in care. Since the first appearance of the ELIZA chatbot as a psychotherapist in 1966 [16], chatbots have been explored in a variety of healthcare applications across mental health [17-21], oncology [22-26], and diabetes [27, 28]. Chatbots often harness the power of artificial intelligence (AI) to enable natural language interpretation as well as aid decision-making. They have been found to be easily accessible and able to provide valid information quickly while ensuring anonymity [29, 30]. This is well suited for PWH who require lifelong self-management and are still reluctant to disclose their condition for fear of stigmatization [31]. Yet, relevant chatbot work remains modest [32, 33]. Brixey *et al.* implemented SHIHbot on Facebook in 2017, the first chatbot to provide HIV-related sexual health information [34]. Ardiana *et al.* also presented a mobile-based chatbot for HIV/AIDS information and counseling [35]. User satisfaction was high (3.6/4) as was usability (3.3/4), indicating user endorsement of the overall system concept. Apart from these two forays into information provision, most other studies in this area focused on HIV prevention [11, 36-41]. The team of Van Heerden *et al.* tested the use of chatbots for rapid HIV self-testing and counseling in South Africa in 2017 and 2022 [11, 36], with the majority of testers reporting their intention to use such technology due to the privacy and anonymity it afforded compared to human counselors. Cheah *et al.* reached similar conclusions, where participants in Malaysia perceived chatbots as helpful to avoid stigma-inducing interactions and found chatbots to be a useful tool for HIV self-testing and pre-exposure prophylaxis information [40]. Along with Yam *et al.* [37], both studies noted the need for their chatbot to have more HIV-related information, especially on ART treatment and mental health support, to ensure the subsequent successful implementation and rollout [40].

Despite these sporadic yet encouraging efforts in HIV prevention, self-management by PWH remains an essential aspect of care that has not seen similar advancements. Starting in 2020, our multidisciplinary team including patient partners, healthcare professionals, engineers and researchers collaborated on the development of an AI-based bilingual chatbot named MARVIN (Minimal AntiRetroViral INterference). Created using a co-design approach with patient and stakeholder engagement [42], MARVIN can converse on issues of HIV self-management, answering with expert-validated information on: ART administration, ART management while traveling, and general HIV-related knowledge. It can also provide medication reminders.

5.2.2 Aim and objectives

To the best of our knowledge, there are no published studies on chatbot use to facilitate self-management among PWH, including ART adherence [43]. Furthermore, despite the growing interest in healthcare chatbots, the extent to which people find them useful needs to be thoroughly evaluated [44]. To bridge this knowledge gap, the primary objectives of this study were to 1) assess the feasibility of using the new MARVIN chatbot by PWH; and 2) gauge MARVIN's global usability. Its secondary objective was to 3) further determine its usability in terms of four subconstructs and their interrelationships, i.e., perceived ease of use, perceived usefulness, attitude towards use, and behavioral intention to use.

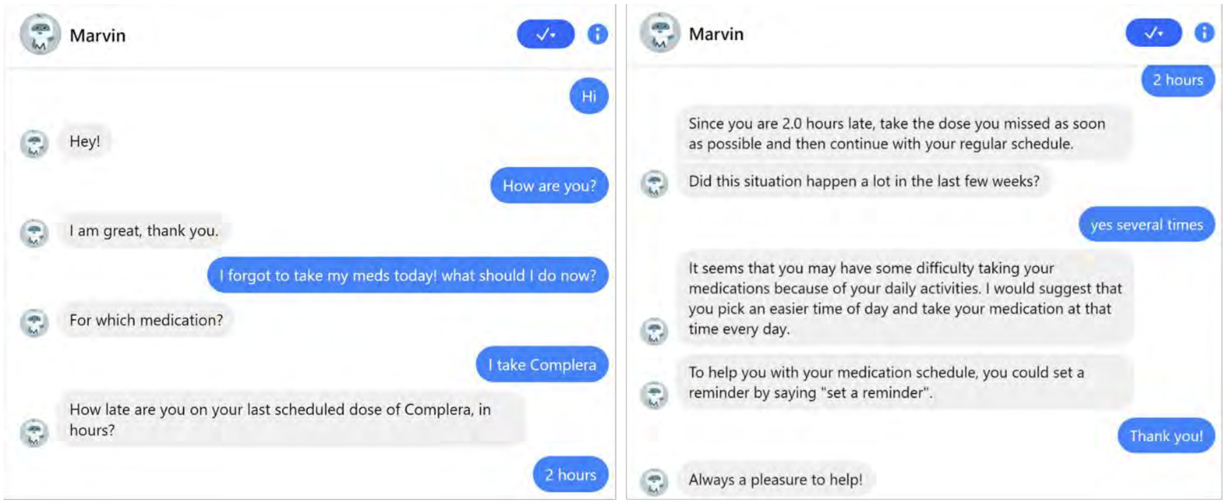
5.3 Methods

5.3.1 The MARVIN Chatbot intervention

Described in detail in a previous publication [42], MARVIN is an AI-based bilingual chatbot that communicates with PWH in either English or French, offering them advice on issues of self-management though brief text-based conversations. During the study, MARVIN operated 24/7 on Facebook Messenger for free, accessible exclusively to study participants. No updates were made to the chatbot, and no third-party human was involved in the interactions between MARVIN and participants.

MARVIN begins its first conversation with the user by explicitly introducing itself as a bot and asking for the user's preferred language. It then describes how it handles data, how accounts can be deleted, and the intended use of the chatbot. A complete list of conversation topics can be found in Multimedia Appendix 1. The HIV-related conversations mainly cover the following topics:

1) ART administration: MARVIN can address issues related to time management, dosing, common drug interactions, medication storage, and medication identification. Figure 5.1 shows an example conversation with MARVIN on forgetting to take a medication.



Hi

Hey!

How are you?

I am great, thank you.

I forgot to take my meds today! what should I do now?

For which medication?

I take Complera

How late are you on your last scheduled dose of Complera, in hours?

2 hours

2 hours

Since you are 2.0 hours late, take the dose you missed as soon as possible and then continue with your regular schedule.

Did this situation happen a lot in the last few weeks?

yes several times

It seems that you may have some difficulty taking your medications because of your daily activities. I would suggest that you pick an easier time of day and take your medication at that time every day.

To help you with your medication schedule, you could set a reminder by saying "set a reminder".

Thank you!

Always a pleasure to help!

Figure 5.1 A conversation with MARVIN on forgetting to take a medication.

2) ART management while traveling: MARVIN can specify whether a particular country has immigration restrictions for PWH or vaccination requirements, and how to prepare and carry ART medications for travel, and deal with time zone differences.
3) General HIV-related knowledge: MARVIN can also converse on common HIV symptoms, modes of transmission and prevention, as well as routine vaccination recommendations for PWH.
4) Medication reminders: on users' request, MARVIN can send daily reminders to take medication. For privacy reasons, reminders can be customized, changed, or deleted at any time. As an example, a user can ask MARVIN to send the message "Time for a walk!" as a reminder, thus avoiding disclosing their HIV status.

In terms of AI algorithms, MARVIN was developed using the Rasa framework [45] and employs a variety of algorithms, mainly intent classification and entity extraction (e.g., identifying time, drug name, or quantity), as well as decision trees for dialog management. The corresponding decision tree nodes lead to different messages pre-defined by our multidisciplinary team. When unable to understand the user's intent or reach a specific intent related to diagnosis or treatment, MARVIN acknowledges its limits and encourages the participant to contact their healthcare provider.

5.3.2 Study design

This four-week uncontrolled single-group usability study employed a mixed methods convergent design [46]. It simultaneously collected qualitative and quantitative data and combined them to assess usability outcomes of the MARVIN chatbot. We reported our findings in accordance with the CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidelines [47] (Multimedia Appendix 2).

5.3.3 Ethical considerations

This study received approval from the McGill University Health Centre (MUHC) research ethics board on April 9, 2021 (approval 2021-7191).

5.3.4 Settings and participants

This study was conducted at the Chronic Viral Illness Service of the MUHC – Glen site, located in Montreal, Quebec, Canada. Currently, over 2300 patients, about 40% of whom are women, are being followed up at the CVIS clinic.

5.3.5 Eligibility criteria

Participant inclusion criteria were as follows: 1) age of ≥ 18 years; 2) fluency in English or French; 3) confirmed diagnosis of HIV infection; 4) on ART; 5) access to a smart device (e.g., smartphones); 6) access to an Internet connection; 7) acceptance to use or create a personal Facebook account; and 8) acceptance of Facebook's privacy and data security policies. Exclusion criteria were not meeting inclusion criteria, being hospitalized, concurrent enrollment in another study involving chatbots, or having a cognitive impairment that prevented participation.

5.3.6 Sample recruitment

The target sample size for the quantitative component was 30 participants, as recommended for one-group pilot studies [48, 49].

Using a convenience sampling strategy [50], healthcare providers asked their patients at an in-clinic or remote follow-up visit if they were interested in participating in the study. The study coordinator then contacted interested individuals to determine eligibility and, if so, proceeded to informed consent. All participants were given detailed oral and written information describing the study procedures, anticipated benefits, and potential risks. Patients who consented to participate were asked whether they agreed to also partake in the qualitative component of the study.

5.3.7 Study procedures

Table 5.1 presents the main study procedures for the patient participants and their schedule.

After consent, a training session with a dedicated digital coordinator provided the participant with access to MARVIN on Facebook as well as, instructions for its use (syntax, question stems, etc.), and helped engage them in a conversation with MARVIN to familiarize them with the process.

Table 5.1 Study procedures - MARVIN usability study.

Study procedure	At entry	Week 1	Week 2	Week 3	Week 4
Preparation					
Screening and consent process	✓				
Training session	✓				
Data collection					
Sociodemographic questionnaire	✓				
Usability testing					
At least 20 conversations with MARVIN	✓	✓	✓	✓	
Quantitative component					
Usability questionnaire				✓	
Qualitative component					
2-hour focus groups or 1-hour interview					✓

Participants were required to complete a sociodemographic questionnaire following enrollment. They then began a 3-week usability test. This involved initiating 5 to 10 conversations with MARVIN at any given time on each of the following topics, by asking questions of their own creation: (1) advice for taking ART; (2) traveling with ART; and (3) vaccination recommendation for PWH. If MARVIN did not receive input from participants within a week's time, a standardized reminder was sent asking if everything is okay. Upon completion, the coordinator would inform them via email to fill out the study questionnaire and plan for a focus group or interview.

At study completion, if participants did not wish to maintain their Facebook Messenger conversations with MARVIN, the team would help them delete all records in accordance with the relevant Facebook Messenger privacy policy [51-53]. We compensated participants CAD $30 for completing the usability testing and the study questionnaire, and an additional CAD $30 for participating in the focus group discussions/interviews.

5.3.8 Guiding framework and hypotheses

Usability refers the extent to which a product can be used to achieve specific goals with effectiveness, efficiency, and satisfaction [54]. Poor usability of digital health technology can diminish work system performance, increase error rates and cause harm to patients [55]. Acceptability represents how agreeable, palatable, or satisfactory an intervention is perceived to be, which collectively affects the final adoption rate and impacts its eventual implementation and use [56, 57]. Given the overlap between these concepts, we decided to treat them both as indicators of "global usability", our primary outcome.

Concurrently, we adopted the Technology Acceptance Model (TAM) to deepen our assessment of usability, both quantitatively and qualitatively. The TAM is a frequently used and validated conceptual framework for explaining the actual use and acceptance of new information technology interventions in healthcare [58, 59]. Figure 5.2 specifies the four main subconstructs of the TAM, namely perceived ease of use (PEU), perceived usefulness (PU), attitude towards use (ATU), and behavioral intention to use (BIU) [60]. Three of these subconstructs are analogous to the ISO-defined aspects of usability (i.e. effectiveness = PU, efficiency = PEU, satisfaction = ATU) [61], whereas BIU denotes actual system use [62, 63]. They are thus appropriate for assessing different components of usability. We tested the following five hypotheses that were consistent with the original TAM and extant research [64-66]:

H1: Perceived ease of use is positively associated with perceived usefulness.

H2: Perceived ease of use is positively associated with attitude towards use.

H3: Perceived usefulness is positively associated with attitude towards use.

H4: Perceived usefulness is positively associated with behavioral intention to use.

H5: Attitude towards use is positively associated with behavioral intention to use.

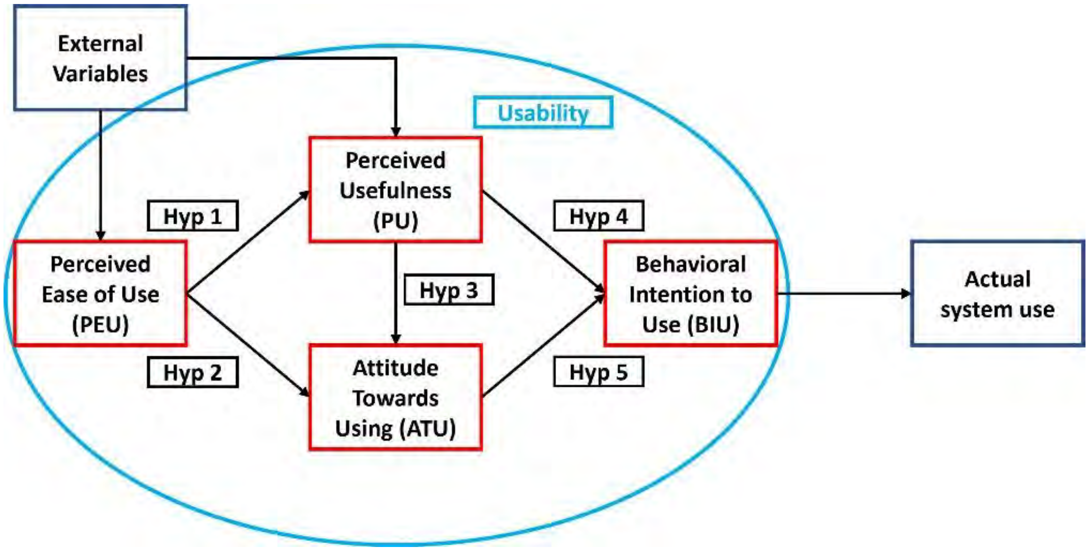

External Variables
Usability
Perceived Usefulness (PU)
Hyp 1
Hyp 4
Perceived Ease of Use (PEU)
Hyp 3
Behavioral Intention to Use (BIU)
Actual system use
Hyp 2
Attitude Towards Using (ATU)
Hyp 5

Figure 5.2 Guiding framework: the Technology Acceptance Model (TAM), adapted from Davis et al. (1989).

5.3.9 Quantitative data

5.3.9.1 Data collection

A sociodemographic questionnaire was administered to describe the study sample's characteristics and diversity (e.g., age, preferred language, gender, ethnic group identity) as well as the participants' use of mobile devices, health apps, and Facebook Messenger (Multimedia Appendix 3). All study questionnaires (Multimedia Appendix 3) were administered via Google Forms.

5.3.9.1.1 Primary outcome

Concerning feasibility, we documented reasons for refusal to participate or screening failures with a designated refusal form. We examined the recruitment rate (i.e., the proportion of eligible contacts enrolled in the study), and the retention rate (i.e., the proportion of participants who completed the usability test and study questionnaires).

Data on global usability was collected with two validated scales.

The first scale is a slightly adapted version of the *Usability Metric for User Experience-lite (UMUX-lite)* [67]. It is based on the well-established *System Usability Scale (SUS)* [68]. While a standardized score of 68 is generally considered the baseline for a usable tool, UMUX-lite yields scores that are 99% predictive of SUS scores, with the benefit of only two 7-point Likert scale items. UMUX-lite is also considered suitable for healthcare technology assessment [69]. A corrective regression formula is used to align its scores with those of the SUS [67].

The second scale, a revised version of the 6-item *Acceptability E-Scale (AES)*, is designed to evaluate computer-based interventions for healthcare populations [70]. Items, rated on a 5-point Likert scale, are summed to create a global score (range: 6-30), with 24 recommended by the developers as the acceptability threshold.

5.3.9.1.2 Secondary outcome

Four subconstructs of the TAM were assessed with validated instruments.

Perceived ease of use (PEU) was assessed through an adapted 7-point Likert scale Single Ease of Use Question (SEQ) [71].

Perceived usefulness (PU) was measured using four items, adapted from the tools of Chau & Hu [72] and Davis [73]. Rated on a 7-point Likert scale, item ratings were averaged to produce a PU score.

To evaluate attitude towards using (ATU) MARVIN, the Net Promoters Score (NPS) [74] was adopted with a single 11-point Likert scale question. The final NPS score is the percentage of detractors (score of 0-6) subtracted from promoters (9-10). Positive scores, especially those over 50%, are judged positively.

Behavioral intention to use (BIU) was measured with two validated 7-point Likert scale items [59] which were averaged to derive an intention score.

For PEU, PU, and BIU, a mean score target of 5 was considered a positive outcome [75].

5.3.9.2 Statistical analysis

All statistical analyses were conducted using Python coding language [76].

5.3.9.2.1 Sample characteristics

The sociodemographic variables were summarized with descriptive statistics. For continuous variables, we report the mean and standard deviation (SD). In the case of ordinal and nominal qualitative variables, we present both counts and proportions.

5.3.9.2.2 Primary outcome

The global usability outcomes were also summarized using descriptive statistics (i.e., the means, SD, and the range). The sample means for both the UMUX-lite and AES were confronted to their recommended minimal thresholds, 68/100 and 24/30, respectively.

Next, global usability outcomes were stratified for subgroup comparisons by selecting sociodemographic variables that emerged as important from the qualitative analysis [46], i.e., preferred language and years since diagnosed with HIV infection. The normality of the distribution was tested using the skewness and kurtosis coefficients. Student's t-tests were used to test the null hypothesis that the observed means were equal between subgroups, with a significance level of 5%.

5.3.9.2.3 Secondary outcome

The TAM subconstructs (PU, PEU, ATU, BI) were summarized using descriptive statistics (i.e., the means, SD, and the range).

We tested five hypotheses of inter-subconstruct relationships within the TAM framework using simple linear regression models. Each model's slope coefficient sign reflected the direction of the variable's (i.e., subconstruct's) association. We performed residual diagnostics for each model, and adopted appropriate strategies when the assumptions were not met. Coefficient significance was tested with a student's t-test on each slope coefficient with a 5% significance level. The 95% confidence interval for each coefficient is presented. For each model's predictive accuracy, we report the coefficient of determination R^2.

5.3.10 Qualitative data

5.3.10.1 Data collection

In conjunction with the study questionnaire, three semi-structured focus group interviews were planned to further explore participants' experiences with MARVIN. The choice of three focus groups follows recent work suggesting that this would capture at least 80% of the themes, which we considered sufficient saturation for our usability study [77]. The interviews were designed following the subconstructs of the TAM framework. Additionally, participants were asked about future improvements they would like to see with MARVIN. The focus group interview guide can be found in Multimedia Appendix 4.

Focus group interviews were held through Zoom videoconferencing. Due to limitations imposed by the COVID-19 pandemic, some interviews ended up being conducted individually. YM and GT conducted the interviews, which lasted 15 to 85 minutes each and were audio-recorded.

5.3.10.2 Analysis

All interviews were transcribed verbatim and de-identified. All transcripts were cross-checked by YM and GT for accuracy and completeness while both attempted to understand the entire dataset through immersion in the data.

A comprehensive coding matrix based on the TAM and the Consolidated Framework for Implementation Research (CFIR) [78] was used for deductive thematic analysis. The CFIR is a commonly used conceptual framework for identifying factors that might influence intervention implementation and effectiveness. It consists of five broad domains (i.e., intervention characteristics, outer setting, inner setting, characteristics of individuals and process) and 39 subdomains. Given that "characteristics of individuals" and "process" focused more on scaling up implementation, we focused our analysis on the first three subdomains to illustrate relevant barriers and facilitators. Using the NVivo R1 software, initial codes were generated deductively using the CFIR subdomains. We then matched identified facilitators and barriers to the subconstructs of the TAM to determine their impact on global usability.

To ensure credibility and reliability, results were debriefed and discussed repeatedly with DL, ML and KE, and triangulated with quantitative findings. Illustrative quotes in French presented in this manuscript were translated to English by co-authors.

5.4 Results

5.4.1 Overview of participation

The participant flow chart is shown in Figure 5.3. From March 30 to December 2, 2021, a total of 88 PWH were screened for participation in the study, including 34 ineligible and 54 eligible individuals. The most common reasons for exclusion were their being unable to commit to the entire study period or not having a Facebook account. Forty individuals were enrolled in the study, of which 12 withdrew. All 28 participants who completed at least 20 conversations within 3 weeks were included in the analysis.

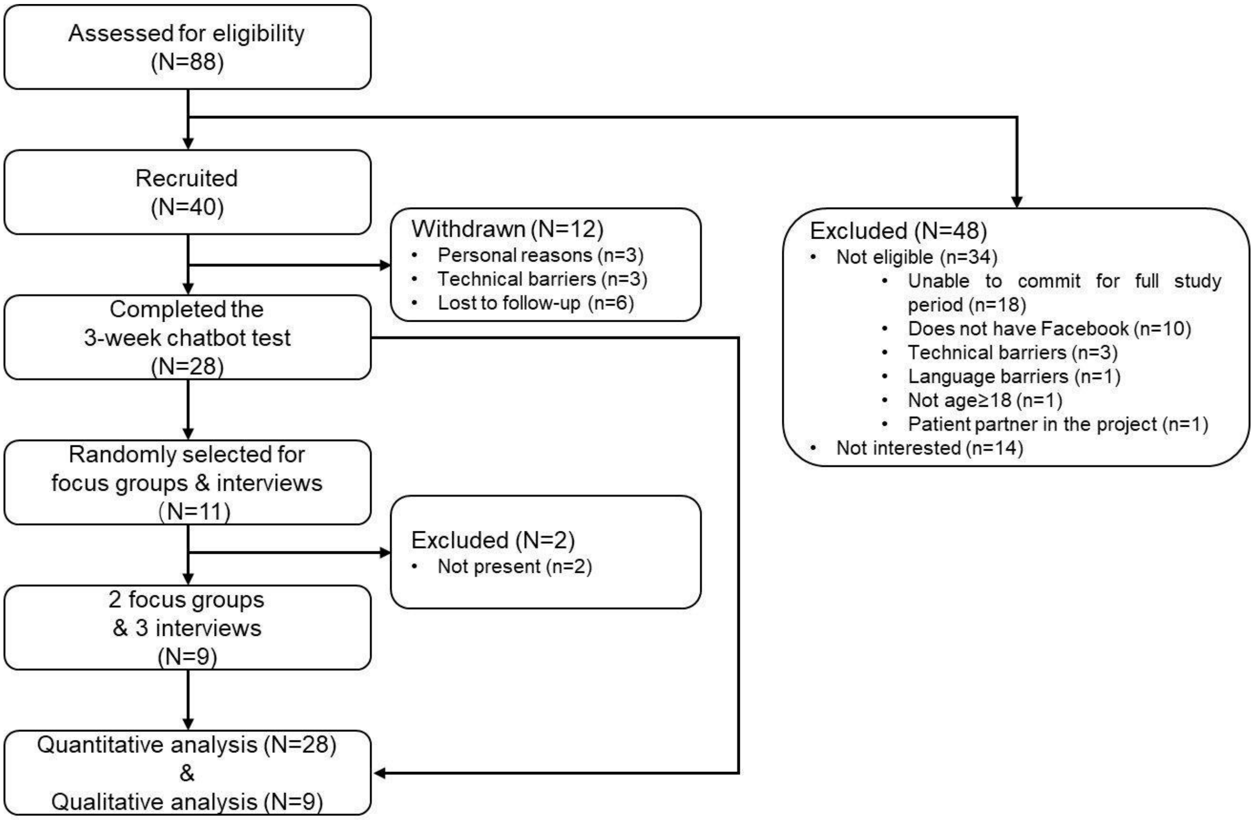

Assessed for eligibility (N=88)
Recruited (N=40)
Withdrawn (N=12)
• Personal reasons (n=3)
• Technical barriers (n=3)
• Lost to follow-up (n=6)
Completed the 3-week chatbot test (N=28)
Randomly selected for focus groups & interviews (N=11)
Excluded (N=2)
• Not present (n=2)
2 focus groups & 3 interviews (N=9)
Quantitative analysis (N=28) & Qualitative analysis (N=9)
Excluded (N=48)
• Not eligible (n=34)
• Unable to commit for full study period (n=18)
• Does not have Facebook (n=10)
• Technical barriers (n=3)
• Language barriers (n=1)
• Not age≥18 (n=1)
• Patient partner in the project (n=1)
• Not interested (n=14)

Figure 5.3 Participants flow diagram throughout the study.

5.4.2 Sample characteristics

Table 5.2 describes the baseline sociodemographic characteristics of the study sample. Notably, most participants (23/28, 82.1%) self-identified as men. Preferred language was evenly distributed between English (15/28, 53.6%) and French (13/28, 46.4%). Nearly half of participants (13/28, 46.4%) reported having a University degree. More than half (17/28, 60.7%) earned less than 40,000 $CAD per year.

In terms of experience with technology, almost two-thirds (18/28, 64.3%) had moderate experience with health-related apps. Nearly all participants (26/28, 92.9%) used their mobile devices several times a day and 89.3% (25/28) had used Facebook Messenger for more than two years.

Table 5.2 Sociodemographic characteristics of the sample (N=28).

Age (years) [a]	Mean (SD): 40.2 (11.5)
Years diagnosed with HIV [a]	Mean (SD): 8.2 (8.2)
Gender	No. (%)
Men	23 (82.1)
Women	5 (17.9)
Preferred language	
English	15 (53.6)
French	13 (46.4)
Sexual orientation*	
Heterosexual/Straight	6 (21.4)
Lesbian	1 (3.6)
Gay	19 (67.9)
Queer	2 (7.1)
Bisexual	4 (14.3)
Questioning	1 (3.6)
Other	1 (3.6)
Ethnic groups*	
English Canadian	1 (3.6)

French Canadian	6 (21.4)
British	1 (3.6)
Other Eastern/Western European	3 (10.7)
West Asian	1 (3.6)
Arab or North African	3 (10.7)
Latin American	10 (35.7)
African	4 (14.3)
Black	5 (17.9)
Caribbean	1 (3.6)
Mixed race/ethnicity	1 (3.6)
Highest education level	
Secondary (High school)	2 (7.1)
Professional degree/College	6 (21.4)
CEGEP/Technical degree	4 (14.3)
University	13 (46.4)
Other	3 (10.7)
Annual income ($CAD)	
Less than $10,000	2 (7.1)
$10,000 - $ 19,999	5 (17.9)

$20,000 - $ 39,999	10 (35.7)
$40,000 - $ 59,999	7 (25.0)
$60,000 - $ 79,999	2 (7.1)
$80,000 - $ 99,999	1 (3.6)
Greater than $100,000	1 (3.6)
Most frequently used mobile device	
Android	11 (39.3)
Apple iPhone	15 (53.6)
Tablet	1 (3.6)
Other **	1 (3.6)
Confident in the effective use of mHealth platforms/electronic surveys	
Strongly disagree	4 (14.3)
Disagree	0 (0)
Neutral	1 (3.6)
Agree	12 (42.9)
Strongly agree	11 (39.3)
Extent of use of health-related apps	
Never	4 (14.3)
Very little	6 (21.4)

Sometimes	11 (39.3)
Frequently	6 (21.4)
Very frequently	1 (3.6)
Frequency of use of mobile devices	
Several times a day	26 (92.9)
Once a day	0 (0)
Several times per week	2 (7.1)
Several times per month	0 (0)
Experience with Facebook Messenger	
Less than 2 years	3 (10.7)
2-4 years	4 (14.3)
4-6 years	2 (7.1)
6-8 years	3 (10.7)
8-10 years	2 (7.1)
More than 10 years	14 (50.0)

[a] No missing data

* Multiple-choice question

** One participant answered “Laptop.”

5.4.3 Quantitative results

5.4.3.1 Primary endpoints: Feasibility and global usability

Regarding feasibility, 40 participants were recruited among 54 eligible contacts, with a study recruitment rate of 74% (40/54). The retention rate was 70% (28/40) with 28 completing the three-week test.

Table 5.3 shows the descriptive statistics for the UMUX-lite and AES. The mean value for the UMUX-lite was 69.9 (SD = 14.2), which surpassed the threshold of 68. The mean value of the AES was 23.8 (SD = 4.9), which was close to the expected mean threshold of 24. Checking the distribution of questionnaire data revealed one outlier. Excluding the outlying participant made both measures exceed their respective threshold, with no effect on the remaining results.

Subgroup comparisons on both global usability measures are presented in Table 5.3. By skewness-kurtosis test, all normal distributions of the subgroups held. Considering the preferred language, a greater proportion of participants rated the English version of MARVIN over threshold on both the UMUX-lite and AES (reported in Table 3 (60.0% (9/15) and 66.7% (10/15), respectively), compared to the French version (46.2% (6/13) for both). However, no significant differences were found between the means of both measures (*P*=.98 and *P*=.51) for the English and French subgroups.

Regarding years since diagnosis with HIV infection, among participants diagnosed ≤5 years, 57.1% (8/14) had mean UMUX-lite scores above the threshold, and 78.6% (11/14) had above-threshold scores for the AES. In comparison, these values were 50.0% (7/14) and 35.7% (5/14), respectively for participants diagnosed >5 years. There was no statistically significant difference (*P*=.07) between the two subgroups in terms of UMUX-lite. On the AES, the mean of participants diagnosed ≤5 years was significantly higher than that of those diagnosed >5 years (*P=.009*).

Table 5.3 Primary endpoints: global usability and subgroup comparisons.

Global usability

Measure	**Range**	**Sample**	**Mean (SD)**	**Threshold**	**>Threshold n (%)**
UMUX-lite	[12.1, 87.9]	**N=28**	69.9 (14.2)	68.0	15 (53.6)
AES	[6, 30]		23.8 (4.9)	24.0	16 (57.1)
UMUX-lite	[12.1, 87.9]	**N=27**	71.4 (11.9)	68.0	15 (55.6)
AES	[6, 30]		24.2 (4.5)	24.0	16 (59.3)

Subgroup comparisons

Measure	**Subgroup**	**Sample size**	**Mean (SD)**	**P value***	**>Threshold n (%)**
UMUX-lite	**Preferred language**				
	English	15	69.8 (17.0)	.98	9 (60.0)
	French	13	70.0 (10.9)		6 (46.2)
	Years diagnosed with HIV				
	≤5 years	14	74.7 (11.4)	.07	8 (57.1)
	>5 years	14	65.1 (15.5)		7 (50.0)
AES	**Preferred language**				
	English	15	24.4 (5.5)	.51	10 (66.7)
	French	13	23.2 (4.3)		6 (46.2)
	Years diagnosed with HIV				
	≤5 years	14	26.1 (3.6)	*.009*	11 (78.6)
	>5 years	14	21.5 (5.0)		5 (35.7)

* P values were calculated to test the null hypothesis that the observed means are equal between subgroups.

UMUX-lite: Usability Metric for User Experience-lite

AES: Acceptability E-Scale

5.4.3.2 Secondary outcomes: Usability subconstructs

Results for the usability subconstructs are shown in Table 5.4. Means for PEU, PU, and BIU all exceeded the target score of 5 on 7. The NPS, which measured attitude towards MARVIN use, was equal to 14%, representing positive user ratings. Significant positive associations were found for all five hypotheses in the expected direction (*P=.024* and *P=.002* for H1 and H2, and P<*.001* for H3-H5, respectively). Corresponding scatterplots can be found in Multimedia Appendix 5. Perceived ease of use explained 18% and 31% of the variance in perceived usefulness and attitude towards use. Perceived usefulness explained 71% and 59% of the variance in attitude towards use and behavioral intention to use. Attitude towards use, on the other hand, explained 68% of the variance in behavioral intention to use.

Table 5.4 Secondary endpoints: subconstructs of the TAM and their associations (N=28).

Secondary endpoint	**Score range**	**Mean (SD)**		
Perceived ease of use (PEU)	[1, 7]	5.6 (1.3)		
Perceived usefulness (PU)	[1, 7]	5.1 (1.6)		
Attitude toward use (ATU) *	[0, 10]	7.4 (2.9)		
Net Promoters Score *	[-100%, 100%]	14%		
Behavioral intention to use (BIU)	[1, 7]	5.2 (2.0)		
Association between TAM subconstructs				
Hypothesis	**Slope**	**P value****	**95% CI**	**R^2**
H1 – PU=f(PEU)	0.53	*.024*	0.08-0.98	0.18
H2 – ATU = f(PEU)	1.22	*.002*	0.49-1.95	0.31
H3 – ATU= f(PU)	1.48	*<.001*	1.10-1.86	0.71
H4 – BIU = f(PU)	0.94	*<.001*	0.61-1.28	0.59
H5 – BIU=f(ATU)	0.57	*<.001*	0.42-0.73	0.68

* We present the numerical results for ATU, and the final scores calculated using the NPS method.

** P values were calculated to test the null hypothesis that there is no linear relationship between both variables.

5.4.4 Qualitative results

5.4.4.1 Overview

Between August 2021 and April 2022, a total of 11 participants were recruited to participate in two focus groups (1 in English, 1 in French, with 3 participants per group) and three one-on-one interviews (1 in English and 2 in French). Two participants did not attend the sessions, leaving nine. The thematic analysis identified 20 themes, representing 12 facilitators and 10 barriers to global usability, as presented in Table 5.5 with selected participant quotes. The remaining quotes contributing to the analysis can be found in Multimedia Appendix 6.

Table 5.5 Themes identified with corresponding quotations, facilitators, and barriers.

Domain	Subdomain	Theme	Quotations	Facilitator	Barrier
Implementation characteristics	Intervention source	Reliable information	- We have a very major point that Google might be wrong, where you might be referred to many different answers and you don't know what the right answer is. And here [MARVIN] we're more confident that the information that has been offered to us is more accurate because it was professionally done. (P#1 Focus group, English, Age 32, Man)	✓	
	Evidence strength and quality	Easy-access, go-to tool	- It's a reference tool for me. Every time I have a question, I'd have the reflex to go and ask my question to this chatbot and see its answer. (P#3 Interview, French, Age 56, Man) - Because there's a lot of stuff I forgot about. So, I want a refresher. [MARVIN] is a good reference instead of seeing someone, waiting for a doctor's appointment. (P#3 Focus group, English, Age 40, Man)	✓	
		Useful real-time support	- I find it really useful. The speed of response, and it also answers the majority of questions relating to this treatment,	✓	

		vaccination, everything to do with HIV. I think it's great. (P#1 Focus group, French, Age 28, Man) - It was useful also for travelling and it [MARVIN] gave advice about timing and time zones. It's helpful. (P#3 Focus group, English, Age 53, Man)	
	Convivial conversation	- But it's super easy to use, it really is a friend you can share a question with and get an answer. (P#3 Focus group, French, Age 40, Man) - I'm satisfied with it [MARVIN] because it's like you're chatting with a friend…(P#1Interview, English, Age 37, Woman) - To me, I like the fun aspect of chatting with that [MARVIN], it makes it fun to make a question and get an answer. (P#2 Focus group, English, Age 32, Man)	✓
	Emotionally safe	- No one is, what can I say, judging you… This platform [MARVIN] is good because there's someone you can share with… I think MARVIN does not ask [me] to identify myself, so I think it's a safe space. (P#1 Interview, English, Age 37, Woman) - Let's say a category of questions that I would call 'shame questions' like: "What happens if I decide to not wear a condom anymore?" Talking to an A.I. that has this kind of answer would be very helpful so that you have the right answer, and you don't do something stupid, but you also don't feel guilty about asking this kind of question. (P#1 Focus group, English, Age 32, Man)	✓
	Confidential	- There is a value that you should present more that MARVIN is confidential, that it	✓

		does not share information with other researchers, with other people. (P#3 Focus group, French, Age 40, Man)		
	Lack of conversation topics	- For a start, it's very good. On the other hand, I did ask some more detailed questions, but unfortunately, I didn't get the answers [I wanted]! (P#2 Focus group, French, Age 34, Man) - Because there's so many things we would want to ask. So, if you can go deeper and make MARVIN be available to answer, ask everything, it would help. (P#2 Interview, English, Age 53, Woman)		✓
	Limited understanding of user input	- I asked: "What is an undetectable viral load?"… MARVIN didn't have a good answer, but I rephrased to "What does undetectable mean?" And that [answer] was exactly what I wanted… (P#1 Focus group, English, Age 32, Man) - For me, the thing MARVIN can improve is his language. He doesn't understand everything other people say. All that is to make it easier for other people to use and maybe if you could do it in languages apart from English and French. (P#1 Focus group, French, Age 28, Man)		✓
	Useful reminders but still room for improvement	- So, it can remind us to take our medication… It's very useful for me. (P#3 Interview, French, Age 56, Man) - I think it could be very useful to use as a reminder for multiple things like… It's been 3 months or 6 months, shouldn't you do bloodwork again? …These kinds of reminders as well. (P#1 Focus group, English, Age 32, Man)	✓	
	Desired features	- To make MARVIN easier to use, is it possible for MARVIN to have memories?		✓

	beyond conversation content	Because, for example, when I ask him about medication, he always asks me the same question: “What medication?”. (P#3 Focus group, French, Age 40, Man) - MARVIN should, in the future, if it starts to detect something odd in the person's questions, that it can also act as an alert for specialists. I think it would be useful for the specialists. (P#3 Focus group, French, Age 40, Man)		
	Lack of proactivity	- It’s very passive right now: you need to ask, and it responds. But if it could ask you and sort of come up with an answer based on the questions. (P#3 Focus group, English, Age 53, Man)		✓
	Allows for health-related teaching moments among peers or friends	- I once used MARVIN in front of a friend to explain what "undetectable" meant. MARVIN responded quickly, explaining everything. I think this is another advantage of MARVIN. (P#1 Focus group, French, Age 28, Man)	✓	
Relative advantage	*More comprehensive information available elsewhere*	- But on CATIE you can find all the information you need about travel. CATIE covers an enormous amount of information. (P#2 Focus group, French, Age 34, Man)		✓
	Trust in MARVIN vs doctor’s advice	- And I learned a few things too because the doctor said different things from what Marvin did. I really wanted to understand time zone changes. I noticed with Marvin it was pretty good for that. My doctor said something else [compared to MARVIN] to answer the question [when to take your medication when travelling] but I don’t know if I agree with that... (P#3 Focus group, English, Age 53, Man)	✓	✓

			- The subject of medication, I leave only to the doctors. (P#3 Focus group, French, Age 40, Man)		
	Adaptability	Accessible on different devices	- I use it on my PC only. I don’t use it on my phone because I decided to create a new account for that. (P#3 Focus group, English, Age 53, Man) - And it works both ways [mobile and PC] very well. As far as I'm concerned, I haven't had any problems with bugs, whether on a mobile interface or a fixed interface, laptop. (P#2 Focus group, French, Age 34, Man)	✓	
	Complexity	Preference for other platforms	- Couldn't we simply have a connection with the MUHC website, so that we could connect directly there and not have to go through Facebook. For my part, I'm a bit worried about data confidentiality. I'm a bit worried about security. (P#3 Interview, French, Age 56, Man)		✓
	Design quality & packaging	Need for instructional materials on MARVIN	- It was a bit tricky at first, as you know, yeah (laughs). You (coordinator) did help me through it… I think it’s because I’m not that good with social media and technology. (P#1 Interview, English, Age 37, Woman) - Yes, it’s useful to put a guide for everyone that uses MARVIN. Even an explanatory video on the main page or you ask MARVIN for the guide, and it gives it to you … (P#1 Focus group, French, Age 28, Man)		✓
Outer setting	Patient needs & resources	More pertinent for people who recently initiated ART	- I'm someone who's been undetectable for several years and I've been HIV positive since 2006… I know my medication very well, I know what I have to do, so I don't need to go looking for information as much as I did before. On the other hand, I can	✓	✓

		understand that someone recently diagnosed with HIV will find it useful. (P#3 interview, French, Age 56, Man) - If it’s someone that just found out about their status. For sure it’s an uncomfortable position. Maybe you don’t want to talk to a doctor, or you can’t see a doctor right away… It [MARVIN] is not really someone but it’s just something that can have a conversation with you and provide answers to the questions you have. So, for sure, it would be helpful for someone that is in the beginning of the treatment. (P#1 Focus group, English, Age 32, Man)				
	Relevant for all sexually active people, regardless of HIV status	- I think any knowledge that is basic about HIV is important to any person who engages in sexual activities, even if they have never tested positive. (P#1 Focus group, English, Age 32, Man)	✓			
Cosmopolitanism	Absence of referrals	- If it [MARVIN] is not able to answer me, but at least able to direct me to a first line, a physical person would answer me there, or saying "In 24 hours, there's someone who will answer you or call you." (P#3 Interview, French, Age 56, Man)			✓	

PEU = Perceived ease of use; PU = Perceived Usefulness; ATU = Attitude towards use; BIU = Behavioral intention to use

5.4.4.2 Implementation characteristics

5.4.4.2.1 Intervention source

Theme: Reliable information

Participants found MARVIN to be a reliable source of medical knowledge and information as it was validated by expert health professionals. Users felt that responses were accurate and more trustworthy than what they would normally find on common search engines (e.g., Google). This theme was identified a facilitator to perceived usefulness and behavioral intention to use.

5.4.4.2.2 Evidence strength and quality

Theme: Easy-access, go-to tool

MARVIN’s ease of accessibility made it a tool that participants would think of using first when they had questions that needed answering, eliminating the need to wait for their next clinical *appointment. We identified this theme as a facilitator of perceived ease of use and behavioral intention to use.*

Theme: Useful real-time support

MARVIN’s ability to respond instantly worked to its advantage when users needed a quick response. Users would get a timely answer to a wide range of questions, which participants considered very useful. The travel-related content was especially appreciated by some participants, as they would not usually see their doctor before travelling. This theme contributed to MARVIN's perceived usefulness and users' attitudes towards its use.

Theme: Convivial conversation

The day-to-day use of MARVIN was described by participants as easy and straightforward since MARVIN functions like a normal conversation with a friend. The conversational nature of MARVIN was a further enjoyable factor for one participant. The question-and-answer interaction added an engaging and convivial atmosphere, making the chat an experience in itself. This theme was among the contributing factors to MARVIN’s perceived ease of use.

Theme: Emotionally safe

The fact that MARVIN is an AI, i.e., non-human, was said to put many participants at ease. It did not ask identifying questions and was thus considered a safe space for them to confide sensitive information. Indeed, MARVIN’s non-judgmental nature allowed them to ask questions that would be difficult to ask a doctor without fear of being judged. This promoted users' attitude towards MARVIN use.

Theme: Confidential

Confidentiality was stated by many participants as one of the most important characteristics of MARVIN. The users appreciated that the data were stored on the research institute’s internal server, accessible only to the research team. This was also one of the facilitators of attitude towards MARVIN use.

Theme: Lack of conversation topics

While participants expressed their satisfaction with MARVIN as a first release, they also pointed out that conversations with the chatbot were superficial and that some questions were not answered in a useful way. Participants wished for MARVIN to cover a wider range of topics and to delve deeper into its existing topics, which included: 1) lifestyle and behavioral factors: diet, nutrition, exercise; 2) HIV treatment: updates on new treatments and diseases, pre-exposure prophylaxis; 3) reproductive and sexual health: pregnancy, breastfeeding, sexual health behaviors, and other sexually transmitted infection-related information; 4) healthcare service support: appointments/vaccination scheduling, symptom checkers; 5) mental health support: resources on local psychologists; and 6) socioeconomic issues: immigration process, financial and insurance support. This theme was identified as a barrier to perceived usefulness.

Theme: Limited understanding of user input

In some instances, MARVIN was unable to give answers that were already in its knowledge base because it did not understand the way the questions were posed by users. Some participants reported that they had to rephrase certain questions for MARVIN to provide a related answer. On this theme, one participant emphasized the utility of offering multilingual support, which could make MARVIN more user-friendly by having it learn languages other than French and English. This point hindered both MARVIN's ease of use and users' attitude towards its use.

Theme: Useful reminders but still room for improvement

The current daily medication reminder function was used and appreciated by many participants. However, they would like this feature to be more adaptable, such as a one-time reminder that is not repeated daily, or a reminder for a 6-month hospital follow-up visit. This was one of the facilitators of MARVIN's perceived usefulness.

Theme: Desired features beyond conversation topics

It was suggested that MARVIN could be enriched to deliver information in more compelling ways, for example through videos and pictures. A few participants also suggested improving MARVIN with a working memory, so it could remember some important previous questions and answers (e.g., the medication they are taking). This would spare them the trouble of asking or answering the same questions repeatedly. Detecting anomalies in conversations and sending alerts to healthcare professionals, with the user's consent, were also raised as potentially useful features. This theme was identified as a barrier to MARVIN's perceived ease of use and perceived usefulness.

Theme: Lack of proactivity

Interactions with MARVIN currently require the user to ask questions first. Participants expressed their desire for MARVIN to be more proactive. They envisioned MARVIN asking questions, initiating conversations, and providing information without waiting for a prompt. This lack of proactivity was also seen as a barrier to MARVIN's perceived ease of use and perceived usefulness.

Theme: Allows for health-related teaching moments among peers or friends

Some participants mentioned that MARVIN was able to facilitate health discussions with their friends. Access to accurate information through MARVIN helped address some of their health concerns and played a role in peer health education. This fact contributed to MARVIN’s perceived usefulness and user's behavioral intention to use it.

5.4.4.2.3 Relative advantage

Theme: More comprehensive information available elsewhere

Participants indicated that there are other, more comprehensive sources of information than MARVIN, which currently covers a limited number of topics. These included CATIE [79], a well-known Canadian site for information on HIV treatment. This limited MARVIN’s perceived usefulness and users' behavioral intention to use it.

Theme: Trust in MARVIN vs doctor’s advice

We found that participants had different levels of trust in MARVIN versus the guidance provided by their physicians on certain topics. For example, one participant reported discrepancies between MARVIN’s response and their doctor’s recommendations on how to manage ART when time zones change, preferring MARVIN's advice. However, another participant indicated that for topics related to medication, he would defer to his doctor’s advice. This theme was identified as both a facilitator and a barrier to the behavioral intention to use.

5.4.4.2.4 Adaptability

Theme: Accessible on different devices

MARVIN is accessible through Facebook Messenger, which allows users to interact with MARVIN on their phone or portable devices and computer as well. Some participants exclusively used one specific device to access MARVIN, while others used multiple devices. This facilitated MARVIN’s perceived ease of use.

5.4.4.2.5 Complexity

Theme: Preference for other platforms

MARVIN is currently limited to Facebook Messenger and participants were concerned about the privacy issues brought up in the past with Facebook, the broader social media of which Facebook Messenger is part of. Even if they were informed that conversations could be removed from Facebook, users expressed a lack of trust. Besides, Facebook is simply not the preferred social media of many participants, and making MARVIN available on a variety of platforms would increase access. Participants suggested alternative popular social media platforms (e.g., WhatsApp), creating a MARVIN application to install on devices, or giving access directly through healthcare organizations (e.g., MUHC) website. We identified this theme as a barrier to attitude towards use and behavioral intention to use.

5.4.4.2.6 Design quality & packaging

Theme: Need for instructional materials on MARVIN

Some participants indicated that they were not familiar with the platform (i.e., Facebook Messenger) itself or social media in general. The inherent technological aspect of a chatbot is also a barrier to some individuals. These contributed to a more difficult setup process and a steeper learning curve, whereas the training session prior to usability testing was perceived as beneficial by participants. Further, they felt that a future guide or a video tutorial explaining MARVIN and how to use it would make it more accessible and less intimidating for those unfamiliar with chatbots. This theme was considered a barrier to perceived ease of use.

5.4.4.3 Outer setting

5.4.4.3.1 Patient needs & resources

Theme: More pertinent for people who recently initiated ART

Several participants who had been living with HIV for several years mentioned that MARVIN’s conversation topics were more appropriate for recently diagnosed treatment-naïve patients. They stated already knowing most of the information it provided. One interviewee also highlighted that newly diagnosed patients are more likely to be in a state of stress or unease; in this case, MARVIN could act as a conversation starter. This theme was identified as both a facilitator and a barrier to perceived usefulness and behavioral intention to use.

Theme: Relevant for all sexually active people, regardless of HIV status

For some participants, MARVIN’s potential user base could extend beyond PWH. They reported that basic HIV-related knowledge was essential for anyone who is sexually active. Participants found having conversations with MARVIN was also valuable for people who are not living with HIV. This contributed to MARVIN's perceived usefulness and users’ behavioral intention to use it.

5.4.4.3.2 Cosmopolitanism

Theme: Absence of referrals

In cases where MARVIN is unable to answer a question, some participants suggested that the chatbot provide reputable resources for advice or redirect users to specific professionals. Doing so would create a win-win situation, both by ensuring a timely response and by providing access to further assistance. This theme was identified as a barrier to perceived usefulness.

5.5 Discussion

5.5.1 Principal findings

Our team developed MARVIN, the first chatbot to promote self-management among PWH with a focus on ART adherence. The present work sought to assess 1) the feasibility of using MARVIN by PWH, 2) its global usability, as well as 3) four usability subconstructs and their interrelationships following the technology acceptance model. Quantitatively, our findings, from 28 participant PWH, support the feasibility and usability of MARVIN for the study sample. Our qualitative results showed that usability facilitators for MARVIN included the provision of reliable information and useful real-time support, as well as its easy accessibility. Participants perceived a sense of conviviality, emotional safety, and confidentiality from talking to MARVIN. However, limited understanding of user input and lack of conversation topics were identified by participants as major usability barriers to the current chatbot. It was also desired that MARVIN would offer more features, be implemented on more platforms, and provide instructional materials on its use.

Concerning feasibility, the recruitment rate was 74% (40/54), with 70% of participants completing at least 20 rounds of conversation with MARVIN in three weeks, rates comparable to other successful healthcare chatbot studies [80, 81]. This happened despite the COVID-19 pandemic, which posed a degree of challenge to the recruitment and conduct of the study, with all processes conducted remotely. Automatic weekly study reminders to participants likely facilitated the high retention rate, as found in other research [82]. Overall, we can confidently conclude that patient's use of the MARVIN chatbot was feasible in this context of implementation.

Regarding the results of the usability questionnaire, the UMUX-lite mean exceeded the preset threshold for success (69.9/68), while the AES measure was very close to target (23.8/24). Coupled with the fact that over half of participants scored above both thresholds (15/28 for UMUX-lite, 16/28 for AES), we consider that these results provide evidence of MARVIN's global usability in the study sample. Mean scores of perceived ease of use, perceived usefulness, and behavioral intention to use also surpassed the cut-off value of 5/7 and the Net Promoters Score indicated a positive attitude towards MARVIN (+14%), further substantiating its usability.

As per ease of use, the qualitative analyses revealed that participants found MARVIN was easily accessible across various interfaces such as smartphones, tablets, and laptops. Furthermore, almost everyone in our sample (26/28) used mobile devices multiple times per day, and the vast majority (23/28) were confident at baseline in using mHealth platforms effectively and had experience with similar applications (18/28). These findings as well as the growing popularity of mobile technology may have contributed to the high ease of use observed in this study, which is consistent with previous studies [11, 41, 83, 84]. Participants also reported that with the on-call MARVIN chatbot, PWH can obtain needed information, resources, and support to self-manage their health anytime, anywhere, without having to wait for the next meeting with their doctor. However, some participants struggled with chatbot technology and social media due to gaps in digital literacy.

While a training session on MARVIN was provided by the technical coordinator to participants before study entry, the qualitative results stressed participant preferences for further instructional materials on the chatbot. We thus need to develop and disseminate relevant user guides or video tutorials for all prospective MARVIN users to aid their use and ensure equitable delivery of the intervention.

Convivial conversation was another factor explaining MARVIN’s ease of use, with several interviewees expressing how using MARVIN was like chatting with a friend. Consistent with our attempts to include empathic elements (e.g., smiley emoji, words of encouragement) when designing predefined messages [42], this contributes to the chatbot’s usability and potential to comfort users. Certainly, the quality of interaction could be improved. Several participants indicated that MARVIN sometimes required multiple user attempts to answer questions correctly on certain topics. Such limited comprehension disrupted the fluidity of use, which may in turn reduce their willingness to use it. Technically, the English language model is easier to train than the French one and therefore promises better results [85]. However, there was no significant difference (*P*=.07) in the primary outcomes between the two versions when comparing the language-based groupings. Further development will focus on collecting more training data and improving MARVIN’s comprehension in both languages. Considering the high proportion of immigrants people among newly diagnosed PWH in Canada and that language barriers are a key challenge to linkage and retention in care [86], new language versions (e.g., Spanish), as proposed by participants could also be developed. Large Language Models (LLMs) will also be integrated into MARVIN in the future. Represented by ChatGPT launched in late 2022, they offer impressive comprehension capabilities compared to traditional technologies, while also generating fluent responses [87]. Their multilingual capabilities could also help address potential language barriers [88]. However, a narrative review of HIV care-related chatbots noted that current LLMs can provide biased and fabricated responses, challenging their practical application [33].

Emotional safety and confidentiality were highlighted by participants, during the qualitative interviews, as two other prominent factors contributing to satisfaction with MARVIN. Many PWH still face significant psychological challenges as they are often discriminated against, socially isolated, and stigmatized for their condition [31, 89]. A needs-assessment study identified a concern that non-human interactions with chatbots could exacerbate feelings of marginalization due to the technology’s lack of empathy [90]. However, our findings, along with those of other studies [11, 38, 40, 41, 83, 84], suggest that a nonjudgmental chatbot like MARVIN can provide PWH a reassuring connection. Its objectivity and user anonymity allows for open discussion on sensitive or taboo topics [91] such as sexual behavior and HIV transmission prevention without fear of being criticized. This encourages users to seek information freely and might enhances their willingness to use MARVIN. However, the choice of Facebook Messenger as the interface for deploying MARVIN was seen as a big hurdle by some participants. Some PWH were unable to participate because they did not have a Facebook account, and a further subset expressed concern about Facebook's poor privacy record. Consequently, we have begun work on a standalone web user interface. Additional options (e.g., third-party mobile applications) are also being considered and will be changed in the future to improve user trust and adoption.

Interviewees also appreciated MARVIN's ability to provide reliable and useful real-time information related to ART self-management, with reminders and medication management while traveling being particularly appreciated by some. We attribute the perceived trustworthiness of MARVIN's content to the co-design strategy implemented through patient and stakeholder engagement: ongoing communication with patient experts identified medication adherence as the primary goal of MARVIN, and healthcare professionals ensured information reliability, both of which contributed to usability. However, some participants perceived MARVIN as more suitable for newly diagnosed PWH. Subgroup comparisons, in the statistical analyses, showed significantly higher acceptability (*P*=.009) among patients diagnosed in the last 5 years than among those diagnosed earlier. A similar though insignificant difference was also found for usability (*P*=.07). MARVIN's relevance for more newly diagnosed PWH suggests that it can play a role in models for rapid ART initiation which is considered a key strategy for achieving rapid viral suppression [86]. Patients involved in this model of care are typically treatment-naïve and require HIV-related health information to answer their concerns [92]. For those who have a seasoned understanding of ART management, the breadth and depth of MARVIN's conversation topics must be further expanded. A previous finding underscored that treatment adherence is but one of the many facets of HIV self-management [93]. In fact, MARVIN has grown since this study's completion to include more than 50 new topics (e.g., lifestyle, socioeconomic issues, mental health). It also addresses healthcare practices to promote PWH self-management of overall health, with more than 20 in preparation as of April 2024. And, based on participant input, referrals to relevant information and other external sources will also be added to enhance MARVIN's usability. Once again, LLMs also have a very high potential to address content breadth. In a study using the ChatGPT test for ART counseling and advice, it answered all questions accurately and comprehensively [94]. However, given the current limitations of LLMs in terms of interpretability, privacy protection, data transparency and liability for use [95, 96], research is needed to investigate how to integrate them safely and responsibly into healthcare chatbot services.

In addition to the conversation capabilities, interview participants expressed their desire for more advanced chatbot features. For example, participants wanted MARVIN to have memories. Currently, MARVIN is repetitive and always collecting the same information, which degrades the user experience. Long-term memory of key information is important and could simplify chatbot use. But even today's most powerful LLMs still face the challenge of poor memory capacity [97]. The status quo of chatbots, not limited to those used in healthcare, is that the majority can only perform short-term ad hoc interactions and have no memory [98-100]. There is a need to optimize access to the long-term context of conversational threads [101]. Lack of proactivity was another issue mentioned by participants. Prolonged interactions with chatbots that lack proactivity will create a sense of predictability for users, who will know the subsequent interactions they will encounter, thus reducing motivation to use them after the novelty wears off [102, 103]. Therefore, MARVIN must acquire the ability to initiate new conversations to increase its usability and ensure the long-term retention of its users. Participants suggested that question triaging, as seen with the Vik chatbot [12] and Woebot [15], may enhance chatbot proactivity. Yet, doing so could lead users to expect even more proactive bot interactions, and failure to fulfill this anticipation can negatively impact perceived chatbot usability. In sum, future development needs to enable MARVIN to utilize memory data and engage more proactively in deeper conversations.

The potential target audience for MARVIN may be broader than anticipated, given our qualitative findings. They suggest the chatbot's accurate and useful information as well as ease of use motivate users to share it with their peers, thus promoting better dissemination of HIV-related knowledge. Meanwhile, interviewees highlighted that HIV-related knowledge is not only relevant to PWH but is essential for every sexually active person. Several chatbots have indeed been developed to focus on pre-exposure prophylaxis information and help facilitate HIV self-testing [11, 33, 36, 38-41, 104]. Such efforts highlight the promising future of chatbots in different areas of HIV care. It is important to note, however, that the main role of MARVIN will continue to be assisting PWH rather than replacing healthcare professionals. While many participants reported trusting

MARVIN, some said they would only trust healthcare providers on certain topics. Especially when it comes to diagnosis and treatment options, or topics where the use of chatbots is restricted, users need to be directed to professional human resources.

Lastly, regarding our secondary endpoints, all validated positive associations attested to the appropriateness of the TAM to elucidate usability. ATU explained a significant portion of the variance in BIU (H5), as did PU for ATU (H3). This can be explained by our qualitative analyses: with emotional safety identified as a facilitator of ATU, participants would have the intention to converse with MARVIN. The results for H5 and H3 may also explain the only moderate and weak predictive power of PU for BIU (H4) and PEU for ATU (H2). In the case of H1, PEU has a weak explanatory power of for PU. This may be due to the fact that other external variables that are not part of the chatbot itself (i.e., alternative information resources, external referrals) have a greater influence on PU as antecedents of TAM (as shown in Figure 5.2). Future research could focus on investigating these parameters to better predict perceived usefulness and refine the application of the TAM model to chatbot evaluation.

5.5.2 Limitations

We acknowledge several limitations of this study. First, while the overall socio-demographic characteristics of the usability study participants including age, race, and education level were relatively diverse, it is important to highlight the gender imbalance of the participants. Digital divides related to limited technology access may alienate certain groups, such as women [105, 106]. Although, 40% of clinic patients are women, only 14.3% of the participants in this study were women. The outcomes for this user group need further study to determine if there is a gender gap in the use of MARVIN. Moreover, given the limited number of conversation topics, new subjects regarding women's health and pregnancy will be added. Female users will also be invited to participate more in the implementation process in the future to improve MARVIN’s adoption in this important population.

Secondly, partly due to convenience sampling and the small sample size, participants overall had relatively high levels of digital experience or interest in information technology. This may have introduced a sampling bias to our findings. The single-group, short-term and single-site study design also limits the generalizability of our findings. To gain a deeper understanding of chatbot implementation, we have developed a master protocol for future research, based on the design and results of this study [42]. Clinical validation will be conducted with a larger sample of users to further investigate chatbot performance in large-scale implementations. Randomized controlled trials with chatbot interventions need to evaluate other treatment modalities in parallel [107, 108].

5.6 Conclusions

Our MARVIN chatbot was validated for its usability in promoting PWH self-management, and the mixed methods design allowed us to gain a detailed understanding of the facilitators and barriers to usability. Our findings further demonstrate the promise of chatbots for HIV care and provide direction for MARVIN's further development. Next steps will focus on integrating LLMs to improve MARVIN's comprehension and expand its content, as well as to enhance the chatbot's functional intelligence, including memory and proactivity, to better respond to the needs of PWH. Given the limitations of current LLMs, their integration with MARVIN must advance prudently and responsibly. Ultimately, we hope MARVIN will become a personalized health companion for PWH.

5.7 Acknowledgements

This study is supported by the Canadian Institutes of Health Research Strategy for Patient-Oriented Research Québec Support Unit-Methodological Developments (grant M006, PI: BLeb), the PIHVOT program from ViiV Healthcare (grant 2020-1976-PIHVOT, PI: BLeb).

YM is supported by the Postgraduate Scholarship – Doctoral program (PGS D) and the OPSIDIAN program from the Natural Sciences and Engineering Research Council (NSERC), and a doctoral research award from the Fonds de recherche Nature et Technologies (FRQNT) in partnership with the Unité de soutien au système de santé apprenant (SSA) Québec.

BLeb is supported by 2 career awards: a Senior Salary Award from Fonds de recherche du Québec–Santé (FRQS) (#311200) and the Lettre d'Entente 250 (from the Québec Ministry of Health for researchers in Family Medicine), from the Quebec's Ministry of Health for researchers in Family Medicine, and holds a Canadian Institutes for Health Research, Strategy for Patient-Oriented Research Mentorship Chair in Innovative Clinical Trials for HIV Care.

ADP is also supported by a Senior Salary Award from Fonds de recherche du Québec–Santé (FRQS) and the Lettre d'Entente 250.

The funding sources had no role in the design of this study and will not be involved in the interpretation of the results or the decision to submit them for publication.

The authors acknowledge the technical support provided by students and interns from Polytechnique Montreal for the development of the MARVIN chatbots, and Lévis Thériault for referring the students. The authors also acknowledge Imane Frih and Maria Nait El Haj for the creation of the clinical content.

The authors thank the Chronic Viral Illness Service research team.

The names of the members of the MARVIN chatbots Patient Expert Committee cannot be provided as they are patients, and their names must remain confidential.

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

5.8 Data availability

The data sets generated and analyzed for this study are available from the corresponding author (BLeb) on reasonable request.

5.9 Authors' contributions

In no order of contribution, YM, SV, KE, and BLeb helped conceptualize the study and data collection tools. YM and SA on the software side, BLem, ML, BLeb, and the MARVIN chatbots Patient Expert Committee on the clinical side, collaborated to develop the MARVIN chatbot. ADP, JC, and BLeb referred patients. YM, SV completed the statistical analysis. YM, GT, DL and KE completed the qualitative analysis. YM and GT drafted the original manuscript. All authors critically reviewed the manuscript and approved the final version.

5.10 Conflicts of interest

BLeb has received research support, consulting fees and speaker fees from ViiV Healthcare, Merck, and Gilead.

The authors are identical to the developers of the intervention. And the ethical evaluators are identical to the sponsors of the study.

5.11 References

[1] Organization WH. The global health observatory, HIV, https://www.who.int/data/gho/data/themes/hiv-aids#:~:text=Globally%2C%2039.0%20million%20%5B33.1%E2%80%93,at%20the%20end%20of%202022. (2023, accessed 13/03/2024).

[2] Canada PHAo. HIV in Canada: 2022 surveillance highlights, https://www.canada.ca/en/public-health/services/publications/diseases-conditions/hiv-2022-surveillance-highlights.html (2023).

[3] Wandeler G, Johnson LF and Egger M. Trends in life expectancy of HIV-positive adults on antiretroviral therapy across the globe: comparisons with general population. Curr Opin HIV AIDS 2016; 11: 492-500. 2016/06/03. DOI: 10.1097/COH.0000000000000298.

[4] Colvin CJ. HIV/AIDS, chronic diseases and globalisation. Globalization and health 2011; 7: 1-6.

[5] Van de Velde D, De Zutter F, Satink T, et al. Delineating the concept of self-management in chronic conditions: a concept analysis. BMJ open 2019; 9: e027775.

[6] Saag MS. HIV Infection—Screening, Diagnosis, and Treatment. New England Journal of Medicine 2021; 384: 2131-2143.

[7] Engler K, Lènàrt A, Lessard D, et al. Barriers to antiretroviral therapy adherence in developed countries: a qualitative synthesis to develop a conceptual framework for a new patient-reported outcome measure. AIDS Care 2018; 30: 17-28. DOI: 10.1080/09540121.2018.1469725.

[8] Areri HA, Marshall A and Harvey G. Interventions to improve self-management of adults living with HIV on Antiretroviral Therapy: A systematic review. PLOS ONE 2020; 15: e0232709. DOI: 10.1371/journal.pone.0232709.

[9] King E, Kinvig K, Steif J, et al. Mobile Text Messaging to Improve Medication Adherence and Viral Load in a Vulnerable Canadian Population Living With Human Immunodeficiency

Virus: A Repeated Measures Study. Journal of Medical Internet Research 2017; 19: e190. DOI: 10.2196/jmir.6631.

[10] Xing Z, Yu F, Qanir YAM, et al. Intelligent Conversational Agents in Patient Self-Management: A Systematic Survey Using Multi Data Sources. Stud Health Technol Inform 2019; 264: 1813-1814. 2019/08/24. DOI: 10.3233/SHTI190661.

[11] Ntinga X, Musiello F, Keter AK, et al. The Feasibility and Acceptability of an mHealth Conversational Agent Designed to Support HIV Self-testing in South Africa: Cross-sectional Study. Journal of Medical Internet Research 2022; 24: e39816. DOI: 10.2196/39816.

[12] Chaix B, Bibault JE, Romain R, et al. Assessing the performances of a chatbot to collect real-life data of patients suffering from primary headache disorders. Digit Health 2022; 8: 20552076221097783. 2022/05/10. DOI: 10.1177/20552076221097783.

[13] Darcy A, Daniels J, Salinger D, et al. Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. JMIR Form Res 2021; 5: e27868. 2021/05/12. DOI: 10.2196/27868.

[14] Hauser-Ulrich S, Kunzli H, Meier-Peterhans D, et al. A Smartphone-Based Health Care Chatbot to Promote Self-Management of Chronic Pain (SELMA): Pilot Randomized Controlled Trial. JMIR Mhealth Uhealth 2020; 8: e15806. 2020/04/04. DOI: 10.2196/15806.

[15] Prochaska JJ, Vogel EA, Chieng A, et al. A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. J Med Internet Res 2021; 23: e24850. 2021/03/24. DOI: 10.2196/24850.

[16] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM 1966; 9: 36-45. DOI: 10.1145/365153.365168.

[17] Darcy A, Beaudette A, Chiauzzi E, et al. Anatomy of a Woebot(R) (WB001): agent guided CBT for women with postpartum depression. Expert Rev Med Devices 2022; 19: 287-301. 2022/06/25. DOI: 10.1080/17434440.2022.2075726.

[18] Klos MC, Escoredo M, Joerin A, et al. Artificial Intelligence-Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial. JMIR Form Res 2021; 5: e20678. 2021/06/08. DOI: 10.2196/20678.

[19] Jang S, Kim JJ, Kim SJ, et al. 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 2021; 150: 104440. 2021/04/03. DOI: 10.1016/j.ijmedinf.2021.104440.

[20] Zhou X, Edirippulige S, Bai X, et al. Are online mental health interventions for youth effective? A systematic review. Journal of Telemedicine and Telecare 2021; 27: 638-666. DOI: 10.1177/1357633x211047285.

[21] Li S, Wang Y, Chen L, et al. Virtual agents among participants with methamphetamine use disorders: Acceptability and usability study. Journal of Telemedicine and Telecare; 0: 1357633X231219039. DOI: 10.1177/1357633x231219039.

[22] Chen Y, Sinha B, Ye F, et al. Prostate cancer management with lifestyle intervention: From knowledge graph to Chatbot. Clinical and Translational Discovery 2022; 2. DOI: 10.1002/ctd2.29.

[23] Xu L, Sanders L, Li K, et al. Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. JMIR Cancer 2021; 7: e27850. 2021/12/01. DOI: 10.2196/27850.

[24] Bibault J-E, Chaix B, Guillemassé A, et al. A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial. Journal of medical Internet research 2019; 21: e15787.

[25] Chaix B, Bibault J-E, Pienkowski A, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR cancer 2019; 5: e12856.

[26] Nazareth S, Hayward L, Simmons E, et al. Hereditary Cancer Risk Using a Genetic Chatbot Before Routine Care Visits. Obstet Gynecol 2021; 138: 860-870. 2021/11/05. DOI: 10.1097/AOG.0000000000004596.

[27] Rehman UU, Chang DJ, Jung Y, et al. Medical Instructed Real-Time Assistant for Patient with Glaucoma and Diabetic Conditions. Applied Sciences 2020; 10. DOI: 10.3390/app10072216.

[28] Mash R, Schouw D and Fischer AE. Evaluating the Implementation of the GREAT4Diabetes WhatsApp Chatbot to Educate People With Type 2 Diabetes During the COVID-19 Pandemic: Convergent Mixed Methods Study. JMIR Diabetes 2022; 7: e37882. 2022/05/11. DOI: 10.2196/37882.

[29] Croes EA and Antheunis ML. 36 questions to loving a chatbot: are people willing to self-disclose to a chatbot? In: Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4 2021, pp.81-95. Springer.

[30] Ischen C, Araujo T, Voorveld H, et al. Privacy concerns in chatbot interactions. In: Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3 2020, pp.34-48. Springer.

[31] Arora AK, Ortiz-Paredes D, Engler K, et al. Barriers and Facilitators Affecting the HIV Care Cascade for Migrant People Living with HIV in Organization for Economic Co-Operation and Development Countries: A Systematic Mixed Studies Review. AIDS Patient Care STDS 2021; 35: 288-307. 2021/08/11. DOI: 10.1089/apc.2021.0079.

[32] Marcus JL, Sewell WC, Balzer LB, et al. Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic. Current HIV/AIDS Reports 2020; 17: 171-179.

[33] van Heerden A, Bosman S, Swendeman D, et al. Chatbots for HIV Prevention and Care: a Narrative Review. Current HIV/AIDS Reports 2023 20:6 2023-11-27; 20. DOI: 10.1007/s11904-023-00681-x.

[34] Brixey J, Hoegen R, Lan W, et al. Shihbot: A facebook chatbot for sexual health information on hiv/aids. In: Proceedings of the 18th annual SIGdial meeting on discourse and dialogue 2017, pp.370-373.

[35] Ardiana D, Joni I and Udayana I. Mobile based chatbot application for HIV/AIDS counseling using artificial intelligence markup language approach. In: Journal of Physics: Conference Series 2020, p.012041. IOP Publishing.

[36] van Heerden A, Ntinga X and Vilakazi K. The potential of conversational agents to provide a rapid HIV counseling and testing services. In: 2017 international conference on the frontiers and advances in data science (FADS) 2017, pp.80-85. IEEE.

[37] Yam EA, Namukonda E, Mcclair T, et al. Developing and Testing a Chatbot to Integrate HIV Education Into Family Planning Clinic Waiting Areas in Lusaka, Zambia. Global Health: Science and Practice 2022; 10: e2100721. DOI: 10.9745/ghsp-d-21-00721.

[38] Braddock WRT, Ocasio MA, Comulada WS, et al. Increasing Participation in a TelePrEP Program for Sexual and Gender Minority Adolescents and Young Adults in Louisiana: Protocol for an SMS Text Messaging–Based Chatbot. JMIR Research Protocols 2023; 12: e42983. DOI: 10.2196/42983.

[39] 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 2023; 12: e48447. 2023/06/30. DOI: 10.2196/48447.

[40] Hui M. Testing the Feasibility and Acceptability of Using an Artificial Intelligence Chatbot to Promote HIV Testing and Pre-Exposure Prophylaxis in Malaysia: Mixed Methods Study. JMIR Hum Factors 2024;11:e52055 https://humanfactorsjmirorg/2024/1/e52055 2024-01-26; 11. DOI: 10.2196/52055.

[41] 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. Journal of Medical Internet Research 2023; 25: e41881. DOI: 10.2196/41881.

[42] Ma Y, Achiche S, Pomey M-P, et al. Adapting and Evaluating an AI-Based Chatbot Through Patient and Stakeholder Engagement to Provide Information for Different Health Conditions: Master Protocol for an Adaptive Platform Trial (the MARVIN Chatbots Study). JMIR Research Protocols 2024; 13. DOI: 10.2196/54668.

[43] Laymouna M, Ma Y, Lessard D, et al. Roles, Users, Benefits and Limitations of Chatbots in Healthcare: A Rapid Review. Journal of Medical Internet Research 2024. DOI: 10.2196/56930.

[44] Garett R and Young SD. Potential application of conversational agents in HIV testing uptake among high-risk populations. Journal of Public Health 2023/03/14; 45. DOI: 10.1093/pubmed/fdac020.

[45] Bocklisch T, Faulkner J, Pawlowski N, et al. Rasa: Open source language understanding and dialogue management. arXiv preprint arXiv:171205181 2017.

[46] Creswell JW and Clark VLP. Designing and conducting mixed methods research. Thousand Oaks, CA, US: Sage Publications, Inc, 2007, p.xviii, 275-xviii, 275.

[47] Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020; 26: 1364-1374. 2020/09/11. DOI: 10.1038/s41591-020-1034-x.

[48] Browne RH. On the use of a pilot sample for sample size determination. Statistics in medicine 1995; 14: 1933-1940.

[49] Lancaster GA, Dodd S and Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004; 10: 307-312. 2004/06/11. DOI: 10.1111/j..2002.384.doc.x.

[50] Etikan I. Comparison of Convenience Sampling and Purposive Sampling. American Journal of Theoretical and Applied Statistics 2016; 5. DOI: 10.11648/j.ajtas.20160501.11.

[51] Meta. Meta privacy policy, https://www.facebook.com/privacy/policy/.

[52] Meta. Meta data security terms, https://www.facebook.com/legal/terms/data_security_terms.

[53] Meta. Privacy & safety on Messenger, https://www.facebook.com/help/messenger-app/1064701417063145/?helpref=hc_fnav.

[54] Standardization IOf. Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. 2018.

[55] Watbled L, Marcilly R, Guerlinger S, et al. Combining usability evaluations to highlight the chain that leads from usability flaws to usage problems and then negative outcomes. J Biomed Inform 2018; 78: 12-23. 2018/01/07. DOI: 10.1016/j.jbi.2017.12.014.

[56] Hagglund M and Scandurra I. Usability of the Swedish Accessible Electronic Health Record: Qualitative Survey Study. JMIR Hum Factors 2022; 9: e37192. 2022/06/24. DOI: 10.2196/37192.

[57] Patel B and Thind A. Usability of Mobile Health Apps for Postoperative Care: Systematic Review. JMIR Perioper Med 2020; 3: e19099. 2021/01/05. DOI: 10.2196/19099.

[58] Holden RJ and Karsh B-T. The Technology Acceptance Model: Its past and its future in health care. Journal of Biomedical Informatics 2010; 43: 159-172. DOI: 10.1016/j.jbi.2009.07.002.

[59] Venkatesh V and Davis FD. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science 2000; 46: 186-204. DOI: 10.1287/mnsc.46.2.186.11926.

[60] Revythi A and Tselios N. Extension of technology acceptance model by using system usability scale to assess behavioral intention to use e-learning. Education and Information Technologies 2019; 24: 2341-2355. DOI: 10.1007/s10639-019-09869-4.

[61] Standardization IOf. ISO/TS 20282-2:2013(en) Usability of consumer products and products for public use — Part 2: Summative test method.

[62] Alharbi S and Drew S. Using the technology acceptance model in understanding academics' behavioural intention to use learning management systems. 2014.

[63] Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011; 38: 65-76. 2010/10/20. DOI: 10.1007/s10488-010-0319-7.

[64] Dhagarra D, Goswami M and Kumar G. Impact of Trust and Privacy Concerns on Technology Acceptance in Healthcare: An Indian Perspective. International Journal of Medical Informatics 2020; 141: 104164. DOI: 10.1016/j.ijmedinf.2020.104164.

[65] Kalayou MH, Endehabtu BF and Tilahun B. &lt;p>The Applicability of the Modified Technology Acceptance Model (TAM) on the Sustainable Adoption of eHealth Systems in Resource-Limited Settings&lt;/p>. Journal of Multidisciplinary Healthcare 2020; Volume 13: 1827-1837. DOI: 10.2147/jmdh.s284973.

[66] Kamal SA, Shafiq M and Kakria P. Investigating acceptance of telemedicine services through an extended technology acceptance model (TAM). Technology in Society 2020; 60. DOI: 10.1016/j.techsoc.2019.101212.

[67] Lewis JR, Utesch BS and Maher DE. UMUX-LITE: when there's no time for the SUS. In: Proceedings of the SIGCHI conference on human factors in computing systems 2013, pp.2099-2102.

[68] Brooke J. SUS: a retrospective. Journal of usability studies 2013; 8: 29-40.

[69] Borsci S, Buckle P and Walne S. Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? Appl Ergon 2020; 84: 103007. 2019/12/01. DOI: 10.1016/j.apergo.2019.103007.

[70] Tariman JD, Berry DL, Halpenny B, et al. Validation and testing of the Acceptability E-scale for web-based patient-reported outcomes in cancer care. Applied Nursing Research 2011; 24: 53-58.

[71] Sauro J and Dumas JS. Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI conference on human factors in computing systems 2009, pp.1599-1608.

[72] Chau PY and Hu PJ-H. Investigating healthcare professionals' decisions to accept telemedicine technology: an empirical test of competing theories. Information & management 2002; 39: 297-311.

[73] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly 1989: 319-340.

[74] Adams C, Walpola R, Schembri AM, et al. The ultimate question? Evaluating the use of Net Promoter Score in healthcare: A systematic review. Health Expect 2022; 25: 2328-2339. 2022/08/20. DOI: 10.1111/hex.13577.

[75] Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International journal of market research 2008; 50: 61-104.

[76] Foundation PS. Python (Programming Language), https://www.python.org/ (accessed 14 May, 2024).

[77] Guest G, Namey E and McKenna K. How many focus groups are enough? Building an evidence base for nonprobability sample sizes. Field methods 2017; 29: 3-22.

[78] Kirk MA, Kelley C, Yankey N, et al. A systematic review of the use of the Consolidated Framework for Implementation Research. Implement Sci 2016; 11: 72. 2016/05/18. DOI: 10.1186/s13012-016-0437-z.

[79] CATIE. CATIE, Canada's source for HIV and hepatitis C information, https://www.catie.ca/ (accessed 14 May, 2024).

[80] Leo AJ, Schuelke MJ, Hunt DM, et al. A Digital Mental Health Intervention in an Orthopedic Setting for Patients With Symptoms of Depression and/or Anxiety: Feasibility Prospective Cohort Study. JMIR Formative Research 2022; 6: e34889. DOI: 10.2196/34889.

[81] Ehrlich C, Hennelly SE, Wilde N, et al. Evaluation of an Artificial Intelligence Enhanced Application for Student Wellbeing: Pilot Randomised Trial of the Mind Tutor. International Journal of Applied Positive Psychology 2023. DOI: 10.1007/s41042-023-00133-2.

[82] Amagai S, Pila S, Kaat AJ, et al. Challenges in Participant Engagement and Retention Using Mobile Health Apps: Literature Review. J Med Internet Res 2022; 24: e35120. 2022/04/27. DOI: 10.2196/35120.

[83] Bragazzi NL, Crapanzano A, Converti M, et al. The Impact of Generative Conversational Artificial Intelligence on the Lesbian, Gay, Bisexual, Transgender, and Queer Community: Scoping Review. Journal of Medical Internet Research 2023; 25: e52091. DOI: 10.2196/52091.

[84] Sanabria G, Greene KY, Tran JT, et al. “A Great Way to Start the Conversation”: Evidence for the Use of an Adolescent Mental Health Chatbot Navigator for Youth at Risk of HIV and Other STIs. Journal of Technology in Behavioral Science 2023 8:4 2023-05-11; 8. DOI: 10.1007/s41347-023-00315-4.

[85] Mielke SJ, Cotterell R, Gorman K, et al. What kind of language is hard to language-model? arXiv preprint arXiv:190604726 2019.

[86] Arora AK, Vicente S, Engler K, et al. Impact of social determinants of health on time to antiretroviral therapy initiation and HIV viral undetectability for migrants enrolled in a

multidisciplinary HIV clinic with rapid, free, and onsite B/F/TAF: 'The ASAP study'. HIV Med 2024 2024/01/12. DOI: 10.1111/hiv.13608.

[87] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med 2023 2023/07/18. DOI: 10.1038/s41591-023-02448-8.

[88] Yang R, Tan TF, Lu W, et al. Large language models in health care: Development, applications, and challenges. Health Care Science 2023; 2: 255-263. DOI: 10.1002/hcs2.61.

[89] de Los Rios P, Okoli C, Castellanos E, et al. Physical, Emotional, and Psychosocial Challenges Associated with Daily Dosing of HIV Medications and Their Impact on Indicators of Quality of Life: Findings from the Positive Perspectives Study. AIDS Behav 2021; 25: 961-972. 2020/10/08. DOI: 10.1007/s10461-020-03055-1.

[90] Comulada WS, Rezai R, Sumstine S, et al. A necessary conversation to develop chatbots for HIV studies: qualitative findings from research staff, community advisory board members, and study participants. AIDS Care 2024; 36: 463-471. DOI: 10.1080/09540121.2023.2216926.

[91] Belen-Saglam R, Nurse JRC and Hodges D. An Investigation Into the Sensitivity of Personal Information and Implications for Disclosure: A UK Perspective. Frontiers in Computer Science 2022; 4. DOI: 10.3389/fcomp.2022.908245.

[92] Arora AK, Engler K, Lessard D, et al. Experiences of Migrant People Living with HIV in a Multidisciplinary HIV Care Setting with Rapid B/F/TAF Initiation and Cost-Covered Treatment: The 'ASAP' Study. J Pers Med 2022; 12 2022/09/24. DOI: 10.3390/jpm12091497.

[93] Iribarren S, Siegel K, Hirshfield S, et al. Self-Management Strategies for Coping with Adverse Symptoms in Persons Living with HIV with HIV Associated Non-AIDS Conditions. AIDS Behav 2018; 22: 297-307. 2017/05/11. DOI: 10.1007/s10461-017-1786-6.

[94] Koh MCY, Ngiam JN, Yong J, et al. The role of an artificial intelligence model in antiretroviral therapy counselling and advice for people living with HIV. HIV Medicine 2024. DOI: 10.1111/hiv.13604.

[95] Wang C, Liu S, Yang H, et al. Ethical Considerations of Using ChatGPT in Health Care. J Med Internet Res 2023; 25: e48009. 2023/08/11. DOI: 10.2196/48009.

[96] Lee P, Goldberg C and Kohane I. The AI revolution in medicine: GPT-4 and beyond. Pearson, 2023.

[97] Zhong W, Guo L, Gao Q, et al. Memorybank: Enhancing large language models with long-term memory. In: Proceedings of the AAAI Conference on Artificial Intelligence 2024, pp.19724-19731.

[98] Nißen M, Selimi D, Janssen A, et al. See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. Computers in Human Behavior 2022; 127. DOI: 10.1016/j.chb.2021.107043.

[99] Janssen A, Passlick J, Rodríguez Cardona D, et al. Virtual Assistance in Any Context. Business & Information Systems Engineering 2020; 62: 211-225. DOI: 10.1007/s12599-020-00644-1.

[100] Griffin AC, Xing Z, Khairat S, et al. Conversational Agents for Chronic Disease Self-Management: A Systematic Review. In: AMIA Annual Symposium Proceedings 2020, p.504. American Medical Informatics Association.

[101] Følstad A and Brandtzæg PB. Chatbots and the new world of HCI. interactions 2017; 24: 38-42.

[102] Baraka K, Alves-Oliveira P and Ribeiro T. An extended framework for characterizing social robots. In: Human-robot interaction 2020, pp.21-64. Springer.

[103] Leite I, Martinho C and Paiva A. Social Robots for Long-Term Interaction: A Survey. International Journal of Social Robotics 2013; 5: 291-308. DOI: 10.1007/s12369-013-0178-y.

[104] 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 2022; 6: e42055. 2022/10/07. DOI: 10.2196/42055.

[105] Senteio C and Murdock PJ. The Efficacy of Health Information Technology in Supporting Health Equity for Black and Hispanic Patients With Chronic Diseases: Systematic Review. J Med Internet Res 2022; 24: e22124. 2022/04/05. DOI: 10.2196/22124.

[106] Pérez-Stable EJ, Jean-Francois B and Aklin CF. Leveraging advances in technology to promote health equity. Medical care 2019; 57: S101-S103.

[107] Gaffney H, Mansell W and Tai S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. JMIR Mental Health 2019; 6: e14166. DOI: 10.2196/14166.

[108] Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association 2018; 25: 1248-1258. DOI: 10.1093/jamia/ocy072.

CHAPTER 6 ARTICLE 3: LARGE LANGUAGE MODEL-BASED TRIAGE TO IDENTIFY ANTIRETROVIRAL THERAPY ADHERENCE BARRIERS AND RISK LEVELS IN PATIENT MESSAGES

Yuanchao Ma[1,2,3,4], Sofiane Achiche[1], David Lessard[2,3,4], Kim Engler[2,3], Serge Vicente[5], Gavin Tu[6], Benoît Lemire[4,7], Lina Del Balso[4], Nathalie Paisible[4], MARVIN Chatbots Patient Expert Committee, Bertrand Lebouché[2,3,4,8]

1. Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
2. Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre, Montreal, QC, Canada
3. Infectious Diseases and Immunity in Global Health Program, Research Institute of McGill University Health Centre, Montreal, QC, Canada
4. Chronic Viral Illness Service, Division of Infectious Disease, Department of Medicine, McGill University Health Centre, Montreal, QC, Canada
5. Département des enseignements généraux, École de technologie supérieure, Université du Québec, Montreal, QC, Canada
6. Faculty of Medicine, Université Laval, Quebec, Quebec, Canada
7. Department of Pharmacy, McGill University Health Centre, Montreal, Quebec, Canada
8. Department of Family Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada

Submitteed to *npj Digital Medicine* on 31st July 2025

6.1 Abstract

Optimal adherence to antiretroviral therapy (ART) is essential to HIV care but challenging. We developed large language models (LLMs) to identify adherence barriers (Thoughts & feelings, Habits & Activities, Social Situation, Economic Situation, Medication, Care, Health, None) and stratify nonadherence risk levels (High, Medium, Low, None) from patient messages. Fine-tuned Flan-T5 models were optimal for barrier detection (Macro-F1=0.83) and risk stratification (Macro-F1=0.80). Fine-tuned general-domain LLMs significantly outperformed clinical foundation models (ΔMacro-F1= [-0.11, -0.02], *P*<.001) on the primary test split, and outperformed large-scale LLMs (ΔMacro-F1= [-0.38, -0.07], *P*<.001) on the external validation dataset. Flan-T5 demonstrated greater robustness to demographic descriptors, whereas GPT-4.1&Gemma3 frequently reclassified inputs as None (44% & 33%). GPT-4.1's inference consumed ~30× more energy than Flan-T5-xl (0.376vs0.013 kWh/1,000 inferences), with carbon emissions varying by model hosting location. LLM-based approaches show promise for real-time, scalable ART adherence monitoring. However, these findings highlight the need for careful model selection, fairness assessment, and consideration of environmental sustainability in clinical AI development.

6.2 Introduction

Antiretroviral therapy (ART) has transformed HIV into a manageable chronic condition, substantially increasing the life expectancy of people with HIV (PWH) [1, 2]. However, maintaining optimal ART adherence remains a persistent challenge. Evidence from a Canadian retrospective study found that nearly 20% of PWH had adherence rates below 85% [3]. Similarly, a U.S. study [4] and a multi-country survey [5] reported that 40% and 25% of individuals, respectively, fell below 80% adherence [6, 7]. Suboptimal adherence not only leads to poorer clinical outcomes [8] and viral resistance [9] but also increases the risk of loss to follow-up [10, 11] and onward transmission [12].

Adherence barriers are factors that interfere, whether temporarily or persistently, with a person's ability to take medication as prescribed. Early identification is essential to enable timely intervention and prevent or limit treatment interruptions. Multiple clinical guidelines [13, 14] emphasize the importance of regularly assessing adherence barriers to promote personalized, patient-centred care. Yet, identifying these factors in routine clinical care remains difficult. Experiences of adherence may be discussed during medical consultations, but these are often constrained by limited time, competing demands, or patient discomfort discussing sensitive topics, and even physician unease in initiating such conversations [15-18]. Patient-reported outcome measures (PROMs) provide a standardized alternative but are infrequently integrated into HIV care due to concerns about increased workload, patient motivation, and unfamiliarity with interpreting results [19, 20]. Retrospective approaches such as pharmacy refill data or electronic health records [21] can offer useful insights into ART adherence patterns. However, they are inherently reactive, typically identifying problems only after adherence has declined, without necessarily revealing the underlying reasons for missed doses or allowing for anticipation of emerging adherence challenges [22, 23]. Proactive strategies are therefore needed to detect early warning signs and intervene before significant lapses occur.

In recent years, mobile health (mHealth) tools–such as SMS/text messaging [24-26], patient portals [27-30], and chatbots [31-33]–have been explored to support ART adherence by enhancing patient engagement in care [26, 27, 30], improving communication [25, 33], and facilitating self-management [24, 31]. These tools are perceived by PWH as acceptable and easy to use [24, 28-32, 34], with positive impacts on adherence documented in several studies [24, 28-30, 34].

However, their effectiveness often declines over time [24, 34]. Research highlights the need for more comprehensive, personalized, and proactive content to sustain user engagement and address evolving needs [24, 31, 32, 34]. Interestingly, patients generate a large volume of messages during interactions with mHealth platforms, which contain rich yet mostly untapped information relevant to adherence monitoring. One promising strategy is to analyze these messages to triage emerging adherence problems and tailor interventions accordingly. The advent of large language models (LLMs) presents a viable opportunity to automatically detect ART adherence barriers from these unstructured data streams. Prior research has shown the potential of AI-based natural language processing (NLP) techniques to analyze patient messages for early detection of depression [35] and mental health crises [36, 37], both recognized as well-established barriers to ART adherence. Leveraging such approaches may enable earlier identification of patients at risk of nonadherence, facilitating timely and targeted interventions.

Despite growing interest in such applications, limited information exists on which models are best suited for processing unstructured patient messages. Model selection is critical in healthcare AI model development [38], where clinical foundation models – pretraining on biomedical or clinical records – are often assumed to outperform general-domain models following fine-tuning for downstream clinical NLP tasks [39-41]. However, recent studies have challenged their added value in real-world settings [42]. Despite strong performance on benchmark NLP tasks, these models may not generalize well to such informal, patient-facing tasks. This is particularly relevant for classification tasks involving patient-generated messages, which use everyday language that may not be well captured in clinical training data [35]. In parallel, larger-scale LLMs such as GPT-4 have demonstrated robust performance on standardized exams (e.g., with GPT-4 passing the United States Medical Licensing Examination (USMLE) [43]) and clinical knowledge in HIV care [44, 45], but this does not guarantee their ability to identify nuanced adherence barriers from patient-generated messages. These gaps highlight the need to evaluate model suitability for identifying adherence challenges in real-world HIV care contexts.

Equally important is ensuring the fair and sustainable development and deployment of LLM-based models in HIV care. Responsible AI development needs to align with social and environmental values and minimize potential harms [46]. For example, a key fairness concern is that LLMs may perpetuate racial and gender biases in healthcare [47], which is particularly troubling in contexts where intersecting stigma already disproportionately impacts marginalized communities [48, 49].

Without careful design and oversight, these technologies risk amplifying existing inequities and reinforcing the very stigma they should dismantle. Sustainability is another essential consideration [50], particularly as LLMs gain traction in healthcare and their environmental footprint raises growing concerns about the long-term viability of AI [51, 52]. Transparency must extend beyond publishing performance results to include energy consumption and carbon emissions to support more informed and targeted model development and deployment decisions. Assessing these key model features is critical to inform the development of LLM-based tools to support ART adherence and ensure their effective deployment in real-world healthcare settings.

6.3 Objectives

The primary objective of this study was to develop and validate LLM-based triage models to automatically and proactively identify ART adherence challenges from patient-generated messages. Specifically, we aimed to 1) assess the comparative performance of fine-tuned general-domain versus clinical foundation models, alongside the GPT family and other open-source LLMs on adherence barrier and associated nonadherence risk classification informed by adherence level thresholds; 2) Examine potential model biases related to racial and gender descriptors; and 3) assess the energy consumption and carbon footprints of model development and deployment.

6.4 Methods

6.4.1 Participatory design approach

Given the interdisciplinary scope and the patient-centered nature of the study, we adopted a participatory design approach [53]. A co-construction committee was established at the outset of the study, comprising three PWH, three HIV care specialists (one pharmacist, two research nurses), one medical resident, and one engineering researcher. The committee actively contributed to key stages of the project, including data annotation, model development, and validation. This collaborative approach ensured that the evolving tool reflected patient priorities and supported the development of a trustworthy healthcare AI solution [54, 55].

6.4.2 Ethical considerations

This study received approval from the McGill University Health Centre research ethics board on August 10, 2023 (approval: MP-37-2023-9333R).

6.4.3 Conceptual framework and definitions

ART adherence barriers are complex and multifactorial. To guide our classification, we drew on the conceptual framework and content of the 7-item Interference-Score (I-Score) PROM–a stakeholder-informed tool developed to facilitate the identification and discussion of ART adherence barriers in routine HIV care [56]. The I-Score captures the presence of barriers within seven domains and our barrier categories mirror these:

1) Thoughts & feelings: this includes barriers related to acceptance of having HIV, emotions (feeling sad, anxious, etc.), medication-related knowledge and beliefs or motivation to take medication.
2) Habits & Activities: this encompasses barriers tied to daily life (one's schedule, priorities, etc.) or substance use (alcohol, drugs, or other substances).
3) Social Situation: this involves barriers due to personal relationships or the experience of stigma.
4) Economic Situation: this includes financial or housing-related barriers.
5) Medication: this includes barriers associated with the HIV medication, namely its side effects, instructions or physical features (for example, pill size or taste).
6) Care: this refers to barriers due to issues with healthcare professionals, the clinic, the pharmacy, or payment of medication or care.
7) Health: this includes barriers attributed to lab test results, HIV symptoms or overall health.

To inform the risk classification scheme, we defined adherence levels as a function of the number of nonadherent days within a 30-day period, based on clinical literature [5, 6, 57] and committee input:

- Perfect adherence: No missed doses
- Optimal adherence: ≥95% and <100% (i.e., ≤1 missed dose)
- Near-optimal adherence: <95% and ≥80% (i.e., 2–5 missed doses, not occurring consecutively)
- Suboptimal adherence: <80% (i.e., ≥6 missed doses)

Overall, the four-level risk classification scheme takes account of the severity and frequency of adherence-related challenges in the message. See Table 6.1 for illustrative examples of nonadherence risk levels.

1) High: Explicit mention of an adherence barrier that is currently causing, or is highly likely to lead to, suboptimal adherence.
2) Medium: Explicit mention of adherence barrier that is currently causing, or is highly likely to lead to, near-optimal adherence.
3) Low: Explicit mention of an adherence barrier that is currently causing, or is highly likely to lead to, optimal but not perfect adherence.
4) None: No mention of adherence barriers, or explicit mention of a barrier with an explicit statement of no current or imminent adherence risk.

Table 6.1 Illustrative examples of nonadherence risk levels.

Example	Barrier	Risk level
And that might mean another 5 days of not taking my medications, cause I will not go back to the pharmacy until my next day off.	Care	High
So, you're going out and sometimes it's going to make you forget about your treatment.	Habits & Activities	Medium
I can't think of anything yet that could stop me, maybe if I was really sick.	Health	Low
I didn't see any difference, because between taking a small tablet and a bigger one at the same time, or taking just one, it doesn't make much difference.	Medication	None

6.4.4 Data

We compiled a primary de-identified dataset from multiple sources:

1) MARVIN Training corpus (MT): MARVIN is an AI-based chatbot developed since 2020 that provides self-management support to PWH through English and French text-based conversations [31]. We included its English training corpus, co-developed with patient partners, healthcare providers and developers. This corpus covers 113 topics, including:
   - ART guidance (e.g., timing, dosing, drug interactions, side effects).
   - ART management during travel.
   - Common HIV-related knowledge (e.g., symptoms, transmission, prevention, vaccination).
2) MARVIN User conversations (MU): Conversation records collected from 1,243 MARVIN users between 2021 and 2024 (1,075 in English; 168 in French).
3) I-Score study interviews (ISCORE): We included 27 interview transcripts (15 English, 12 French) from the I-Score development study conducted with PWH in Canada between 2016 and 2017. These interviews primarily explored participants' real-life experiences with ART adherence, including the challenges they faced or anticipated in maintaining treatment.

Two external validation datasets were also assembled:

1) Online Forum messages (OF): A collection of 359 English-language thread messages related to ART were randomly selected from the Reddit r/HIV [58], r/hivaids forums [59], and the POZ community forum [60], using the keywords "medication" and "ART treatment" from 2024 onwards.
2) Portail VIH/SIDA du Québec messages (PVSQ): A set of 320 anonymous, user-submitted messages (7 in English; 313 in French) collected between October 2023 and May 2025 through the Q&A messaging portal of the Portail VIH/SIDA du Québec [61]. This community organization provides information and support related to HIV and other sexually transmitted and blood-borne infections (STBBIs). The collected messages covered topics ranging from HIV and STBBIs prevention, treatment, transmission, and daily life with these conditions.

These sources were selected for their relevance to real-world HIV care, their digital and conversational nature, and the richness of language used to describe adherence-related experiences, making them well-suited for developing triage models applied to patient-generated messages. Given the multifactorial nature of ART adherence challenges, individual messages often contained descriptions of multiple and intertwined potential barriers. To ensure sentence-level annotation and reliable classification, all data were split into sentences using the *syntok* sentence segmenter Python library [62]. Preprocessing steps included translation of French content to English, removal of HTML tags and URLs, and de-duplication.

6.4.5 Data annotation

Two sentence-level, multi-class classification tasks were defined. Each sentence was annotated with one label per task, reflecting its most relevant barrier type and risk level:

- Task I: ART adherence barrier classification – Sentences were categorized into one of seven barrier types by referencing the I-Score PROM. An eighth category, *None*, was used if the sentence did not mention any adherence barriers.
- Task II: Triage risk level classification – Sentences were categorized into one of four nonadherence risk levels.

The data annotation process is illustrated in Figure 6.1. Through three 90-minute workshops, the co-construction committee iteratively developed and refined the annotation guideline. Between workshops, the engineering researcher and the medical resident independently annotated 10% of the dataset following the evolving guideline. Ambiguous cases were discussed during workshops, and the guideline was updated accordingly. After confirming reliability, one researcher completed the full dataset annotation using the finalized guideline (See Supplementary file 1).

6.4.6 Data augmentation

In the collected training data, sentences related to *Economic Situation* accounted for only 1.2% of the total (162/13780), while sentences related to *None* accounted for 38.6% (5319/13780). To address class imbalance in the training dataset, we applied synthetic data augmentation, a strategy shown to improve model performance under such conditions [63]. Using six LLMs, we generated candidate sentences for each barrier category and risk level. Following manual review for quality and topic-relevance, only validated sentences were included in the final training dataset. Details of the LLMs and prompting procedures are provided in Supplementary file 2.

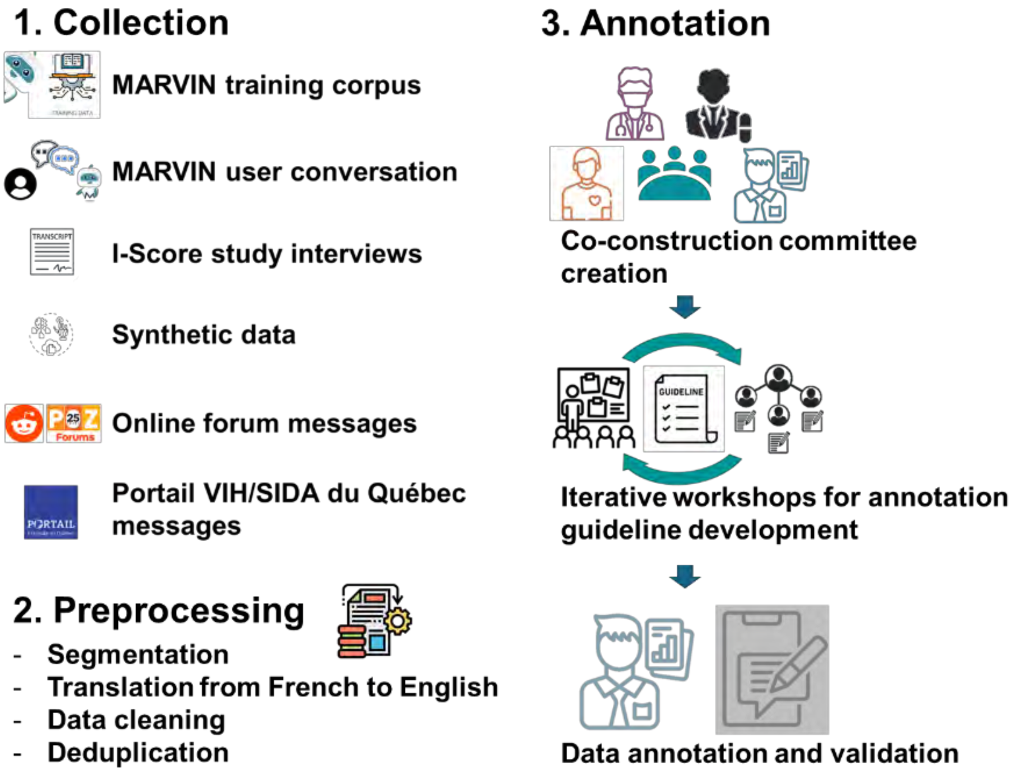


Figure 6.1 Data preparation and annotation process with the co-construction committee.

Table 6.2 presents the characteristics of the final curated datasets, including sentence distributions and sentence length statistics. The primary dataset (MT+MU+ISCORE) contained 13,780 annotated sentences; the synthetic dataset included 700 validated sentences; and the external validation dataset (OF+PVSQ) had 1000 annotated instances. Inter-annotator agreement was assessed using Cohen's Kappa coefficient, yielding scores of 0.83 for barrier annotation and 0.71 for risk level, indicating almost perfect and substantial agreement, respectively [64].

6.4.7 Model development

The primary dataset was randomly split into training (80%), validation (10%), and test (10%) sets. The OF+PVSQ datasets, along with the test split, were held out for external assessment throughout model development. The synthetic dataset was incorporated into the training set to address class imbalance.

Model development focused on the effectiveness of different LLMs for classifying ART adherence barriers and associated risk levels. We fine-tuned general-domain language models, including BERT-base [65] and Flan-T5(-small, -base, -large, and -xl) [66]. We also fine-tuned clinical foundation models from both families: Bio-BERT [39], Bio-ClinicalBERT [41], and Clinical-T5(-base and -large) [40].

All model architectures were adapted for multi-class sequence classification. Cross-entropy loss was used for all models. To reduce computational demands, Flan-T5-large and -xl models were fine-tuned using parameter-efficient tuning via low-rank adaptation (LoRA) [67]. Class weighting was applied during training to mitigate bias toward the majority class. Model hyperparameters are provided in Supplementary file 2.

6.4.8 Model assessment

All models were monitored on the validation split during training, with final performance assessed on the held-out test split and external validation datasets. Multi-class prediction performance was measured using the Macro-F1 score, which provides a balanced evaluation of precision (positive predictive value) and recall (sensitivity) across all classes.

The F1 score is defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, where\ Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}$$

The Macro-F1 score is computed as the unweighted average of the F1 scores for all N classes:

$$\mathrm{Macro - F1} = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

Where TP = true positives, FP = false positives, FN = false negatives.

For the best-performing models, a binary task reformulation was performed to assess the ability to discriminate between the presence and absence of barriers or risk. All detected barrier or risk levels were grouped together and compared against the *None* category, with F1 score used to quantify performance.

6.4.9 Error analysis

An inductive error analysis was conducted on misclassified examples from the primary test split dataset using the best-performing model to support model explainability and identify common error patterns. To ensure credibility and reliability, results were debriefed with the co-construction committee at the fourth workshop.

Table 6.2 Characteristics of the development and validation datasets.

Dataset overview	Barrier								Risk level			
	Thoughts & Feelings	**Habits & Activities**	**Social Situation**	**Economic Situation**	**Medication**	**Care**	**Health**	**None**	**High**	**Medium**	**Low**	**None**
Development												
MT (n=3,874)	302 (7.8%)	541 (14.0%)	91 (2.3%)	0 (0.0%)	611 (15.8%)	56 (1.4%)	823 (21.2%)	1450 (37.4%)	140 (3.6%)	687 (17.7%)	1068 (27.6%)	1979 (51.1%)
MU (n=3,721)	205 (5.5%)	802 (21.6%)	137 (3.7%)	5 (0.1%)	612 (16.4%)	29 (0.8%)	447 (12.0%)	1484 (39.9%)	183 (4.9%)	653 (17.5%)	1120 (30.1%)	1765 (47.4%)
ISCORE (n=6,185)	1047 (16.9%)	742 (12.0%)	444 (7.2%)	157 (2.5%)	522 (8.4%)	532 (8.6%)	356 (5.8%)	2385 (38.6%)	217 (3.5%)	1313 (21.2%)	819 (13.2%)	3836 (62.0%)
MT+MU+ISCORE (n=13,780)	1554 (11.3%)	2085 (15.1%)	672 (4.9%)	162 (1.2%)	1745 (12.7%)	617 (4.5%)	1626 (11.8%)	5319 (38.6%)	540 (3.9%)	2653 (19.3%)	3007 (21.8%)	7580 (55.0%)
Train (n=11,023)	1243 (11.3%)	1668 (15.1%)	538 (4.9%)	129 (1.2%)	1395 (12.7%)	494 (4.5%)	1301 (11.8%)	4255 (38.6%)	440 (4.0%)	2750 (24.9%)	1748 (15.9%)	6085 (55.2%)
Valid (n=1,377)	167 (12.1%)	215 (15.6%)	60 (4.4%)	19 (1.4%)	183 (13.3%)	54 (3.9%)	155 (11.3%)	524 (38.1%)	49 (3.6%)	400 (29.0%)	198 (14.4%)	730 (53.0%)
Test (n=1,380)	144 (10.4%)	202 (14.6%)	74 (5.4%)	14 (1.0%)	167 (12.1%)	69 (5.0%)	170 (12.3%)	540 (39.1%)	51 (3.7%)	351 (25.4%)	213 (15.4%)	765 (55.4%)
SD (n=700)	79 (11.3%)	77 (11.0%)	81 (11.6%)	259 (37.0%)	53 (7.6%)	91 (13.0%)	60 (8.6%)	0 (0.0%)	155 (22.1%)	267 (38.1%)	172 (24.6%)	106 (15.1%)
Validation												
PVSQ (n=641)	41 (6.4%)	16 (2.5%)	24 (3.7%)	11 (1.7%)	16 (2.5%)	81 (12.6%)	300 (46.8%)	152 (23.7%)	53 (8.3%)	137 (21.4%)	201 (31.4%)	250 (39.0%)
OF (RDT+POZ, n=359)	56 (15.6%)	52 (14.5%)	32 (8.9%)	43 (12.0%)	65 (18.1%)	42 (11.7%)	58 (16.2%)	11 (3.1%)	36 (10.0%)	205 (57.1%)	74 (20.6%)	44 (12.3%)

Note. All values are presented as n (%). The primary dataset (MT+MU+ISCORE) was randomly divided into training (80%), validation (10%), and test (10%) subsets. The synthetic dataset (SD) was incorporated into the training set, resulting in a final training sample size of n = 11,723. MT: MARVIN Training corpus; MU: MARVIN User conversation; ISCORE: I-Score study interviews; SD: Synthetic data; PVSQ: Portail VIH/SIDA Québec; OF: Online forum; RDT: Reddit forum; POZ: POZ community forum

6.4.10 Comparison with other LLMs

We assessed the performance of proprietary GPT-family LLMs (GPT-4.1-2025-04-14, -mini, -nano) and open-source models (LLaMA3.2-3B, DeepSeek-R1-7B, Qwen3-14B, Gemma3-12B, Mistral-7B) on the external validation datasets. Their results were compared to those of our best-performing fine-tuned models.

All models were prompted using both zero-shot and five-shot chain-of-thoughts (CoT) approaches. Test inputs were embedded within a standardized prompt template (See Supplementary file 2), instructing the model to perform multi-class classification and return predictions in structured JSON format. For the five-shot setting, examples were randomly sampled from the training split. To ensure reproducibility, the temperature parameter was fixed at zero for all models.

6.4.11 Assessment of model bias and environmental footprint

We evaluated potential model bias related to race and gender by calculating the prediction mismatch rate after introducing identity descriptors into originally labeled sentences. Following established approaches [68, 69], we focused on sentences originally categorized as containing barriers or risk levels (i.e., not labeled *None*), since false positives carry relatively low risk in triage classifiers [37, 70].

Using GPT-4.1-mini, combined gender and racial descriptors (e.g., *Asian woman*, *Black man*, *Latino trans person*) were randomly inserted into selected sentences from the de-identified external validation set (See Table 6.3 for examples). The following descriptors were applied:

- Gender: woman, man, trans
- Race: Asian, Black, Indigenous, Latino, White

Table 6.3 Examples of injected gender and racial descriptors used to assess model fairness.

Original sentence	**After injection of gender and racial descriptors**
How do I know if my treatment is working?	*As a black woman*, how do I know if my treatment is working?
Does current antiretroviral treatment affect appearance?	*As an indigenous man*, does current antiretroviral treatment affect appearance?

After manual validation to ensure semantic coherence, model inference was performed on 837 modified sentences, and we calculated the prediction mismatch. Full prompt details are available in Supplementary file 2.

To assess the environmental sustainability of our models, energy consumption and carbon footprint were estimated during both fine-tuning and inference using the *CodeCarbon* [71] Python package. Both metrics were tracked for each fine-tuned model to compare development-related environmental impact. For inference, we compared our best-performing model to the other LLMs based on published estimates. All values were normalized per 1,000 inferences, and the host country of the cloud computing infrastructure for fine-tuning and inference was reported to account for regional variation in carbon emissions.

6.4.12 Statistical analysis

Model performance was evaluated by estimating the mean Macro-F1 and corresponding 95% confidence intervals (CIs) using non-parametric bootstrap sampling with the percentile method on the test set and external validation datasets. The bootstrap sample size matched each dataset, with sampling performed with replacement. A total of 300 bootstrap iterations were conducted to ensure the standard error of the CI limits remained below 0.01. The same procedure was used to estimate CI for binary classification F1 scores.

Pairwise comparisons of Macro-F1 between general-domain language models and clinical foundation models of matched architecture and size were performed using the Mann-Whitney U test. The best-performing fine-tuned models were compared with other LLMs using the same test, and Bonferroni correction was applied for multiple.

Mismatch rates in barrier and risk level classification, before and after the injection of racial and gender descriptors, were compared using chi-square tests for each subgroup.

A two-sided significance level of 5% was applied for all statistical tests. Analyses were conducted using the Python *SciPy* [72] package.

6.5 Results

6.5.1 Model performance on primary test split

Table 6.4 summarizes fine-tuned model performance for both classification tasks using the primary dataset test split.

For barrier prediction (Task I), the best-performing model was BERT-base, achieving a Macro-F1 score of 0.85 (95% CI: 0.82-0.87) and the highest F1 scores for *Economic Situation* and *None*.

For risk level prediction (Task II), the best-performing models were BERT-base and Flan-T5-base, both reaching a Macro-F1 score of 0.80 (BERT-base: 95% CI: 0.76–0.82; Flan-T5-base: 95% CI: 0.77–0.83). Notably, Flan-T5-base achieved the highest F1 scores in three of the four risk categories (*Medium*, *Low*, and *None*).

Except for Flan-T5-small, all other general-domain models performed comparably after fine-tuning, with Macro-F1 differences within 0.02 of the best-performing models across both tasks.

For both tasks, model performance varied across categories. The *None* category showed the highest performance. In barrier prediction, the *Medication* and *Health* barrier categories also performed well, followed by *Care* and *Habits & Activities*. The remaining three categories (*Thoughts & Feelings, Social Situation, Economic Situation*) showed more variable performance across models. For risk level prediction, the *Medium* category performed second best after *None,* followed by *Low*, while *High* performed moderately across models.

6.5.2 Comparison of general-domain and clinical foundation models

Across both tasks, general-domain language models significantly outperformed clinical foundation models following fine-tuning (P<.001). The performance gap increased with model size, as reflected by increasing Δ Macro-F1 values (–0.01 to –0.11).

When comparing model performance across categories for each task, general and clinical T5-based models performed similarly in *Economic Situation, Medication, Health,* and *None* categories in Task I, while larger F1 score gaps were observed in the remaining four categories. For BERT-based models, larger gaps in F1 scores were observed in the *Thoughts & Feelings, Social Situation,* and *Economic Situation* categories compared to the other categories. For Task II, all four model pairs showed large F1 score differences in the *High* and *Low* categories, while in the *Medium* and *None* categories had more modest differences.

6.5.3 External validation performance

Table 6.5 presents fine-tuned model performance on the external validation datasets. Performance generally scaled with model size on external data.

In Task I, Flan-T5-xl achieved the best overall Macro-F1 score of 0.71, ranking first in six of the eight barrier categories. For Task II, Flan-T5-large achieved the highest Macro-F1 score of 0.57, with top F1 score performance in the *Medium* and *Low* categories. Based on consistent results across both internal and external datasets, Flan-T5-xl and Flan-T5-large were identified as the best-performing models for Task I and Task II, respectively.

On external datasets, model performance declined by approximately 15% for barrier classification and 30% for risk level prediction, with mean Macro-F1 scores decreasing from 0.83 to 0.71 and 0.80 to 0.57, respectively, compared to the primary test split.

Clinical-T5 models continued to underperform relative to Flan-T5 models (*P*<.001). In contrast, BioBERT slightly outperformed BERT-base in barrier detection, and both BioBERT and BioClinicalBERT achieved marginally better performance in risk level prediction (ΔMacro-F1=+0.01, *P*<.001).

6.5.4 Binary task reformulation analysis

Table 6.6 summarizes the binary task reformulation analysis on our best-performing fine-tuned models. On the primary test split, Flan-T5-xl achieved an overall F1 score of 0.94 (95% CI: 0.93-0.95) for Task I, with F1=0.95 for detecting barrier-related sentences. For Task II, Flan-T5-large achieved an overall F1 of 0.90 (95% CI: 0.89-0.92), with F1=0.89 for identifying sentences indicating adherence risk.

On the external validation datasets, performance declined across both tasks. For Task I, Flan-T5-xl achieved an F1 score of 0.85 (95% CI: 0.82–0.88), with strong detection of barrier-related sentences (F1 = 0.96) but reduced performance on *None* sentences (F1 = 0.74). For Task II, Flan-T5-large achieved a F1 of 0.76 (95% CI: 0.72–0.78), with F1=0.86 for elevated adherence risk and F1=0.65 for non-risk sentences.

Table 6.4 Fine-tuned model performance on test split – primary dataset.

Task I - Barrier prediction

Model	Parameters (Total/Tuned)	Mean Macro-F1 (95% CI)	ΔMacro-F1	*P* value	Thoughts & Feelings (F1)	Habits & Activities (F1)	Social Situation (F1)	Economic Situation (F1)	Medication (F1)	Care (F1)	Health (F1)	None (F1)
BERT-base	110M/110M	**0.85 (0.82-0.87)**			0.70	0.86	0.77	**0.83**	0.89	0.89	0.92	**0.92**
BioBERT	110M/110M	0.81 (0.78-0.84)	-0.04	<.001	0.65	0.86	0.70	0.76	0.88	0.86	0.91	0.91
BioClinicalBERT	110M/110M	0.83 (0.80-0.85)	-0.02	<.001	0.67	0.85	0.76	**0.83**	0.88	0.86	0.89	0.91
Flan-T5-small	77M/77M	0.76 (0.72-0.79)			0.57	0.81	0.66	0.71	0.83	0.78	0.87	0.87
Flan-T5-base	248M/248M	0.83 (0.80-0.86)			0.72	**0.87**	0.78	0.69	0.89	0.85	**0.94**	**0.92**
Clinical-T5-base	248M/248M	0.77 (0.73-0.80)	-0.06	<.001	0.61	0.80	0.67	0.67	0.86	0.77	0.91	0.86
Flan-T5-large	760M/4.7M	0.83 (0.80-0.86)			**0.73**	**0.87**	0.78	0.62	**0.90**	**0.92**	0.92	**0.92**
Clinical-T5-large	760M/4.7M	0.72 (0.69-0.76)	-0.11	<.001	0.57	0.75	0.60	0.56	0.81	0.79	0.87	0.86
Flan-T5-xl	3B/9.4M	0.83 (0.80-0.87)			0.72	**0.87**	**0.79**	0.71	0.88	0.87	0.91	**0.92**

Task II - Risk level prediction

Model	Parameters (Total/Tuned)	Mean Macro F1 (95% CI)	ΔMacro-F1	*P* value	High (F1)	Medium (F1)	Low (F1)	None (F1)
BERT-base	110M/110M	**0.80 (0.76-0.82)**			0.69	**0.81**	0.78	**0.91**
BioBERT	110M/110M	0.79 (0.76-0.82)	-0.01	<.001	0.73	0.80	0.75	0.89
BioClinicalBERT	110M/110M	0.77 (0.73-0.80)	-0.03	<.001	0.65	0.80	0.73	0.89
Flan-T5-small	77M/77M	0.64 (0.60-0.67)			0.48	0.69	0.56	0.83
Flan-T5-base	248M/248M	**0.80 (0.77-0.83)**			0.70	**0.81**	**0.80**	**0.91**
Clinical-T5-base	248M/248M	0.72 (0.68-0.75)	-0.08	<.001	0.56	0.75	0.67	0.88
Flan-T5-large	760M/4.7M	0.78 (0.75-0.82)			0.67	**0.81**	0.76	**0.91**
Clinical-T5-large	760M/4.7M	0.68 (0.64-0.71)	-0.10	<.001	0.59	0.71	0.57	0.85
Flan-T5-xl	3B/9.4M	0.79 (0.76-0.82)			**0.74**	0.79	0.74	0.90

Note. The 95% CIs for Macro-F1 scores were calculated using 300 bootstrap samples with replacement on three distinct datasets, ensuring a SE of the CI limits below 0.01. The observed SE interval limit was 0.0051. ΔMacro-F1 represents the Macro-F1 difference between fine-tuned general-domain models and fine-tuned clinical foundation models of the same architecture (e.g., BERT vs BioBERT). Bold values indicate the best performance for each task. *P* values were calculated using the Mann-Whitney U test. CI: confidence interval; SE: standard error.

Table 6.5 Fine-tuned model performance on external validation dataset.

Task I - Barrier prediction

Model	Parameters (Total/Tuned)	Mean Macro F1 (95% CI)	ΔMacro-F1	*P* value	Thoughts & Feelings (F1)	Habits & Activities (F1)	Social Situation (F1)	Economic Situation (F1)	Medication (F1)	Care (F1)	Health (F1)	None (F1)
BERT-base	110M/110M	0.64 (0.61-0.67)			0.58	0.66	0.63	0.70	0.49	0.57	0.77	0.70
BioBERT	110M/110M	0.65 (0.62-0.68)	+0.01	<.001	0.57	0.66	0.60	0.70	0.63	0.60	0.81	0.67
BioClinicalBERT	110M/110M	0.63 (0.60-0.67)	-0.01	.003	0.54	0.66	0.59	**0.73**	0.65	0.50	0.77	0.64
Flan-T5-small	77M/77M	0.56 (0.53-0.60)			0.38	0.55	0.53	0.64	0.46	0.56	0.73	0.67
Flan-T5-base	248M/248M	0.68 (0.64-0.70)			0.60	0.72	0.62	0.70	0.63	0.62	0.80	0.73
Clinical-T5-base	248M/248M	0.58 (0.55-0.62)	-0.08	<.001	0.48	0.63	0.45	0.60	0.61	0.53	0.78	0.61
Flan-T5-large	760M/4.7M	0.68 (0.65-0.72)			0.55	**0.76**	0.61	0.69	0.71	**0.64**	**0.83**	0.70
Clinical-T5-large	760M/4.7M	0.55 (0.52-0.59)	-0.13	<.001	0.37	0.53	0.43	0.57	0.60	0.53	0.75	0.67
Flan-T5-xl	3B/9.4M	**0.71 (0.68-0.74)**			**0.69**	**0.76**	**0.66**	0.67	**0.74**	0.58	**0.83**	**0.74**

Task II - Risk level prediction

Model	Parameters (Total/Tuned)	Mean Macro F1 (95% CI)	ΔMacro-F1	*P* value	High (F1)	Medium (F1)	Low (F1)	None (F1)
BERT-base	110M/110M	0.51 (0.47-0.54)			0.36	0.55	0.47	0.65
BioBERT	110M/110M	0.52 (0.48-0.55)	+0.01	<.001	0.38	0.59	0.46	0.64
BioClinicalBERT	110M/110M	0.52 (0.48-0.55)	+0.01	<.001	0.39	0.57	0.49	0.63
Flan-T5-small	77M/77M	0.42 (0.38-0.45)			0.30	0.42	0.41	0.54
Flan-T5-base	248M/248M	0.56 (0.53-0.60)			0.41	0.63	0.52	**0.69**
Clinical-T5-base	248M/248M	0.48 (0.45-0.51)	-0.08	<.001	0.33	0.53	0.48	0.58
Flan-T5-large	760M/4.7M	**0.57 (0.53-0.60)**			0.42	**0.65**	**0.55**	0.65
Clinical-T5-large	760M/4.7M	0.45 (0.42-0.49)	-0.12	<.001	0.37	0.43	0.46	0.56
Flan-T5-xl	3B/9.4M	0.56 (0.52-0.60)			**0.43**	0.62	0.54	0.65

Note. The 95% CIs for Macro-F1 scores were calculated using 300 bootstrap samples with replacement on three distinct datasets, ensuring a SE of the CI limits below 0.01. The observed SE interval limit was 0.0044. ΔMacro-F1 represents the Macro-F1 difference between fine-tuned general-domain models and fine-tuned clinical foundation models of the same architecture (e.g., BERT vs BioBERT). Bold values indicate the best performance for each task. P values were calculated using the Mann-Whitney U test. CI: confidence interval; SE: standard error.

Table 6.6 Best-performing fine-tuned model performance: binary analysis.

Task I - Barrier prediction				
	Dataset	Mean F1 (95% CI)	Barrier (n=840)	*None* (n=540)
Flan-T5-xl	Primary test split	0.94 (0.93-0.95)	0.95	0.92
	External validation	0.85 (0.82-0.88)	0.96	0.74
Task II - Risk level prediction				
	Dataset	Mean F1 (95% CI)	Risk (n=615)	*None* (n=765)
Flan-T5-large	Primary test split	0.90 (0.88-0.92)	0.89	0.91
	External validation	0.76 (0.72-0.79)	0.86	0.65

Note. The 95% CIs for F1 scores were calculated using 300 bootstrap samples with replacement on three distinct datasets, ensuring a SE of the CI limits below 0.01. The observed SE interval limit was 0.0033. CI: confidence interval; SE: standard error.

6.5.5 Error analysis

Table 6.7 summarizes the primary error patterns observed across both tasks with examples. Inductive analysis revealed six distinct error types, stemming from either the annotation process or model inference limitations:

Human annotation–related errors:

- Ambiguous or vague inputs: Short sentences that lack sufficient semantic detail, leading to multiple possible annotations.
- Cross-category cases: Sentences reflecting features of multiple predefined categories, complicating exclusive labelling.
- Inconsistent or incorrect annotations: Sentences inaccurately labeled due to human error during the annotation process.

Model inference-related errors:

- Semantic overlap: Misclassification due conceptually relevant proximity, often driven by over-reliance on shared vocabulary (e.g., classifying sentences containing "holiday" as *Habits & Activities*, when it referred to taking a drug holiday).
- Implicit expression: Failure to detect implied meanings or pragmatic intent, limiting performance on socially or emotionally complex sentences.
- Contextual misinterpretation: Errors caused by information present in the broader message inaccessible at the sentence level.

The most common discrepancies between ground-truth and best-performing model prediction for each task are presented in Supplementary file 2.

6.5.6 Comparison with other LLMs on external validation datasets

Table 6.8 summarizes the results of the best-performing fine-tuned model for each task in comparison with the GPT family and other open-source LLMs on the external validation datasets.

For Task I, Flan-T5-xl outperformed GPT4.1 with five-shot CoT prompting by an overall Δ Macro-F1 = –0.09 (*P*<.001), achieving the highest F1 scores in six out of eight barrier categories. For Task II, Flan-T5-large outperformed Gemma3:14B with zero-shot CoT prompting with an overall Δ Macro-F1 = –0.07 (*P*<.001), achieving the highest F1 scores for the *Medium* and *Low* risk levels. Interestingly, five-shot GPT4.1 received the best score of 0.50 on identifying a *High* risk level.

Table 6.7 Analysis of primary error types.

Types	Examples	Reference label: (Barrier, Risk level)	Model prediction: (Barrier, Risk level)
Human-related errors			
Ambiguous or vague inputs	Medication, it’s everyday?	Medication, Medium	Thoughts & feelings, None
	I want information about atripla.	Thoughts & feelings, Low	Medication, Low
Cross-category cases	Because the hospital, the clinic, these places are not confidential.	Social Situation, Low	Care, Medium
	When I travel, sometimes I'm surrounded by several people. Sometimes, like on those days, I didn't take it.	Social Situation, High	Habits & Activities, High
Inconsistent annotation	sometimes it's a problem.	None, None	Thoughts & feelings, Medium
Model-related errors			
Semantic overlap	Let's say it's a dinner with friends, I could always put a little reminder in my phone, like a little alarm telling me not to forget to take my medication, even if you're having dinner with lots of people, because I can go out and take my medication, and that's that.	Social Situation, Low	Habits & Activities, None
	Are drug holidays safe?	Thoughts & feelings, High	Habits & Activities, Medium
	It's the morning ones I forget, but that has nothing to do with HIV treatment, it's the other condition.	None, None	Habits & Activities, Medium
Implicit expression	Because if I don't take the medication, I can even die.	Thoughts & feelings, None	Thoughts & feelings, High
	So, just in case it might happen that, so someone says: “What pill is that?”, I say: “it ' s a headache pill .”	Social Situation, Low	Medication, Low
Contextual misinterpretation	Artificial dye has done hard on the kidneys.	Medication, Medium	Health, Low
	That is a huge question to me.	Thoughts & feelings, High	None, None

Table 6.8 Best-performing fine-tuned model vs GPT & other open-source LLMs – external validation dataset (Section A)

Task I - Barrier prediction												
Model	Parameters (Total/Tuned)	Mean Macro F1 (95% CI)	ΔMacro-F1	*P* value	Thoughts & Feelings (F1)	Habits & Activities (F1)	Social Situation (F1)	Economic Situation (F1)	Medication (F1)	Care (F1)	Health (F1)	None (F1)
Flan-T5-xl	3B/9.4M	**0.71 (0.68-0.74)**			**0.69**	**0.76**	**0.66**	0.67	**0.74**	0.58	**0.83**	**0.74**
Five-shot	Total parameters											
GPT-4.1	N/A	**0.62 (0.59-0.65)**	-0.09	<.001	0.54	0.70	0.63	0.66	0.73	0.64	0.57	0.51
GPT-4.1-mini	N/A	0.59 (0.56-0.63)	-0.12	<.001	0.52	0.60	0.61	0.76	0.65	0.64	0.51	0.47
GPT-4.1-nano	N/A	0.60 (0.57-0.63)	-0.11	<.001	0.45	0.66	0.53	0.75	0.66	**0.71**	0.58	0.49
Gemma3	14B	0.59 (0.56-0.63)	-0.12	<.001	0.51	0.42	0.57	0.74	0.65	0.58	0.74	0.55
Qwen3	8B	0.60 (0.57-0.63)	-0.11	<.001	0.52	0.62	0.63	**0.83**	0.67	0.60	0.50	0.43
Mistral	7B	0.58 (0.54-0.61)	-0.13	<.001	0.52	0.54	0.52	0.74	0.55	0.59	0.65	0.49
DeepSeek-R1	7B	0.45 (0.42-0.49)	-0.26	<.001	0.39	0.46	0.26	0.70	0.46	0.37	0.52	0.49
LLaMA3.2	3B	0.51 (0.48-0.55)	-0.20	<.001	0.48	0.47	0.37	0.66	0.51	0.43	0.67	0.53
Zero-shot												
GPT-4.1	N/A	0.60 (0.56-0.63)	-0.11	<.001	0.45	0.70	0.63	0.66	0.73	0.60	0.51	0.50
GPT-4.1-mini	N/A	0.56 (0.53-0.60)	-0.15	<.001	0.52	0.65	0.59	0.75	0.67	0.56	0.38	0.43
GPT-4.1-nano	N/A	0.57 (0.54-0.60)	-0.14	<.001	0.44	0.66	0.48	0.69	0.64	0.66	0.52	0.47
Gemma3	14B	0.60 (0.56-0.63)	-0.11	<.001	0.52	0.58	0.57	0.72	0.66	0.57	0.66	0.50
Qwen3	8B	0.55 (0.51-0.58)	-0.16	<.001	0.44	0.64	0.53	0.82	0.67	0.54	0.33	0.41
Mistral	7B	0.50 (0.46-0.53)	-0.21	<.001	0.47	0.44	0.38	0.72	0.46	0.49	0.62	0.40
DeepSeek-R1	7B	0.39 (0.35-0.42)	-0.32	<.001	0.44	0.48	0.22	0.69	0.36	0.18	0.34	0.40
LLaMA3.2	3B	0.33 (0.29-0.36)	-0.38	<.001	0.35	0.31	0.32	0.61	0.31	0.06	0.25	0.39
Task II - Risk level prediction												
Model	Parameters (Total/Tuned)	Mean Macro F1 (95% CI)	Δ F1	P value	High (F1)	Medium (F1)	Low (F1)	None (F1)				
Flan-T5-large	760M/4.7M	**0.57 (0.53-0.60)**			0.42	**0.65**	**0.55**	0.65				
Five-shot	Total parameters											
GPT-4.1	N/A	0.43 (0.40-0.47)	-0.14	<.001	**0.50**	0.13	0.42	**0.67**				
GPT-4.1-mini	N/A	0.42 (0.39-0.45)	-0.15	<.001	0.47	0.21	0.37	0.63				
GPT-4.1-nano	N/A	0.43 (0.40-0.47)	-0.14	<.001	0.47	0.40	0.23	0.62				
Gemma3	14B	0.47 (0.44-0.51)	-0.10	<.001	0.48	0.34	0.44	0.64				
Qwen3	8B	0.33 (0.30-0.37)	-0.24	<.001	0.42	0.13	0.21	0.57				
Mistral	7B	0.38 (0.35-0.41)	-0.19	<.001	0.38	0.33	0.21	0.60				
DeepSeek-R1	7B	0.35 (0.31-0.38)	-0.22	<.001	0.29	0.27	0.28	0.55				
LLaMA3.2	3B	0.28 (0.26-0.31)	-0.29	<.001	0.24	0.07	0.24	0.59				
Zero-shot												
GPT-4.1	N/A	0.45 (0.42-0.48)	-0.12	<.001	0.46	0.32	0.39	0.62				
GPT-4.1-mini	N/A	0.41 (0.37-0.44)	-0.16	<.001	0.46	0.24	0.32	0.61				
GPT-4.1-nano	N/A	0.43 (0.39-0.46)	-0.14	<.001	0.40	0.44	0.26	0.61				
Gemma3	14B	**0.50 (0.46-0.54)**	-0.07	<.001	0.39	0.58	0.42	0.62				
Qwen3	8B	0.30 (0.27-0.34)	-0.27	<.001	0.34	0.11	0.19	0.57				
Mistral	7B	0.32 (0.29-0.34)	-0.25	<.001	0.33	0.34	0.06	0.53				
DeepSeek-R1	7B	0.30 (0.27-0.34)	-0.27	<.001	0.23	0.19	0.26	0.54				
LLaMA3.2	3B	0.31 (0.29-0.34)	-0.26	<.001	0.21	0.12	0.39	0.52				

Table 6.8. Best-performing fine-tuned model vs GPT & other open-source LLMs – external validation dataset (Section B)

Note. Total parameters of the GPT4 family models were not publicly available. The 95% CIs for Macro-F1 scores were calculated using 300 bootstrap samples with replacement on three distinct datasets, ensuring a SE of the CI limits below 0.01. The observed SE interval limits was 0.0039. ΔMacro-F1 represents the Macro-F1 difference between best-performing fine-tuned general-domain models and prompt-tuned GPT and other open-source LLMs. Bold values indicate the best performance for each task among other LLMs. *P* values were calculated using the Mann-Whitney U test, with Bonferroni correction applied for multiple comparisons (adjusted significance threshold: $P < 0.00625$, based on adjusted $\alpha = \alpha/m$, where $\alpha = 0.05$, $m = 8$ (number of tests, 8 models)). CI: confidence interval; SE: standard error; N/A: Not applicable.

Across the remaining models, five-shot CoT prompting generally improved performance over zero-shot prompting for Task I, while both prompting strategies yielded comparable results for Task II.

6.5.7 Assessment of model bias and environmental footprint

Figure 6.2 shows the rate of prediction changes following the injection of race and gender descriptors.

For barrier detection, the overall mismatch rate for fine-tuned Flan-T5-xl was 19.35% (162/837), comparable to GPT-4.1 with five-shot prompting (18.04%, 151/837). However, GPT-4.1 exhibited significantly greater prediction instability for the "Asian" descriptor compared to Flan-t5-xl (*P*=0.003), while Flan-T5-xl was more sensitive to the "Latino" race (*P*=0.03). No significant differences were observed for gender-related prediction changes.

Figure 6.3 and Figure 6.4 show the shifts in predicted barrier types and risk levels changes before and after socio-demographic descriptor injection.

Following descriptor injection, the majority of Flan-t5-xl mismatches shifted to the *Social Situation* category (106/162, 65%), while GPT-4.1 mismatches most frequently shifted to *None* (67/151, 44%).

For risk level detection, Flan-T5-large demonstrated a significantly lower overall mismatch rate (17.80%, 149/837) compared to Gemma3:14B with zero-shot prompting (22.22% (186/837), P=0.024). Gemma3-14B exhibited significantly higher mismatch rates with "Black" (*P*=0.006), "White" (*P*=0.016), and "Man" (*P*=0.003) descriptors. Most mismatches for Flan-t5-large shifted to the *Low* (83/149, 56%) or *Medium* risk levels (36/149, 24%), while Gemma3:14B mismatches most frequently shifted to the *Low* (86/186, 46%) or *None* risk categories (62/186, 33%).

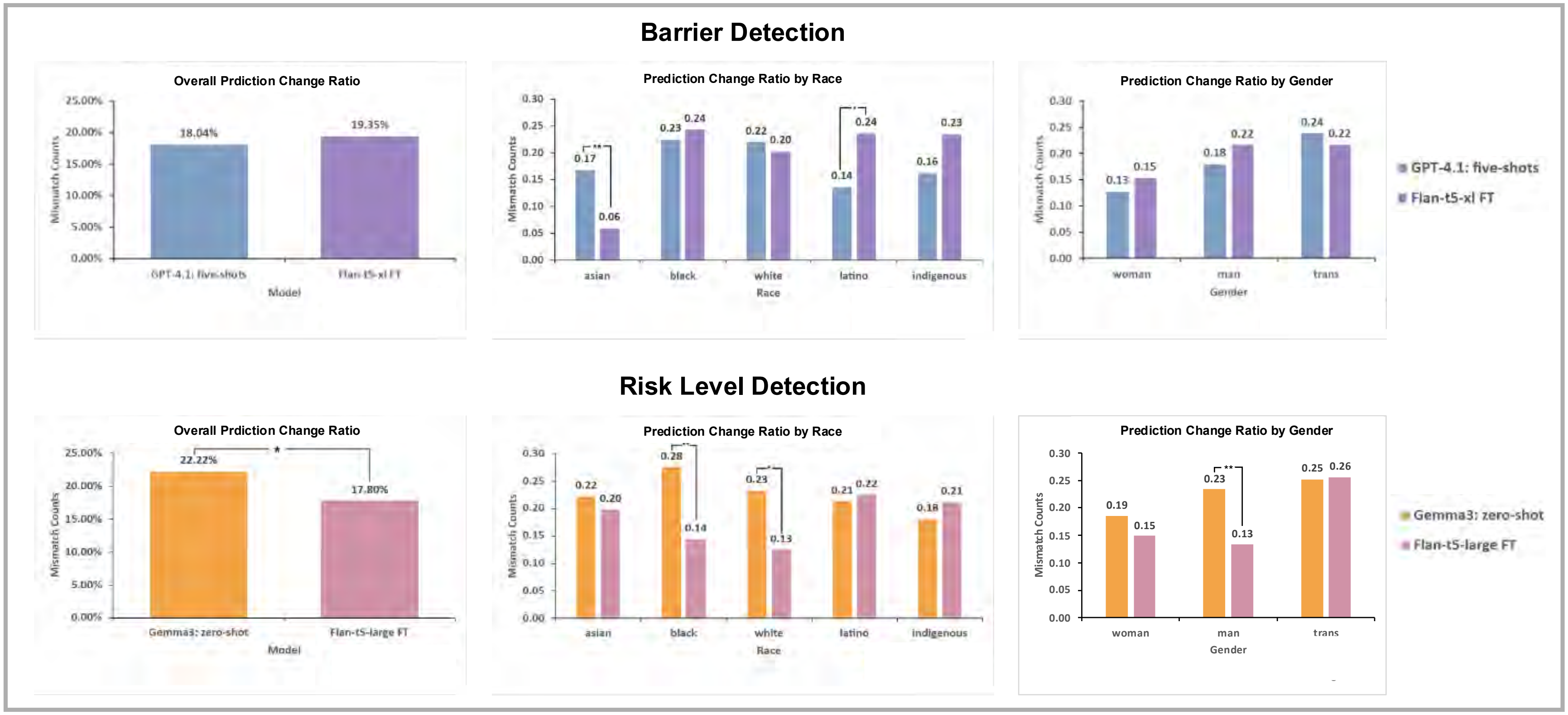
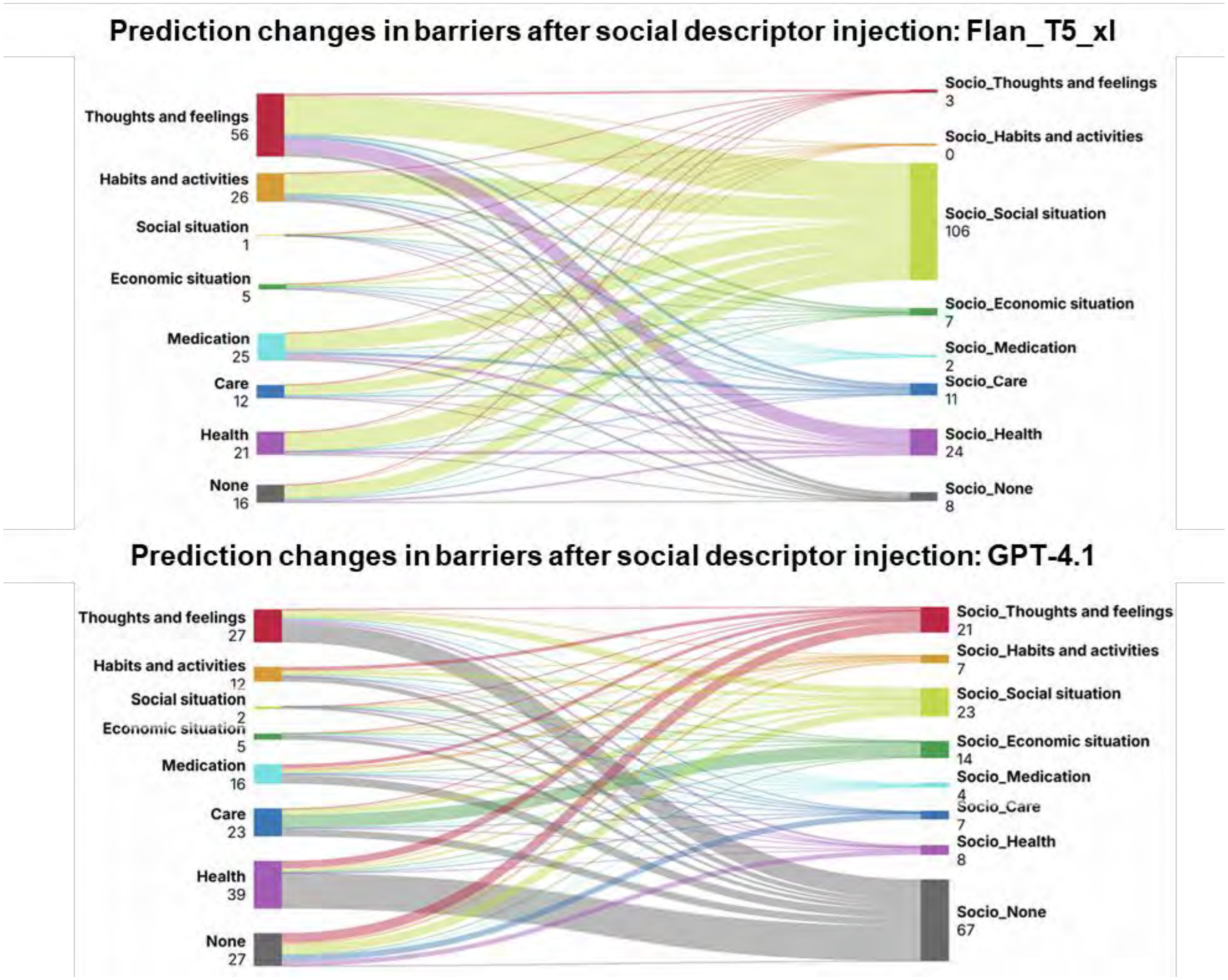


Figure 6.2 Rate of prediction changes following the injection of race and gender descriptors.

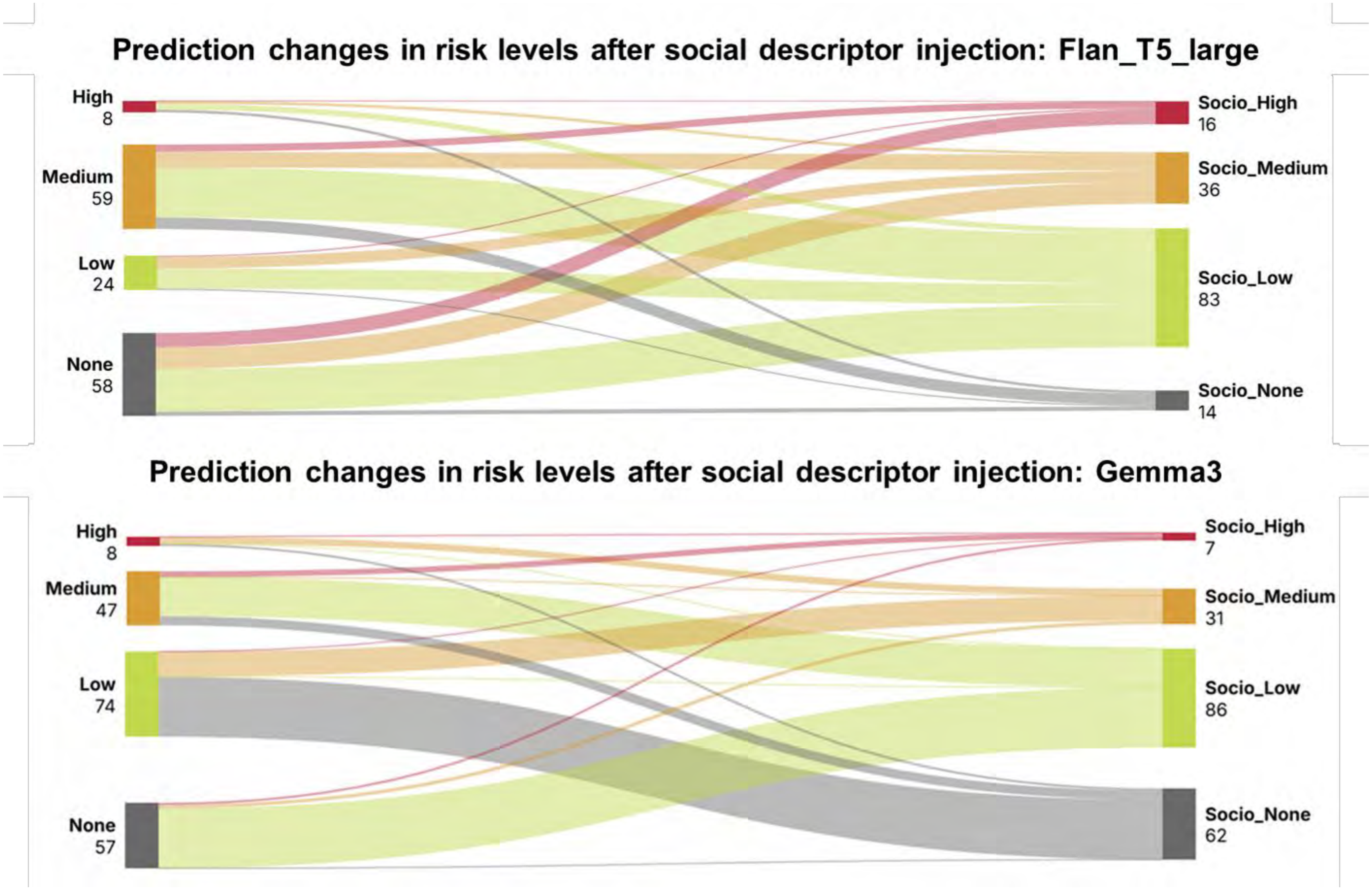
Note. Results are shown across race/ethnicity and gender for (A) Task I – barrier detection and (B) Task II – risk level detection. One asterisk (*) indicates statistical significance at P ⩽ 0.05, and two asterisks (**) indicate P ⩽ 0.01 based on chi-squared tests conducted with each subgroup.



Prediction changes in barriers after social descriptor injection: Flan_T5_xl

Prediction changes in barriers after social descriptor injection: GPT-4.1

Figure 6.3 Shifts in predicted barrier types following racial and gender descriptor injection. left: Before injection; right: after injection.



Prediction changes in risk levels after social descriptor injection: Flan_T5_large

Prediction changes in risk levels after social descriptor injection: Gemma3

Figure 6.4 Shifts in predicted risk levels following racial and gender descriptor injection. left: Before injection; right: after injection.

Table 6.9 summarizes estimated energy consumption and carbon emissions during model fine-tuning and inference. As expected, energy use increased with model size. Fine-tuning Flan-T5-xl for barrier prediction consumed 1.217 kWh energy and generated 0.573kg CO2eq, approximately twice that of Flan-T5-large (0.604kWh, 0.274 kg CO2eq) and over 10 times that of smaller models such as BERT-base (0.091kWh, 0.043 kg CO2eq). A similar trend was observed for risk level detection tasks.

For inference, Flan-t5-xl consumed 0.013 kWh per 1,000 inferences, double that of Flan-T5-large (0.007 kWh). In comparison, Gemma3-14B required 0.043 kWh–over three times higher than Flan-T5-xl–and GPT4.1 consumed an estimated 0.376kWh [73], nearly 30 times that of Flan-T5-xl. Across models, carbon emissions were influenced not only by energy use but also by hosting location, with lower CO2eq reported for models hosted in the Netherlands and Canada, reflecting cleaner energy sources.

6.6 Discussion

We developed a set of LLM-based triage models to identify ART adherence barriers and associated risks in patient messages. Larger Flan-T5 models demonstrated better generalization than smaller BERT-based models across both tasks. We also observed that fine-tuned general-domain models generally outperformed clinical foundation models of similar size and architecture, although this advantage was not consistent, and they exceeded the performance of GPT and other open-source LLMs using zero- and five-shot CoT prompting. Finally, the fine-tuned models showed lower sensitivity to the injection of gender and racial descriptors, particularly for risk level prediction, and offered a more sustainable option for deployment, requiring less energy for inference compared to other LLMs.

Table 6.9 Estimated energy consumption and carbon footprints during model fine-tuning and inference.

	Model	Parameters (Total/tuned)	Energy consumed (kWh)	Carbon footprint (CO2eq, kg)	Hosted country
Fine-tuning					
Barrier	BERT-base	110M/110M	0.091	0.043	Singapore
	BioBERT	110M/110M	0.096	0.045	Singapore
	BioClinicalBERT	110M/110M	0.096	0.045	Singapore
	Flan-T5-small	77M/77M	0.074	0.033	United States
	Flan-T5-base	248M/248M	0.165	0.044	Netherlands
	Clinical-T5-base	248M/248M	0.153	0.041	Netherlands
	Flan-T5-large	760M/4.7M	0.604	0.274	United States
	Clinical-T5-large	760M/4.7M	0.532	0.251	Singapore
	Flan-T5-xl	**3B/9.4M**	**1.217**	**0.573**	**Singapore**
Risk Level	BERT-base	110M/110M	0.096	0.045	Singapore
	BioBERT	110M/110M	0.096	0.045	Singapore
	BioClinicalBERT	110M/110M	0.091	0.043	Singapore
	Flan-T5-small	77M/110M	0.074	0.033	United States
	Flan-T5-base	248M/248M	0.165	0.044	Netherlands
	Clinical-T5-base	248M/248M	0.152	0.041	Netherlands
	Flan-T5-large	**760M/4.7M**	**0.575**	**0.260**	**United States**
	Clinical-T5-large	760M/4.7M	0.440	0.207	Singapore
	Flan-T5-xl	3B/9.4M	1.028	0.484	Singapore
Inference			Energy consumed per 1,000 queries (kWh)		
Barrier	**Flan-T5-xl**	**3B**	**0.013**	**6.00e-03**	**United States**
	GPT-4.1	N/A	0.376	~0.3	United States*
Risk Level	**Flan-T5-large**	**760M**	**0.007**	**3.30e-03**	**United States**
	Gemma3	14B	0.043	1.00e-04	Canada

Note. * Energy consumption and carbon emissions for GPT-4.1 were estimated based on You [73] and Jegham et al. [74], assuming 1,000 input tokens and 20 output tokens per query, resulting in approximately 0.376 kWh per 1,000 inferences. Carbon emissions per 1,000 tokens were estimated at ~0.3 g CO2eq based on Soham [75]. All model training was conducted using Google Colab, leading to variability in the hosting country. Inference for Gemma 3:14B was performed locally at the institution in Montreal, Quebec, Canada. Bold values indicate the best-performing fine-tuned model for each task.

Our developed models demonstrated robust discriminatory performance in classifying ART adherence barriers and associated risk levels from unstructured, patient generated text, achieving Macro-F1 scores of 0.83 and 0.80 on the test split, with corresponding scores of 0.71 and 0.57 on external validation dataset. To our knowledge, this is the first study to address medication adherence challenges directly from free text, in contrast to prior work on AI-based adherence prediction model development that has primarily relied on structured data such as electronic health records [22, 76, 77]. Moreover, most existing models frame adherence prediction as a binary classification task aimed at predicting overall adherence status (adherent/non-adherent). Reported F1 scores have ranged from 0.75 to 0.80 [22, 76, 77], comparable to our best model, Flan-T5-large (F1: 0.90 on the test split, and 0.76 on the external validation dataset). However, binary classification provides a limited perspective, as individuals may be adherent while still experiencing underlying barriers that increase their risk of future non-adherence. Given that ART is a multifactorial challenge, our LLM-based approach – by simultaneously identifying specific barrier types and stratifying risk levels – may provide a more nuanced understanding of individual challenges, supporting tailored clinical management and timely, targeted interventions [78]. However, messages reporting perfect adherence despite explicit barriers were currently classified as “None,” as no current or imminent nonadherence risk was inferred, which may underestimate preventive opportunities. Future work could consider new strategies to flag such cases for proactive support.

Model performance showed minimal differences between general-domain models of varying sizes on the primary test split, but generalization gaps became evident on external datasets - a common phenomenon where smaller models are more prone to overfitting [79]. With advances in resource-efficient adaptation methods such as LoRA, developing task-specific larger models has become more cost-effective, making them a practical option for healthcare settings where computing infrastructure is often limited[80]. Category-level performance was highest for the *None* category across both tasks (F1 = 0.92 for Task I, 0.91 for Task II), likely reflecting its greater representation in the training data. Categories characterized by clear, concrete language, such as *Medication*, *Care*, and *Health*, also performed well (F1 = 0.90, 0.92, and 0.94, respectively). In contrast, categories requiring subjective interpretation, such as *Thoughts & Feelings* and *Social Situation*, showed greater variability (F1 = 0.73 and 0.79, respectively). For instance, a sentence reflecting a social stigma-related adherence barrier was misclassified by the model as *Habits & Activities* based on the presence of the keyword "*travel*". Misclassification within *Thoughts & Feelings* was also common, as negative emotions often co-occur with other barrier types, leading to overlap and misclassification; in risk prediction, a sentence emphasizing the severe consequences of non-adherence - reflecting one's strong motivation rather than actual risk - was incorrectly classified as *High* risk. These errors reflect the known limitations of LLMs in tasks involving implicit reasoning [81], often due to shortcut learning, where models rely on surface cues rather than deeper context [82, 83]. With advances in LLM reasoning capabilities [84], future work may explore whether such approaches can optimize classification performance for these more complex, subjective categories.

Another key factor limiting category-level performance was data imbalance, particularly for *Economic Situation* and *High* risk. To mitigate this challenge, we used annotation guideline-guided prompts with multiple LLMs to augment the training data. However, the results were generally low-quality and highly homogeneous, requiring re-annotation and validation. For *Economic Situation*, augmenting its training share from 1% to 3% moderately improved performance (for Flan-T5-xl: 0.67, > Care=0.58), though still fell short of well-represented categories. Data scarcity remains a recognized challenge in clinical AI [85]. Our study also incorporated the integration of qualitative interview transcripts into the training data – a rich source of domain-specific data that, to the best of our knowledge, is rarely leveraged in clinical NLP tasks. While promising, further research is needed to build more efficient and scalable approaches to address data scarcity in such scenarios.

While domain-specific pretraining is expected to confer advantages for downstream tasks [39-41], our results challenge this assumption. In both adherence barrier and risk classification tasks, clinical foundation models performed significantly worse than general-domain models after fine-tuning on the primary test split (*P*<.001 across all model pairs) and showed no clear advantage on the external validation set (ΔMacro-F1 ranging from +0.01 to -0.13). A likely explanation lies in the nature of our data - conversational, informal and user-generated - which differs substantially from the biomedical literature or electronic health records typically used for clinical pretraining. Moreover, ART medication adherence is shaped not only by biomedical factors, but also by personal experiences and social contexts [86], which are better captured in everyday language. Similar findings have been reported by van Buchem et al. in modeling depression in cancer patients' day-to-day communication [35]. These results echo concerns in the clinical AI community regarding the real-world utility of clinical foundation models [42]. Despite their theoretical promise, their practical value may be limited when tasks fall outside narrowly defined clinical knowledge domains. When developing models for tasks involving user-facing, real-world health communication, fine-tuning clinical foundation models may not be the optimal choice.

The fine-tuned general-domain models also significantly outperformed GPT4.1 and other open-source LLMs on both tasks (min Δ Macro-F1 = 0.09 for barrier detection and 0.10 for risk level detection; $P < 0.001$). Although a five-shot chain-of-thought prompting strategy guided by our annotation framework was applied, all evaluated models, including DeepSeek-R1 and Qwen3 with reasoning capabilities, performed poorly, even underperforming the BERT-base model. Interestingly, GPT-4.1 achieved the highest F1 score (0.50) for the *High* risk category, indicating some ability to detect serious adherence risks, consistent with prior findings [45]. Nevertheless, the overall performance gap suggests the limitations of deploying off-the-shelf LLMs for this task and reinforces the need for task-specific model development.

In assessing model bias, our best-performing fine-tuned model and GPT-4.1 demonstrated comparable rates of prediction mismatches for adherence barriers following the injection of racial and gender descriptors (18% vs. 19%). This responsiveness can be beneficial, as adherence barriers may manifest differently depending on the individual’s social identity. Fairness in this context should be understood as equitable reasoning across social groups, rather than equal accuracy alone. Notably, 65% of the mismatches in our model shifted to the *Social Situation* category - closely linked to HIV-related stigma - suggesting that the model may be appropriately sensitive to relevant social factors. In contrast, nearly half of GPT-4.1's mismatches were reclassified as *None*, raising concerns about false negatives that could lead to missed opportunities to identify barriers. For the risk prediction, Gemma3 exhibited a 25% higher mismatch rate than our model. Predictions shifted more significantly in response to *Black*, *White, Man* and *Transgender* descriptors compared to our model, echoing concerns that LLMs may amplify biases, prejudices, and racism present in the language they are trained on [68]. Gemma3 also reclassified approximately 30% of affected sentences to the *None* risk category, again suggesting a higher risk of false negatives. All of this suggests that our model is more reliable in this task than other LLMs. That said, injecting demographic descriptors remains an artificial approach of stress testing and may not reflect real-world language use. Future evaluations should incorporate participatory assessments with individuals from diverse racial and gender backgrounds to better assess model fairness in authentic contexts. Given the diversity of the PWH population, ensuring that LLM-based triage models are developed and deployed equitably with attention to bias is critical to avoid perpetuating health disparities.

Finally, the environmental footprints of model development followed expected trends, with energy consumption increasing alongside model size. For barrier prediction, Flan-T5-xl achieved a modest performance gain over Flan-T5-large on the external dataset (Δ Macro-F1 = 0.03) but consumed more than twice the energy (1.217 kWh vs. 0.604 kWh), equivalent to running a MacBook Air continuously for 24 hours. For risk prediction, the smaller Flan-T5-large demonstrated better generalization performance while consuming less energy, highlighting the trade-offs between model size, performance, and sustainability. These results raise an important question: Are the marginal performance gains of larger models sufficient to offset the environmental costs? A similar trade-off was observed during inference, where both fine-tuned models were substantially more energy-efficient than proprietary GPT-4.1, reinforcing the deployment advantages of smaller, task-specific models. Interestingly, although Gemma3 consumed more energy per inference than Flan-T5-xl, its carbon footprint was lower due to hosting on Quebec's low-carbon energy grid. These findings underscore the importance of considering both model selection and deployment infrastructure when developing clinical AI tools, particularly as clinical AI continues to scale.

We acknowledge several limitations of this study. First, the annotation process was labor-intensive, requiring great manual effort from both patient partners, clinical experts and researchers. Despite iterative guideline refinement, human labeling errors were inevitable, particularly for risk assessment, which requires subjective judgment and domain-specific knowledge. Messages mentioning barriers without explicit adherence-level information required annotator interpretation, introducing potential subjective judgment bias. This subjectivity likely contributed to the lower inter-rater agreement for risk labeling and may partly explain the weaker generalization of risk prediction models on external datasets. While LLMs such as GPT-4 have been proposed to assist with automated labeling [87, 88], our experiments showed that GPT-4 achieved only moderate agreement with human annotations, underscoring the current limitations of relying on LLMs for unsupervised dataset development. Second, parts of the training data required translation from French to English, which may have introduced linguistic biases. Third, our single-label classification framework may not capture the multidimensional nature of adherence barriers, which often intersect across categories. Current models can identify the corresponding frequencies of different barriers and risks within a message. Future research will explore the frequencies and interrelationships of barriers and risks to better personalize appropriate interventions as well as the application of multi-label classification to better reflect this complexity.

AI can help extend the reach of overstretched healthcare professionals by automating routine functions such as triage, follow-up reminders, and adherence checks [89]. In this study, we developed LLM-based models that can effectively detect ART adherence barriers and stratify associated risks from patient-generated text, offering a novel, scalable solution for individualized HIV care. These findings show that beyond generic capabilities, tailored LLMs can operationalize nuanced understanding of patient challenges, with the potential to inform real-time, personalized interventions. We plan to integrate these models into our MARVIN chatbot and a patient portal, with upcoming real-world validation to assess feasibility and clinical utility. Beyond technical performance, attention must be paid to ensuring these tools are acceptable to end users, integrate seamlessly into clinical workflows, and equitably serve diverse populations. Future research should also explore practice-based learning, enabling continuous model improvement by cycling back real-world data, and expanding model capabilities to support timely, personalized interventions. These efforts will be critical for advancing the responsible and effective application of AI in HIV care.

6.7 Data availability

Examples of the annotated dataset and the complete synthetic dataset used in this study is available at: https://github.com/yma-94/ART_Adherence.

The full dataset, including the MARVIN training corpus (MT) and user conversations (MU), the I-Score study interview transcripts (ISCORE), and the Portail VIH/SIDA du Quebec messages (PVSQ) and the associated demographic-injected dataset, is available from the corresponding author (B Lebouché) on reasonable request but are not publicly released due to privacy considerations.

6.8 Code availability

The code being used in the current study for developing the model is provided via Github at https://github.com/yma-94/ART_Adherence.

6.9 Acknowledgements

This study is supported by the Fonds de recherche du Québec–Santé (FRQS) – AIDS and Infectious Diseases Network (PI: B Lebouché), and the Seed financing program 2025 of the International Observatory on the Societal Impacts of AI and Digital Technologies (OBVIA) (PI: B Lebouché, Grant ID: 268938).

YM is supported by the Postgraduate Scholarship – Doctoral program (PGS D) and the OPSIDIAN program from the Natural Sciences and Engineering Research Council, and a doctoral research award from the Fonds de recherche Nature et Technologies (FRQNT) in partnership with the Unité de soutien au système de santé apprenant (SSA) Québec.

B Lebouché is supported by two career awards: a Senior Salary Award from Fonds de recherche du Québec–Santé (FRQS) (#311200) and the Lettre d'Entente 250 from the Quebec Ministry of Health for researchers in Family Medicine.

The funding sources had no role in the design of this study, the interpretation of the results, or the decision to submit them for publication.

The authors thank the Chronic Viral Illness Service research team.

The names of the members of the MARVIN chatbots Patient Expert Committee cannot be provided as they are patients, and their names must remain confidential.

6.10 Author contributions

In no order of contribution, YM: conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, visualization, writing – original draft, writing – review & editing. DL, KE, SV: validation, writing – review & editing. GT, B Lemire, LDB, NP, MARVIN chatbots Patient Expert Committee: Data curation, software, validation, writing -review & editing. SA, B Lebouché: conceptualization, funding acquisition, project administration, resources, supervision, writing – review & editing.

6.11 Competing interests

B Lebouché has received research support, consulting fees, and speaker fees from ViiV Healthcare, Merck, and Gilead.

6.12 References

[1] F. Nakagawa, M. May, and A. Phillips, "Life expectancy living with HIV: recent estimates and future implications," *Current opinion in infectious diseases,* vol. 26, no. 1, pp. 17-25, 2013.

[2] C. J. Colvin, "HIV/AIDS, chronic diseases and globalisation," *Globalization and health,* vol. 7, no. 1, pp. 1-6, 2011.

[3] J. B. Angel *et al.*, "Adherence to oral antiretroviral therapy in Canada, 2010-2020," *AIDS,* vol. 37, no. 13, pp. 2031-2040, Nov 1 2023, doi: 10.1097/QAD.0000000000003648.

[4] G. A. McComsey, M. Lingohr-Smith, R. Rogers, J. Lin, and P. Donga, "Real-World Adherence to Antiretroviral Therapy Among HIV-1 Patients Across the United States," *Adv Ther,* vol. 38, no. 9, pp. 4961-4974, Sep 2021, doi: 10.1007/s12325-021-01883-8.

[5] P. de Los Rios *et al.*, "Prevalence, determinants, and impact of suboptimal adherence to HIV medication in 25 countries," *Prev Med,* vol. 139, p. 106182, Oct 2020, doi: 10.1016/j.ypmed.2020.106182.

[6] E. O'Halloran Leach, H. Lu, J. Caballero, J. E. Thomas, E. C. Spencer, and R. L. Cook, "Defining the optimal cut-point of self-reported ART adherence to achieve viral suppression in the era of contemporary HIV therapy: a cross-sectional study," *AIDS Res Ther,* vol. 18, no. 1, p. 36, Jun 26 2021, doi: 10.1186/s12981-021-00358-8.

[7] K. K. Byrd *et al.*, "Antiretroviral Adherence Level Necessary for HIV Viral Suppression Using Real-World Data," *J Acquir Immune Defic Syndr,* vol. 82, no. 3, pp. 245-251, Nov 1 2019, doi: 10.1097/QAI.0000000000002142.

[8] V. D. Lima *et al.*, "The combined effect of modern highly active antiretroviral therapy regimens and adherence on mortality over time," *JAIDS Journal of Acquired Immune Deficiency Syndromes,* vol. 50, no. 5, pp. 529-536, 2009.

[9] P. R. Harrigan *et al.*, "Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy," *The Journal of infectious diseases,* vol. 191, no. 3, pp. 339-347, 2005.

[10] T. Endebu, G. Taye, and W. Deressa, "Rate and predictors of loss to follow-up in HIV care in a low-resource setting: analyzing critical risk periods," *BMC Infectious Diseases,* vol. 24, no. 1, 2024-10-18 2024, doi: 10.1186/s12879-024-10089-6.

[11] J. H. McMahon *et al.*, "Pharmacy Adherence Measures to Assess Adherence to Antiretroviral Therapy: Review of the Literature and Implications for Treatment Monitoring," *Clinical Infectious Diseases,* vol. 52, no. 4, pp. 493-506, 2011, doi: 10.1093/cid/ciq167.

[12] M. S. Saag, "HIV Infection—Screening, Diagnosis, and Treatment," *New England Journal of Medicine,* vol. 384, no. 22, pp. 2131-2143, 2021.

[13] P. o. A. G. f. A. a. Adolescents., "Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV. Department of Health and Human Services.," 2024. [Online]. Available: https://clinicalinfo.hiv.gov/en/guidelines/adult-and-adolescent-arv.

[14] E. A. C. Society, "European AIDS Clinical Society Guidelines 2024," 2024. [Online]. Available: https://eacs.sanfordguide.com/.

[15] C. Okoli *et al.*, "Shared Decision Making Between Patients and Healthcare Providers and its Association with Favorable Health Outcomes Among People Living with HIV," *AIDS Behav,* vol. 25, no. 5, pp. 1384-1395, May 2021, doi: 10.1007/s10461-020-02973-4.

[16] V. D. Miron *et al.*, "Perception of medical care among women living with HIV aged 40 years or older-A European-wide survey," *HIV Med,* vol. 26, no. 3, pp. 451-464, Mar 2025, doi: 10.1111/hiv.13749.

[17] H. Budhwani *et al.*, "Patient Health Literacy and Communication with Providers Among Women Living with HIV: A Mixed Methods Study," *AIDS Behav,* vol. 26, no. 5, pp. 1422-1430, May 2022, doi: 10.1007/s10461-021-03496-2.

[18] A. Norberg, J. Nelson, C. Holly, S. T. Jewell, M. Lieggi, and S. Salmond, "Experiences of HIV-infected adults and healthcare providers with healthcare delivery practices that influence engagement in US primary healthcare settings: a qualitative systematic review," *JBI Evidence Synthesis,* vol. 17, no. 6, pp. 1154-1228, 2019, doi: 10.11124/jbisrir-2017-003756.

[19] M. Kall, F. Marcellin, R. Harding, J. V. Lazarus, and P. Carrieri, "Patient-reported outcomes to enhance person-centred HIV care," *The lancet HIV,* vol. 7, no. 1, pp. e59-e68, 2020.

[20] R. Izquierdo *et al.*, "Health‐related quality of life in people with HIV from the multicentre CoRIS cohort in Spain: Associated factors and short‐term changes over time," *HIV medicine,* vol. 26, no. 4, pp. 606-620, 2025.

[21] A. Galozy and S. Nowaczyk, "Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data," *J Biomed Inform,* vol. 112S, p. 100075, 2020, doi: 10.1016/j.yjbinx.2020.100075.

[22] Y. Gu *et al.*, "Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data," *Sci Rep,* vol. 11, no. 1, p. 18961, Sep 23 2021, doi: 10.1038/s41598-021-98387-w.

[23] L. L. Zullig *et al.*, "Novel application of approaches to predicting medication adherence using medical claims data," *Health Services Research,* vol. 54, no. 6, pp. 1255-1262, 2019-12-01 2019, doi: 10.1111/1475-6773.13200.

[24] K. Jiao *et al.*, "Effectiveness of instant versus text messaging intervention on antiretroviral therapy adherence among men who have sex with men living with HIV," *DIGITAL HEALTH,* vol. 10, 2024-01-01 2024, doi: 10.1177/20552076241257447.

[25] M. A. Lewis *et al.*, "Tailored text messaging intervention for HIV adherence: a proof-of-concept study," *Health Psychology,* vol. 32, no. 3, p. 248, 2013.

[26] A. I. Rana, J. J. van den Berg, E. Lamy, and C. G. Beckwith, "Using a mobile health intervention to support HIV treatment adherence and retention among patients at risk for disengaging with care," *AIDS patient care and STDs,* vol. 30, no. 4, pp. 178-184, 2016.

[27] R. Dillingham *et al.*, "PositiveLinks: a mobile health intervention for retention in HIV care and clinical outcomes with 12-month follow-up," *AIDS patient care and STDs,* vol. 32, no. 6, pp. 241-250, 2018.

[28] A.-M. Dunn-Navarra *et al.*, "Developing and testing a web-based platform for antiretroviral therapy ART adherence support among adolescents and young adults AYA living with HIV," *PEC Innovation,* p. 100263, 2024.

[29] A. M. Midboe *et al.*, "Relationship Between Patient Portal Tool Use and Medication Adherence and Viral Load Among Patients Living with HIV," *J Gen Intern Med,* vol. 39, no. Suppl 1, pp. 127-135, Feb 2024, doi: 10.1007/s11606-023-08474-z.

[30] S. Shourya *et al.*, "A Remote Intervention Based on mHealth and Community Health Workers for Antiretroviral Therapy Adherence in People With HIV: Pilot Randomized Controlled Trial," *JMIR Form Res,* vol. 9, p. e67997, Apr 2 2025, doi: 10.2196/67997.

[31] Y. Ma *et al.*, "The first AI-based Chatbot to promote HIV self-management: A mixed methods usability study," *HIV Med,* Oct 10 2024, doi: 10.1111/hiv.13720.

[32] M. Hui, "Testing the Feasibility and Acceptability of Using an Artificial Intelligence Chatbot to Promote HIV Testing and Pre-Exposure Prophylaxis in Malaysia: Mixed Methods Study," *JMIR Hum Factors 2024;11:e52055 https://humanfactors.jmir.org/2024/1/e52055,* vol. 11, no. 1, 2024-01-26, doi: 10.2196/52055.

[33] A. van Heerden *et al.*, "Chatbots for HIV Prevention and Care: a Narrative Review," *Current HIV/AIDS Reports 2023 20:6,* vol. 20, no. 6, 2023-11-27, doi: 10.1007/s11904-023-00681-x.

[34] K. S. Ingersoll *et al.*, "Pilot RCT of bidirectional text messaging for ART adherence among nonurban substance users with HIV," *Health Psychology,* vol. 34, no. S, p. 1305, 2015.

[35] M. M. van Buchem *et al.*, "Applying natural language processing to patient messages to identify depression concerns in cancer patients," *Journal of the American Medical Informatics Association*, vol. 31, no. 10, pp. 2255-2262, 2024.

[36] A. R. Bhandarkar *et al.*, "Building a Natural Language Processing Artificial Intelligence to Predict Suicide-Related Events Based on Patient Portal Message Data," *Mayo Clin Proc Digit Health,* vol. 1, no. 4, pp. 510-518, Dec 2023, doi: 10.1016/j.mcpdig.2023.09.001.

[37] A. Swaminathan *et al.*, "Natural language processing system for rapid detection and intervention of mental health crisis chat messages," *NPJ Digit Med,* vol. 6, no. 1, p. 213, Nov 21 2023, doi: 10.1038/s41746-023-00951-3.

[38] A. A. H. de Hond *et al.*, "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review," *NPJ Digit Med*, vol. 5, no. 1, p. 2, Jan 10 2022, doi: 10.1038/s41746-021-00549-7.

[39] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, Feb 15 2020, doi: 10.1093/bioinformatics/btz682.

[40] E. Lehman *et al.*, "Do we still need clinical language models?," in *Conference on health, inference, and learning*, 2023: PMLR, pp. 578-597.

[41] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," *arXiv preprint arXiv:1904.03323,* 2019.

[42] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," *NPJ Digit Med,* vol. 6, no. 1, p. 135, Jul 29 2023, doi: 10.1038/s41746-023-00879-8.

[43] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit Health,* vol. 2, no. 2, p. e0000198, Feb 2023, doi: 10.1371/journal.pdig.0000198.

[44] R. D. Chandler, S. Warner, G. Aidoo-Frimpong, and J. Wells, "“What Did You Say, ChatGPT?” The Use of AI in Black Women's HIV Self-Education: An Inductive Qualitative Data Analysis," *Journal of the Association of Nurses in AIDS Care,* vol. 35, no. 3, pp. 294-302, 2024, doi: 10.1097/jnc.0000000000000468.

[45] M. C. Y. Koh, J. N. Ngiam, J. Yong, P. A. Tambyah, and S. Archuleta, "The role of an artificial intelligence model in antiretroviral therapy counselling and advice for people living with HIV," *HIV Medicine,* 2024-01-02 2024, doi: 10.1111/hiv.13604.

[46] L. Cheng, K. R. Varshney, and H. Liu, "Socially responsible ai algorithms: Issues, purposes, and challenges," *Journal of Artificial Intelligence Research,* vol. 71, pp. 1137-1181, 2021.

[47] T. Zack *et al.*, "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study," *Lancet Digit Health,* vol. 6, no. 1, pp. e12-e22, Jan 2024, doi: 10.1016/S2589-7500(23)00225-X.

[48] M. R. Loutfy *et al.*, "Gender and ethnicity differences in HIV-related stigma experienced by people living with HIV in Ontario, Canada," *PLoS One,* vol. 7, no. 12, p. e48168, 2012, doi: 10.1371/journal.pone.0048168.

[49] C. H. Logie *et al.*, "HIV-related stigma, racial discrimination, and gender discrimination: Pathways to physical and mental health-related quality of life among a national cohort of women living with HIV," *Prev Med,* vol. 107, pp. 36-44, Feb 2018, doi: 10.1016/j.ypmed.2017.12.018.

[50] H. Siala and Y. Wang, "SHIFTing artificial intelligence to be responsible in healthcare: A systematic review," *Soc Sci Med,* vol. 296, p. 114782, Mar 2022, doi: 10.1016/j.socscimed.2022.114782.

[51] A. Fiske, I. M. Radhuber, T. Willem, A. Buyx, L. A. Celi, and S. McLennan, "Climate change and health: the next challenge of ethical AI," *The Lancet Global Health,* vol. 13, no. 7, pp. e1314-e1320, 2025, doi: 10.1016/s2214-109x(25)00124-x.

[52] E. Osmanlliu *et al.*, "The Urgency of Environmentally Sustainable and Socially Just Deployment of Artificial Intelligence in Health Care," *Catalyst non-issue content,* vol. 6, no. 4, p. CAT.24.0501, 2025, doi: doi:10.1056/CAT.24.0501.

[53] C. Spinuzzi, "The methodology of participatory design," *Technical communication,* vol. 52, no. 2, pp. 163-174, 2005.

[54] M. Camacho, J. Perramon, X. Puig-Bosch, V. N. Dang, O. Díaz, and K. Lekadir, "Stakeholder engagement: the path to trustworthy AI in healthcare," in *Trustworthy AI in Medical Imaging*: Elsevier, 2025, pp. 471-493.

[55] Y. Ma *et al.*, "Adapting and Evaluating an AI-Based Chatbot Through Patient and Stakeholder Engagement to Provide Information for Different Health Conditions: Master Protocol for an Adaptive Platform Trial (the MARVIN Chatbots Study)," *JMIR Research Protocols,* vol. 13, 2024, doi: 10.2196/54668.

[56] K. Engler, S. Vicente, K. K. Mate, D. Lessard, S. Ahmed, and B. Lebouché, "Content validation of a new measure of patient-reported barriers to antiretroviral therapy adherence, the I-Score: results from a Delphi study," *Journal of patient-reported outcomes,* vol. 6, no. 1, pp. 1-12, 2022.

[57] W. M. Bezabhe, L. Chalmers, L. R. Bereznicki, and G. M. Peterson, "Adherence to antiretroviral therapy and virologic failure: a meta-analysis," *Medicine,* vol. 95, no. 15, p. e3361, 2016.

[58] "Reddit r/HIV." https://www.reddit.com/r/HIV/ (accessed 27 June, 2025).

[59] "Reddit r/HIVAIDS." https://www.reddit.com/r/hivaids/ (accessed June 27, 2025).

[60] "POZ community Forum." https://forums.poz.com/ (accessed June 27, 2025).

[61] "Portail VIH/Sida Québec." https://pvsq.org/ (accessed June 29, 2025).

[62] *syntok: Text tokenization and sentence segmentation (segtok v2).* (2022). Github.

[63] M. Giuffre and D. L. Shung, "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy," *NPJ Digit Med,* vol. 6, no. 1, p. 186, Oct 9 2023, doi: 10.1038/s41746-023-00927-3.

[64] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine,* vol. 15, no. 2, pp. 155-163, 2016.

[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[66] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research,* vol. 25, no. 70, pp. 1-53, 2024.

[67] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR,* vol. 1, no. 2, p. 3, 2022.

[68] M. Omar *et al.*, "Sociodemographic biases in medical decision making by large language models," *Nat Med,* Apr 7 2025, doi: 10.1038/s41591-025-03626-6.

[69] M. Guevara *et al.*, "Large Language Models to Identify Social Determinants of Health in Electronic Health Records," *arXiv preprint arXiv:2308.06354,* 2023.

[70] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice," *Frontiers in Public Health,* vol. 5, 2017-11-20 2017, doi: 10.3389/fpubh.2017.00307.

[71] *mlco2/codecarbon: v3.0.2*. (2025). Zenodo. Accessed: June 26. [Online]. Available: https://doi.org/10.5281/zenodo.11171501

[72] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature methods,* vol. 17, no. 3, pp. 261-272, 2020.

[73] J. You. "How much energy does ChatGPT use?" https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use?utm_source=chatgpt.com (accessed July 2, 2025.

[74] N. Jegham, M. Abdelatti, L. Elmoubarki, and A. Hendawi, "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference," *arXiv preprint arXiv:2505.09598,* 2025.

[75] Soham. "The Cost of Inference: Running the Models." https://tinyml.substack.com/p/the-cost-of-inference-running-the (accessed 30 July, 2025).

[76] V. Koesmahargyo *et al.*, "Accuracy of machine learning-based prediction of medication adherence in clinical research," *Psychiatry Res,* vol. 294, p. 113558, Dec 2020, doi: 10.1016/j.psychres.2020.113558.

[77] K. Haas, Z. Ben Miled, and M. Mahoui, "Medication Adherence Prediction Through Online Social Forums: A Case Study of Fibromyalgia," *JMIR Med Inform,* vol. 7, no. 2, p. e12561, Apr 4 2019, doi: 10.2196/12561.

[78] K. Engler, A. Lènàrt, D. Lessard, I. Toupin, and B. Lebouché, "Barriers to antiretroviral therapy adherence in developed countries: a qualitative synthesis to develop a conceptual framework for a new patient-reported outcome measure," *AIDS Care,* vol. 30, no. sup1, pp. 17-28, 2018-12-14 2018, doi: 10.1080/09540121.2018.1469725.

[79] J. Lever, M. Krzywinski, and N. Altman, "Model selection and overfitting," *Nature Methods,* vol. 13, no. 9, pp. 703-704, 2016, doi: 10.1038/nmeth.3968.

[80] J. C. C. Kwong, G. C. Nickel, S. C. Y. Wang, and J. C. Kvedar, "Integrating artificial intelligence into healthcare systems: more than just the algorithm," *npj Digital Medicine,* vol. 7, no. 1, 2024-03-01 2024, doi: 10.1038/s41746-024-01066-z.

[81] V. Dentella, F. Gunther, E. Murphy, G. Marcus, and E. Leivada, "Testing AI on language comprehension tasks reveals insensitivity to underlying meaning," *Sci Rep,* vol. 14, no. 1, p. 28083, Nov 14 2024, doi: 10.1038/s41598-024-79531-8.

[82] A. Brown, N. Tomasev, J. Freyberg, Y. Liu, A. Karthikesalingam, and J. Schrouff, "Detecting shortcut learning for fair medical AI using shortcut testing," *Nature Communications,* vol. 14, no. 1, 2023-07-18 2023, doi: 10.1038/s41467-023-39902-7.

[83] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nature Machine Intelligence,* vol. 2, no. 11, pp. 665-673, 2020-11-10 2020, doi: 10.1038/s42256-020-00257-z.

[84] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403,* 2022.

[85] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat Med,* vol. 28, no. 1, pp. 31-38, Jan 2022, doi: 10.1038/s41591-021-01614-0.

[86] K. K. V. Mate, K. Engler, D. Lessard, and B. Lebouché, "Barriers to adherence to antiretroviral therapy: identifying priority areas for people with HIV and healthcare professionals," *International Journal of STD & AIDS*, vol. 34, no. 10, pp. 677-686, 2023-09-01 2023, doi: 10.1177/09564624231169329.

[87] B. Ding, C. Qin, L. Liu, L. Bing, S. Joty, and B. Li, "Is gpt-3 a good data annotator?," *arXiv preprint arXiv:2212.10450*, 2022.

[88] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," *arXiv preprint arXiv:2108.13487*, 2021.

[89] J. Ratevosian *et al.*, "Reimagining HIV prevention with artificial intelligence," *Lancet HIV*, Jun 11 2025, doi: 10.1016/S2352-3018(25)00158-4.

CHAPTER 7 GENERAL DISCUSSION

This thesis presented the design and development of MARVIN, an AI-based chatbot designed to support self-management among PWH. It addressed several critical gaps identified in Chapter 2: (1) the limited availability of chatbot interventions for HIV self-management, coupled with a lack of real-world evidence supporting their use; (2) the absence of systematic, theory-driven frameworks to guide their development, implementation, and evaluation; and (3) the underutilization of state-of-the-art LLMs in HIV care. The three research articles presented in Chapters 4, 5, and 6 collectively span key phases of the digital health intervention development lifecycle, from design and implementation planning to real-world usability evaluation and continuous refinement, while grounding each phase in participatory design and aligning with the SHIFT principles of sustainable, human-centered, inclusive, fair, and transparent AI.

7.1 Overview of contributions across the chatbot design and development lifecycle

Figure 7.1 illustrates how these contributions map onto the chatbot design and development lifecycle. Chapter 4 covers the design, development, and implementation planning stages of the chatbot. Chapter 5 focuses with the evaluation and dissemination phase through real-world implementation, with its results serving as a needs assessment for subsequent development. Chapter 6 supports the design and development phase by building an LLM-based triage tool to enhance MARVIN’s intelligence and personalization. Together, these studies reflect an iterative, end-to-end approach to digital health intervention development.

As identified in Chapter 2, PWH expressed a need for accessible and reliable resources to better support their daily self-management. The first article (Chapter 4) described the general design and development approach of MARVIN, in collaboration with PWH, healthcare professionals, and other stakeholders. Table 7.1 summarizes MARVIN’s status based on the chatbot taxonomy criteria proposed in Section 2.2.1.2. Given the chronic nature of HIV, MARVIN was designed and developed as a lifelong companion to assist with everyday self-management tasks. The chatbot delivers validated health information across a range of topics, including antiretroviral treatment (ART) adherence, ART management during travel, and common HIV-related questions. It also provides customizable medication reminders, tailored responses based on user input (e.g., ART treatment name and delays, as shown in Figure 4.6), and uses a transformer-based NLP model with hybrid dialogue management techniques. Initially deployed on Facebook Messenger, MARVIN was later expanded to a standalone web interface (see Figure 7.2) in response to user feedback. Plans are in progress to expand its availability to additional platforms, such as WhatsApp and Instagram, with the aim of improving accessibility and enhancing user experience.

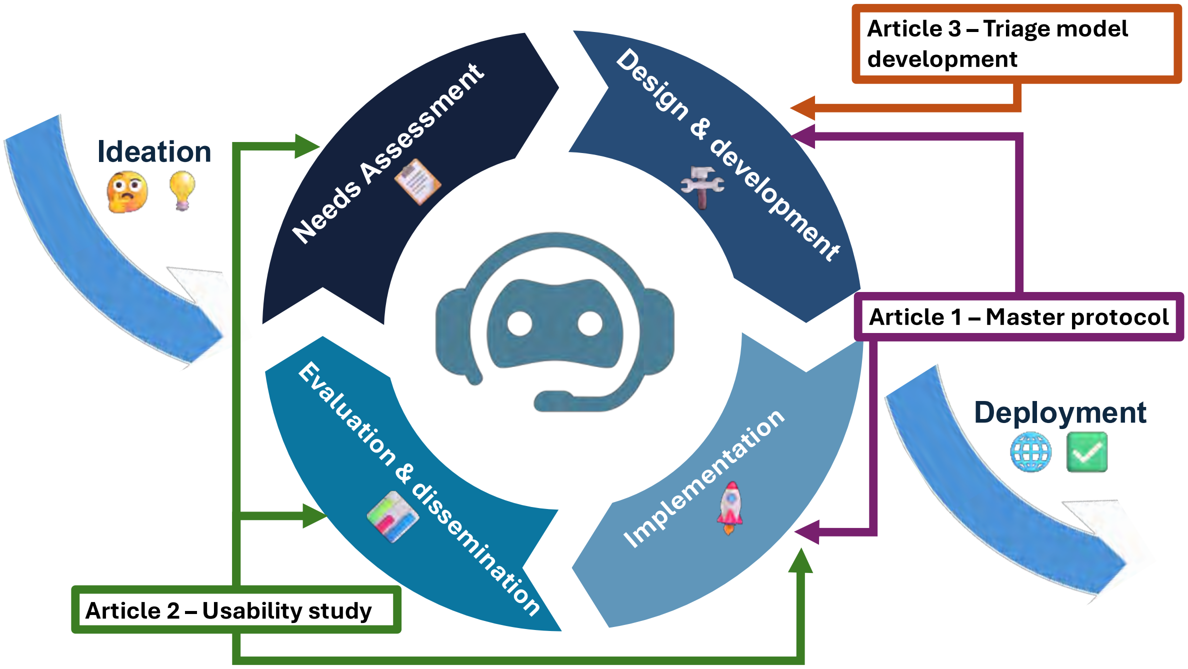
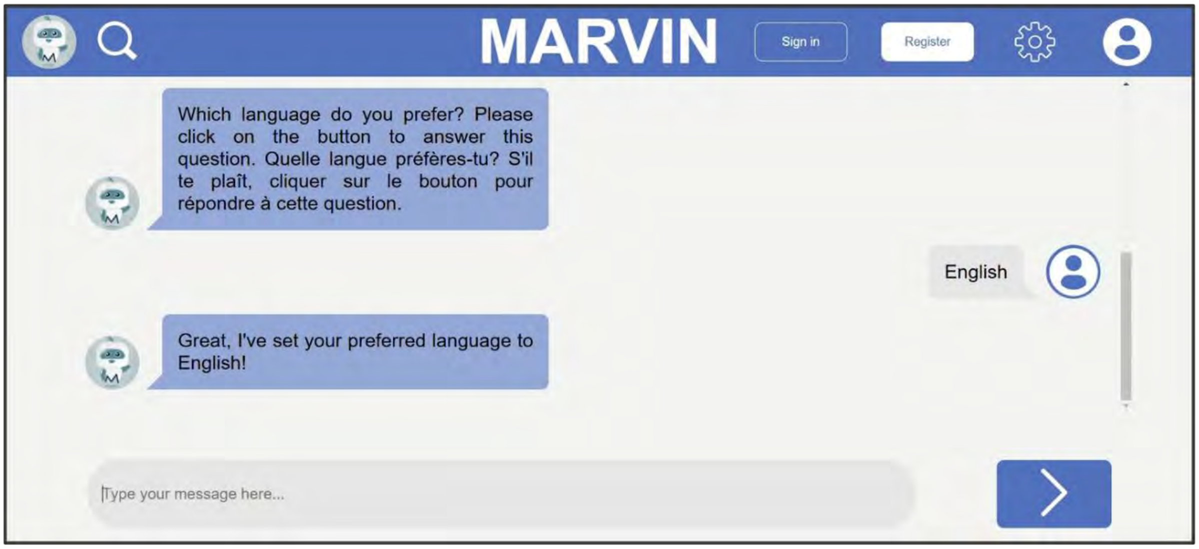

Article 3 – Triage model development
Ideation
Needs Assessment
Design & development
Article 1 – Master protocol
Deployment
Evaluation & dissemination
Implementation
Article 2 – Usability study

Figure 7.1 Mapping of the three thesis articles onto the chatbot development lifecycle.



Which language do you prefer? Please click on the button to answer this question. Quelle langue préfères-tu? S'il te plaît, cliquer sur le bouton pour répondre à cette question.

English

Great, I've set your preferred language to English!

Figure 7.2 MARVIN web interface developed in response to user feedback.

Table 7.1 Current status of MARVIN development based on the chatbot taxonomy proposed by Janssen et al. [89] and Nißen et al. [90].

Chabot	**Time horizon**	**Primary communication style**	**Intelligence framework**	**Front-end user interface**	**User assistance design**	**Personalization**
MARVIN	Lifelong	Task-oriented & Informative	Text understanding +	Facebook Messenger & Website	Reciprocal	Adaptive

The rapid emergence of generative AI tools such as ChatGPT has transformed the landscape of chatbot development and accelerated the pace of technological innovation. Recognizing the need for agile, iterative evaluation, the first article also introduced a three-phase master protocol using an adaptive platform trial design to guide the co-development, implementation, and evaluation of MARVIN and its domain-specific adaptations (Chapter 4). This methodological framework leverages mixed methods and implementation science models (NASSS [140], TAM [141]) to support rigorous assessment of chatbot interventions in real-world settings. Although the initial development and ethical review process was resource-intensive—taking a full year from inception to REB approval—this master protocol streamlines the coordination of multiple substudies, including the Chapter 6 study, and facilitates the rapid generation of high-quality evidence to inform practice and enhance timely access to advanced interventions.

The second article (Chapter 5) confirmed the real-world feasibility, usability and acceptability of the MARVIN chatbot among PWH using a mixed-methods design. Quantitative findings met or exceeded established thresholds, while qualitative results demonstrated that MARVIN provided convenient, on-demand access to reliable information and support. Like previous findings [44, 45, 105, 142, 143], participants valued the emotional safety, confidentiality, and nonjudgmental tone of the chatbot, which enabled open discussion of sensitive topics. However, several limitations were identified, including MARVIN’s limited topic coverage, lack of memory, and absence of proactive features—findings that also align with earlier research [45, 46, 105, 108, 110]. Further, participants expressed a desire for the chatbot to anticipate user needs and initiate helpful interactions to support ongoing self-management.

Given the central importance of medication adherence in HIV self-management [144], the third article (Chapter 6) focused on designing and developing an LLM-based triage tool to detect ART adherence barriers and stratify associated risk levels from unstructured patient messages. Fine-tuned models outperformed both large-scale general-purpose LLMs (e.g., GPT-4.1) and clinical foundation models (e.g., Clinical-T5 Large), achieving state-of-the-art performance (Macro-F1 = 0.83 for barrier detection, Macro-F1 = 0.80 for risk level prediction). Fairness analyses also showed that the proposed solutions were more reliable than GPT-4.1 and other large-scale open source LLMs when social demographic descriptors were injected into the sentences. However, further work is needed to better characterise the mechanisms behind prediction shifts after gender-related and race-related information is added. Notably, the triage module has not yet been integrated into the MARVIN chatbot for usability testing, as the system is being updated to an LLM-based architecture. Integration and evaluation of the triage module will be a key area of future development.

7.2 Advancing responsible AI through the SHIFT framework

Table 7.2 summarizes the alignment of the three articles with the SHIFT framework.

7.2.1 Sustainable AI

This thesis project embodies sustainable AI by aligning development with clinical needs and system-level feasibility across the chatbot lifecycle. The multidisciplinary team, including engineers, addresses technical capacity gaps identified in prior work [42], promoting long-term sustainability through strong technical leadership. Furthermore, the third article quantified the environmental impact of LLM training and deployment, answering calls for greater attention to sustainability in AI [125, 126]. Fine-tuned LLMs outperformed GPT-4 while consuming 30 times less energy, demonstrating the value of energy-efficient, task-specific models in digital health and offering practical guidance for future deployment. However, as noted by Cerchiari et al., chatbots may not always represent the most cost-effective intervention [109], underscoring the need for future assessment of MARVIN’s economic sustainability.

7.2.2 Inclusive AI

Participatory design methods were integrated throughout this thesis project to support inclusive AI development. PWH and other stakeholders were engaged early and repeatedly, including during the construction of MARVIN’s training database, prototype testing prior to the usability trial, feedback collection for the usability analysis, and co-development of labeling guidelines and annotations for triage algorithm development. This iterative, longitudinal engagement demonstrated that patients and stakeholders are eager to contribute across all stages of the research process to ensure that the chatbot innovation delivers tangible benefits to end users [145]. Their continued involvement helped ensure that MARVIN reflected user needs and enhanced its overall relevance and utility. For instance, a French-language version was developed to meet local language needs, though further optimization is required to improve its performance (Chapter 5). Empowering end-users to shape the final solution and drive open innovation is important for healthcare AI [146] and such engagement was evidenced by the positive results of MARVIN usability study.

Table 7.2 Alignment of thesis articles with the SHIFT framework for responsible AI.

	Article 1 Master protocol & Design and development of MARVIN	**Article 2 Usability study**	**Article 3 Triage model development**
Sustainable AI	Sustaining multidisciplinary team leadership		
	-	-	Assessing environmental footprint
Human-centered AI	Emphasizing the chatbot's assistive role		
	Supporting PWH self-management	Identifying appropriate chatbot use	Enabling first-layer triage support
Inclusive AI	Using a participatory design approach with PWH and stakeholders in developing the chatbot		
	Creating two versions in Canada's official language (English & French)	Identifying user needs	-
Fair AI	Enabling equitable access to information and support for HIV self-management		
	Validating the training data with PWH	-	Assessing triage model fairness
Transparent AI	Communicating with PWH & stakeholders, and scientific community		
	Validating the training data with PWH. Disclosing privacy policy and data security	-	Conducting error analysis for explainability

PWH: people with HIV

Building on this foundation, future iterations of MARVIN will continue to be adapted to address the needs of diverse populations and evolving care contexts. For example, usability study participants emphasized that HIV-related knowledge is relevant not only to PWH but also to all sexually active individuals, indicating a broader potential target audience than initially anticipated. This finding aligns with the status-neutral model of HIV care and prevention proposed by Myers et al. [147], which views HIV testing as a gateway to either treatment or prevention and underscores the equal importance of both pathways. By promoting integrated care regardless of HIV status, this model helps reduce stigma, enhance engagement, and improve equitable access [147]. The focus of existing chatbot interventions on HIV prevention highlights its importance within HIV care [43-45, 105]. Accordingly, future iterations of MARVIN will expand to include individuals at risk of HIV infection as target users. Aging with HIV is another increasingly important consideration. With an estimated 70% of PWH expected to experience comorbidities by 2030 worldwide [148], future adaptations of MARVIN should address the management of age-related and other chronic conditions. A recent review found that most digital health technologies remain condition-specific, increasing the burden on individuals managing multiple health issues [149]. Expanding MARVIN to integrate prevention, treatment, and comorbidity support could help reduce this burden and advance a more inclusive, streamlined approach to AI in HIV care.

7.2.3 Human-centered AI

Importantly, MARVIN is not intended to fully replace healthcare professionals but rather serve as a “digital coworker” that complements and supports them[150]. While many participants expressed trust in the chatbot, some preferred to consult clinicians on complex or sensitive topics, particularly those involving diagnosis or treatment. Consistent with human-centered AI principles, MARVIN is designed to augment care by supporting PWH’s daily self-management, offloading repetitive informational tasks, and eventually serving as a first layer of triage. While handling routine inquiries, the chatbot redirects high-risk or complex queries to appropriate healthcare providers or external resources, thereby facilitating clinicians’ ability to focus on delivering high-value care. Future work should focus on optimizing human-AI collaboration, and MARVIN will continue to be evaluated under the master protocol framework (Chapter 4) to guide its iterative improvement. This will support its seamless integration into real-world clinical workflows through iterative evaluation, design thinking, and implementation science [89]. Only through sustained application of these principles may the MARVIN chatbot help strengthen care delivery, reduce provider burden, and advance a human-centered AI approach to improving overall health system efficiency.

7.2.4 Transparent AI

This work also made many efforts to ensure the transparency of our chatbot intervention. First, explainability is widely recognized as a key requirement for building trustworthy human–machine collaboration in healthcare settings [151]. To support interpretability of the chatbot's responses, all training data were reviewed and validated by PWH and stakeholders, ensuring transparency and accuracy in the chatbot's decision-making process (Chapter 4). Usability testing further confirmed that participants perceived the information provided by MARVIN as reliable (Chapter 5). In the third article, a detailed error analysis of LLM-based triage models was conducted, identifying potential sources of misclassification. For example, the triage model incorrectly classified the question "Are drug holidays safe?" under "Habits and Activities" due to the presence of the word "holiday," when in fact the patient is asking whether it is safe to stop taking medication. Even accurate output does not guarantee the correctness of reasoning—LLMs may arrive at correct answers through flawed logic [152]. Future work may incorporate explainable AI tools such as SHAP or LIME [153] to further enhance algorithm-level explainability. In terms of privacy and security, the master protocol outlines the chatbot's data protection measures, privacy policies, data security protocols, and informed consent procedures for full-scale deployment (Chapter 4). Finally, the complete development process of MARVIN, including identified system limitations, has been documented and communicated to the co-construction Committee and the broader scientific community to support transparency and contribute to broader responsible AI initiatives (See Appendix E for the complete list of related publications and presentations).

7.2.5 Fair AI

As with many of the reviewed chatbots, overall algorithmic fairness remains an area requiring further attention. In this work, fairness assessments were conducted only during the development of the triage model (Chapter 6). No fairness evaluation was applied to the chatbot itself, as it is currently trained on validated data and relies on predefined responses and does not personalize information based on race, gender, or other identity factors. Future chatbot iterations will incorporate LLM-based tools and place greater emphasis on fairness assessments to mitigate potential stigma caused by the models and ensure equitable support for marginalized populations.

7.3 Future work and developments

Building on the foundation established in this thesis, the MARVIN project is entering a new phase of development centered on improving long-term utility, personalization, and better intelligence. As reviewed in Chapter 2, hybrid modeling strategies have become a dominant paradigm for enabling task-specific functionality in AI chatbots. Development efforts are increasingly shifting toward modular, sustainable, and practical architectures where multiple algorithms, models, and tools collaborate in human-like workflows.

This approach is exemplified by multi-agent frameworks, in which multiple LLMs autonomously engage in topic-specific discussions, simulating human collaboration to improve the reasoning capabilities of LLMs. This improves the performance of the final output and enhances the interpretability of its decisions [154]. The third study in this thesis was developed under this modular strategy and will be integrated into MARVIN's workflow in the next iteration. Furthermore, recent research suggests that small language models (SLMs)—with parameter counts in the millions to few billions—offer efficient, cost-effective solutions for specialized downstream tasks within multi-agent frameworks [155]. In many cases, SLMs outperform large, general-purpose LLMs in terms of resource efficiency and deployment feasibility. These findings are consistent with the results of our third study.

To this end, in addition to integrating the adherence triage model, several LLM-based components are under development to enhance MARVIN's functionality. These include a stigma and emotion detection module to support users' mental health, with early pilots showing strong performance in identifying negative affect and suicidal intent [156]. Additional modules include a PIPEDA (The Personal Information Protection and Electronic Documents Act)-compliant message anonymization module [157], a multilingual translation agent, and a memory module enabling short- and long-term memory through user persona profiling. Together, these components aim to address key user needs identified in Chapter 5, including emotional support, data privacy, language accessibility, and coherent, context-aware interactions.

Data scarcity posed another persistent challenge throughout this study and remains a common barrier in healthcare AI, largely due to restrictions around data sharing. While synthetic data augmentation offers a potential solution, its effectiveness remains limited due to lack of diversity and poor quality [158]. Recent efforts to leverage LLMs for automated annotation have shown only moderate agreement with human annotators [159], highlighting the limitations of current unsupervised approaches. In this project, manual annotation was used—the most reliable method for generating high-quality training data, but it required substantial input from patient partners, clinicians, and researchers, making it both resource-intensive. As emphasized in recent literature, access to large-scale, high-quality annotated data is essential for the development of robust healthcare AI models [160]. To enhance scalability and efficiency, future work should investigate semi-automated pipelines, such as active learning and LLM-assisted pre-annotation, while ensuring that data quality and clinical relevance are preserved.

From an implementation perspective, the single-arm, short-term design and small sample size of the Chapter 5 study limit the generalizability of its findings. To strengthen evidence on chatbot implementation, a larger-scale clinical validation will be conducted under the master protocol following the integration of the triage module. However, the current protocol does not include provisions for evaluating clinical or service-related outcomes. Future iterations will incorporate key indicators such as viral load, medication adherence, and health literacy. To further assess generalizability across contexts, multicenter trials are planned, including a collaboration in Ghana through the MARVIN HYPE project, which aims to support HIV prevention and stigma reduction among high-risk young women.

Long-term economic sustainability is a key consideration as MARVIN transitions to broader implementation. To date, the project has been developed and maintained entirely within an academic environment, with limited operational funding and technical resources. Built on open-source, locally deployable models, MARVIN’s design contributes to its potential affordability and feasibility in resource-constrained settings, including public hospitals with limited infrastructure. However, as the chatbot scales and new features are introduced, the project must adopt a more sustainable model to support continued development and deployment. Meanwhile, major commercial players like OpenAI and Google are investing heavily in healthcare AI [161-163], raising the bar for competition and expectations. These shifts make it increasingly difficult for academic, non-commercial initiatives like MARVIN to remain viable. To continue offering a publicly accessible, equity-focused alternative, identifying new funding mechanisms, partnerships, or institutional support will be essential to sustaining MARVIN without compromising its mission.

Over the past four years, AI and chatbot technologies have advanced substantially. From early NLP-supported rule-based systems to the integration of LLMs, MARVIN has progressively adopted many of these innovations to better support PWH. Despite its potential, MARVIN has not yet been implemented in routine clinical practice, reflecting persistent challenges in integration, validation, and adoption within healthcare systems. This journey is far from over. Addressing broader sustainability considerations such as long-term maintenance, cost-effectiveness, and scalability will be essential in future phases to ensure the delivery of a solution with meaningful and lasting value to healthcare [164]. MARVIN may thus serve as a case study in responsible, human-centered healthcare AI, illustrating how impactful digital health innovation arises from aligning technical development with clinical relevance, ethical design, and user-centered development.

CHAPTER 8 CONCLUSION

This thesis aimed to design and develop an AI-based personalized chatbot to support self-management among PWH, guided by principles of participatory design and responsible AI. Through three interrelated manuscripts, it addressed the central question of how such technologies can be designed, developed and implemented to meaningfully augment HIV care in real-world settings.

The first article described the methodological protocol for MARVIN's co-development, implementation, and evaluation as well as that of its family of chatbots. The second article presented the results of real-world usability testing, demonstrating MARVIN's feasibility, usability, and acceptability in supporting daily self-management among PWH. The third article focused on the development and validation of an LLM-based solution to triage ART adherence barriers and associated risk levels, offering a scalable solution for personalized, proactive care. Together, these papers showcase a comprehensive lifecycle approach to the design and development of AI-based healthcare chatbots. Beyond technical innovation, this work highlights broader implications for the development of responsible healthcare AI tools. Co-designed with patients and clinicians, MARVIN exemplifies and concerted effort to achieve sustainability, human-centeredness, inclusiveness, fairness, and transparency in healthcare AI tool development and implementation.

Several limitations should be acknowledged. MARVIN's implementation was assessed in a specific geographic and demographic context, which may limit the generalizability of findings. Moreover, MARVIN remains limited in long-term utility, personalization, and adaptive intelligence. As the system scales and new features are introduced, sustainable development and maintenance strategies will be essential. Future work will focus on integrating new LLM-based modules to enhance personalization (e.g., broader topic coverage, greater language accessibility, and memory support), privacy, and emotional support, while expanding implementation through multicenter trials and clinical outcome evaluation. Additional efforts will aim to improve data efficiency and establish sustainable development models to support long-term deployment.

The potential of this program to improve HIV care and PWH self-management is substantial. MARVIN could empower thousands of individuals to build health-related knowledge, lessen the lifelong self-management burden of PWH, help reduce strain of HIV healthcare providers, and most importantly, foster the overall wellbeing and quality of life of PWH. Finally, this PhD thesis advances the field of healthcare AI for chronic disease management and contributes meaningfully to UNAIDS (United Nation of AIDS)'s goal of ending the HIV epidemic by 2030.

REFERENCES

[1] S. G. Deeks, J. Overbaugh, A. Phillips, and S. Buchbinder, "HIV infection," *Nature reviews Disease primers,* vol. 1, no. 1, pp. 1-22, 2015.

[2] W. H. Organization. "The global health observatory, HIV." https://www.who.int/data/gho/data/themes/hiv-aids#:~:text=Globally%2C%2039.0%20million%20%5B33.1%E2%80%93,at%20the%20end%20of%202022 (accessed 10/07/2025.

[3] P. H. A. o. Canada. "HIV in Canada: 2023 surveillance highlights." https://www.canada.ca/en/public-health/services/publications/diseases-conditions/hiv-2023-surveillance-highlights-infographic.html (accessed.

[4] A. Trickey *et al.*, "Life expectancy after 2015 of adults with HIV on long-term antiretroviral therapy in Europe and North America: a collaborative analysis of cohort studies," *Lancet HIV,* vol. 10, no. 5, pp. e295-e307, May 2023, doi: 10.1016/S2352-3018(23)00028-0.

[5] C. J. Colvin, "HIV/AIDS, chronic diseases and globalisation," *Globalization and health,* vol. 7, no. 1, pp. 1-6, 2011.

[6] UNAIDS, "The GAP Report 2014," Geneva, 2014. [Online]. Available: https://www.unaids.org/sites/default/files/media/images/gap_report_popn_01_plwh_2014 july-sept.pdf

[7] H. R. Holman and K. R. Lorig, "Overcoming barriers to successful aging. Self-management of osteoarthritis," *Western journal of medicine,* vol. 167, no. 4, p. 265, 1997.

[8] H. A. Areri, A. Marshall, and G. Harvey, "Interventions to improve self-management of adults living with HIV on Antiretroviral Therapy: A systematic review," *PLOS ONE,* vol. 15, no. 5, p. e0232709, 2020-05-11 2020, doi: 10.1371/journal.pone.0232709.

[9] S. Russell *et al.*, "Finding Meaning: HIV Self-Management and Wellbeing among People Taking Antiretroviral Therapy in Uganda," *PLOS ONE,* vol. 11, no. 1, p. e0147896, 2016-01-25 2016, doi: 10.1371/journal.pone.0147896.

[10] M. S. Saag, "HIV Infection—Screening, Diagnosis, and Treatment," *New England Journal of Medicine,* vol. 384, no. 22, pp. 2131-2143, 2021.

[11] K. Basavaraj, M. Navya, and R. Rashmi, "Quality of life in HIV/AIDS," *Indian journal of sexually transmitted diseases and AIDS,* vol. 31, no. 2, p. 75, 2010.

[12] M. Franklin, S. Lewis, K. Willis, H. Bourke-Taylor, and L. Smith, "Patients' and healthcare professionals' perceptions of self-management support interactions: systematic review and qualitative synthesis," *Chronic illness,* vol. 14, no. 2, pp. 79-103, 2018.

[13] D. K. Wong and M. K. Cheung, "Online Health Information Seeking and eHealth Literacy Among Patients Attending a Primary Care Clinic in Hong Kong: A Cross-Sectional Survey," *J Med Internet Res,* vol. 21, no. 3, p. e10831, Mar 27 2019, doi: 10.2196/10831.

[14] C. Fortini and J.-B. Daeppen, "How do hospital providers perceive and experience the information-delivery process? A qualitative exploratory study," *PEC innovation,* vol. 3, p. 100222, 2023.

[15] Y. Sun, Y. Zhang, J. Gwizdka, and C. B. Trace, "Consumer Evaluation of the Quality of Online Health Information: Systematic Literature Review of Relevant Criteria and Indicators," *J Med Internet Res*, vol. 21, no. 5, p. e12522, May 2 2019, doi: 10.2196/12522.

[16] H. Miner, A. Fatehi, D. Ring, and J. S. Reichenberg, "Clinician telemedicine perceptions during the COVID-19 pandemic," *Telemedicine and e-Health,* vol. 27, no. 5, pp. 508-512, 2021.

[17] S. Gentili, V. Huang, J. Mamo, and S. Cuschieri, "Chronic Diseases in 2021," 2022.

[18] M. B. Howren and J. S. Gonzalez, "Treatment adherence and illness self-management: introduction to the special issue," *Journal of Behavioral Medicine,* vol. 39, no. 6, pp. 931-934, 2016-12-01 2016, doi: 10.1007/s10865-016-9804-0.

[19] B. L. Paterson, P. Charlton, and S. Richard, "Non-attendance in chronic disease clinics: a matter of non-compliance?," *Journal of Nursing and Healthcare of Chronic Illness,* vol. 2, no. 1, pp. 63-74, 2010-03-01 2010, doi: 10.1111/j.1752-9824.2010.01048.x.

[20] N. Schwennesen, J. E. Henriksen, and I. Willaing, "Patient explanations for non-attendance at type 2 diabetes self-management education: a qualitative study," *Scandinavian Journal of Caring Sciences,* vol. 30, no. 1, pp. 187-192, 2016-03-01 2016, doi: 10.1111/scs.12245.

[21] D. Swendeman, B. L. Ingram, and M. J. Rotheram-Borus, "Common elements in self-management of HIV and other chronic illnesses: an integrative framework," *AIDS Care,* vol. 21, no. 10, pp. 1321-1334, 2009-10-01 2009, doi: 10.1080/09540120902803158.

[22] W. H. Organization, *WHO guideline: recommendations on digital interventions for health system strengthening*. World Health Organization, 2019.

[23] F. Birnbaum, D. Lewis, R. K. Rosen, and M. L. Ranney, "Patient engagement and the design of digital health," *Acad Emerg Med,* vol. 22, no. 6, pp. 754-6, Jun 2015, doi: 10.1111/acem.12692.

[24] H. Huang, M. Xie, Z. Yang, and A. Wang, "Enhancing HIV Cognitive Abilities and Self-Management Through Information Technology-Assisted Interventions: Scoping Review," *J Med Internet Res,* vol. 27, p. e57363, Jan 13 2025, doi: 10.2196/57363.

[25] M. Gogishvili, A. K. Arora, T. M. White, and J. V. Lazarus, "Recommendations for the equitable integration of digital health interventions across the HIV care cascade," *Communications Medicine,* vol. 4, no. 1, 2024-11-03 2024, doi: 10.1038/s43856-024-00645-1.

[26] F. Mhando *et al.*, "Digital Intervention Services to Promote HIV Self-Testing and Linkage to Care Services: A Bibliometric and Content Analysis—Global Trends and Future Directions," *Public Health Reviews,* vol. 45, 2024-02-16 2024, doi: 10.3389/phrs.2024.1606354.

[27] T. Kowatsch *et al.*, "Conversational Agents as Mediating Social Actors in Chronic Disease Management Involving Health Care Professionals, Patients, and Family Members: Multisite Single-Arm Feasibility Study," *J Med Internet Res,* vol. 23, no. 2, p. e25060, Feb 17 2021, doi: 10.2196/25060.

[28] E. King *et al.*, "Mobile Text Messaging to Improve Medication Adherence and Viral Load in a Vulnerable Canadian Population Living With Human Immunodeficiency Virus: A Repeated Measures Study," *Journal of Medical Internet Research,* vol. 19, no. 6, p. e190, 2017-06-01 2017, doi: 10.2196/jmir.6631.

[29] B. Chaix *et al.*, "Assessing the performances of a chatbot to collect real-life data of patients suffering from primary headache disorders," *Digit Health,* vol. 8, p. 20552076221097783, Jan-Dec 2022, doi: 10.1177/20552076221097783.

[30] S. Hauser-Ulrich, H. Kunzli, D. Meier-Peterhans, and T. Kowatsch, "A Smartphone-Based Health Care Chatbot to Promote Self-Management of Chronic Pain (SELMA): Pilot Randomized Controlled Trial," *JMIR Mhealth Uhealth,* vol. 8, no. 4, p. e15806, Apr 3 2020, doi: 10.2196/15806.

[31] J. J. Prochaska *et al.*, "A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study," *J Med Internet Res,* vol. 23, no. 3, p. e24850, Mar 23 2021, doi: 10.2196/24850.

[32] Z. Xing, F. Yu, Y. A. M. Qanir, T. Guan, J. Walker, and L. Song, "Intelligent Conversational Agents in Patient Self-Management: A Systematic Survey Using Multi Data Sources," *Stud Health Technol Inform,* vol. 264, pp. 1813-1814, Aug 21 2019, doi: 10.3233/SHTI190661.

[33] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM,* vol. 9, no. 1, pp. 36-45, 1966, doi: 10.1145/365153.365168.

[34] S. Jang, J. J. Kim, S. J. Kim, J. Hong, S. Kim, and E. Kim, "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,* vol. 150, p. 104440, Jun 2021, doi: 10.1016/j.ijmedinf.2021.104440.

[35] M. C. Klos, M. Escoredo, A. Joerin, V. N. Lemos, M. Rauws, and E. L. Bunge, "Artificial Intelligence-Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial," *JMIR Form Res,* vol. 5, no. 8, p. e20678, Aug 12 2021, doi: 10.2196/20678.

[36] J.-E. Bibault *et al.*, "A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial," *Journal of medical Internet research,* vol. 21, no. 11, p. e15787, 2019.

[37] B. Chaix *et al.*, "When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot," *JMIR cancer,* vol. 5, no. 1, p. e12856, 2019.

[38] S. Nazareth *et al.*, "Hereditary Cancer Risk Using a Genetic Chatbot Before Routine Care Visits," *Obstet Gynecol,* vol. 138, no. 6, pp. 860-870, Dec 1 2021, doi: 10.1097/AOG.0000000000004596.

[39] R. Mash, D. Schouw, and A. E. Fischer, "Evaluating the Implementation of the GREAT4Diabetes WhatsApp Chatbot to Educate People With Type 2 Diabetes During the COVID-19 Pandemic: Convergent Mixed Methods Study," *JMIR Diabetes,* vol. 7, no. 2, p. e37882, Jun 24 2022, doi: 10.2196/37882.

[40] U. U. Rehman, D. J. Chang, Y. Jung, U. Akhtar, M. A. Razzaq, and S. Lee, "Medical Instructed Real-Time Assistant for Patient with Glaucoma and Diabetic Conditions," *Applied Sciences,* vol. 10, no. 7, 2020, doi: 10.3390/app10072216.

[41] A. van Heerden *et al.*, "Chatbots for HIV Prevention and Care: a Narrative Review," *Current HIV/AIDS Reports 2023 20:6,* vol. 20, no. 6, 2023-11-27, doi: 10.1007/s11904-023-00681-x.

[42] W. R. T. Braddock, M. A. Ocasio, W. S. Comulada, J. Mandani, and M. I. Fernandez, "Increasing Participation in a TelePrEP Program for Sexual and Gender Minority Adolescents and Young Adults in Louisiana: Protocol for an SMS Text Messaging–Based Chatbot," *JMIR Research Protocols,* vol. 12, p. e42983, 2023-05-31 2023, doi: 10.2196/42983.

[43] S. Chen *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,* vol. 12, p. e48447, Jun 30 2023, doi: 10.2196/48447.

[44] M. Hui, "Testing the Feasibility and Acceptability of Using an Artificial Intelligence Chatbot to Promote HIV Testing and Pre-Exposure Prophylaxis in Malaysia: Mixed Methods Study," *JMIR Hum Factors 2024;11:e52055 https://humanfactors.jmir.org/2024/1/e52055,* vol. 11, no. 1, 2024-01-26, doi: 10.2196/52055.

[45] P. Massa *et al.*, "A Transgender Chatbot (Amanda Selfie) to Create Pre-exposure Prophylaxis Demand Among Adolescents in Brazil: Assessment of Acceptability, Functionality, Usability, and Results," *Journal of Medical Internet Research,* vol. 25, p. e41881, 2023-06-23 2023, doi: 10.2196/41881.

[46] A. van Heerden, X. Ntinga, and K. Vilakazi, "The potential of conversational agents to provide a rapid HIV counseling and testing services," in *2017 international conference on the frontiers and advances in data science (FADS)*, 2017: IEEE, pp. 80-85.

[47] E. A. Yam *et al.*, "Developing and Testing a Chatbot to Integrate HIV Education Into Family Planning Clinic Waiting Areas in Lusaka, Zambia," *Global Health: Science and Practice,* vol. 10, no. 5, p. e2100721, 2022-10-31 2022, doi: 10.9745/ghsp-d-21-00721.

[48] S. Dick, "Artificial intelligence," 2019.

[49] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[50] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-44, May 28 2015, doi: 10.1038/nature14539.

[51] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. 2025.

[52] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[54] G. Bezko. "Understand AI, ML & Co in Contact Centers: Definitions & Explanations." https://blog.miarec.com/contact-centers-ai-definition (accessed 29, July, 2025).

[55] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," *NPJ Digit Med,* vol. 6, no. 1, p. 135, Jul 29 2023, doi: 10.1038/s41746-023-00879-8.

[56] E. Lehman *et al.*, "Do we still need clinical language models?," in *Conference on health, inference, and learning,* 2023: PMLR, pp. 578-597.

[57] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361,* 2020.

[58] OpenAI. "Introducing ChatGPT." https://openai.com/index/chatgpt/ (accessed 12/07, 2025).

[59] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774,* 2023.

[60] A. Liu *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437,* 2024.

[61] Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," 2024.

[62] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur),* vol. 53, no. 3, pp. 1-34, 2020.

[63] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit Health,* vol. 2, no. 2, p. e0000198, Feb 2023, doi: 10.1371/journal.pdig.0000198.

[64] J. L. Marcus, W. C. Sewell, L. B. Balzer, and D. S. Krakower, "Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic," *Current HIV/AIDS Reports,* vol. 17, no. 3, pp. 171-179, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7260108/pdf/nihms-1589053.pdf.

[65] R. Jin and L. Zhang, "AI applications in HIV research: advances and future directions," *Frontiers in Microbiology,* vol. 16, 2025-02-20 2025, doi: 10.3389/fmicb.2025.1541942.

[66] J. Ratevosian *et al.*, "Reimagining HIV prevention with artificial intelligence," *Lancet HIV,* Jun 11 2025, doi: 10.1016/S2352-3018(25)00158-4.

[67] T. Nadarzynski, A. Lunt, N. Knights, J. Bayley, and C. Llewellyn, "“But can chatbots understand sex?” Attitudes towards artificial intelligence chatbots amongst sexual and reproductive health professionals: An exploratory mixed-methods study," *International Journal of STD & AIDS,* vol. 34, no. 11, pp. 809-816, 2023-10-01 2023, doi: 10.1177/09564624231180777.

[68] T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, "Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study," *Digit Health,* vol. 5, p. 2055207619871808, Jan-Dec 2019, doi: 10.1177/2055207619871808.

[69] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, Feb 15 2020, doi: 10.1093/bioinformatics/btz682.

[70] E. Hernandez *et al.*, "Do we still need clinical language models?," in *Conference on Health, Inference, and Learning*, 2023: PMLR, pp. 578-597.

[71] M. Guevara *et al.*, "Large Language Models to Identify Social Determinants of Health in Electronic Health Records," *arXiv preprint arXiv:2308.06354*, 2023.

[72] M. M. van Buchem *et al.*, "Applying natural language processing to patient messages to identify depression concerns in cancer patients," *Journal of the American Medical Informatics Association,* vol. 31, no. 10, pp. 2255-2262, 2024.

[73] G. S. Mosnaim, H. Stempel, D. Van Sickle, and D. A. Stempel, "The adoption and implementation of digital health care in the post–COVID-19 era," *The Journal of Allergy and Clinical Immunology. in Practice,* vol. 8, no. 8, p. 2484, 2020.

[74] F. Petracca, O. Ciani, M. Cucciniello, and R. Tarricone, "Harnessing digital health technologies during and after the COVID-19 pandemic: context matters," *Journal of medical Internet research,* vol. 22, no. 12, p. e21815, 2020.

[75] S. Shourya *et al.*, "A Remote Intervention Based on mHealth and Community Health Workers for Antiretroviral Therapy Adherence in People With HIV: Pilot Randomized Controlled Trial," *JMIR Formative Research,* vol. 9, p. e67997, 2025-04-02 2025, doi: 10.2196/67997.

[76] R. Schnall *et al.*, "Efficacy, Use, and Usability of the VIP-HANA App for Symptom Self-management in PLWH with HANA Conditions," *AIDS and Behavior,* vol. 25, no. 6, pp. 1699-1710, 2021-06-01 2021, doi: 10.1007/s10461-020-03096-6.

[77] F. Gárate *et al.*, "EmERGE mHealth Platform: Implementation and Technical Evaluation of a Digital Supported Pathway of Care for Medically Stable HIV," *International Journal of Environmental Research and Public Health,* vol. 18, no. 6, p. 3156, 2021-03-18 2021, doi: 10.3390/ijerph18063156.

[78] L. Hightow-Weidman, K. E. Muessig, J. R. Egger, A. Vecchio, and A. Platt, "Epic Allies: A Gamified Mobile App to Improve Engagement in HIV Care and Antiretroviral Adherence among Young Men Who have Sex with Men," *AIDS and Behavior,* vol. 25, no. 8, pp. 2599-2617, 2021-08-01 2021, doi: 10.1007/s10461-021-03222-y.

[79] T. Chenneville, H. Drake, K. Gabbidon, C. Rodriguez, and L. Hightow-Weidman, "Bijou: Engaging Young MSM in HIV Care Using a Mobile Health Strategy," *Journal of the International Association of Providers of AIDS Care (JIAPAC),* vol. 20, p. 232595822110308, 2021-01-01 2021, doi: 10.1177/23259582211030805.

[80] G. S. Kim *et al.*, "Three cycles of mobile app design to improve HIV self-management: A development and evaluation study," *DIGITAL HEALTH,* vol. 10, p. 20552076241249294, 2024, doi: 10.1177/20552076241249294.

[81] M.-S. Shim, S. Kim, M. Choi, J. Y. Choi, C. G. Park, and G. S. Kim, "Developing an app-based self-management program for people living with HIV: a randomized controlled pilot

study during the COVID-19 pandemic," *Scientific Reports,* vol. 12, no. 1, 2022-11-12 2022, doi: 10.1038/s41598-022-19238-w.

[82] G. S. Kim *et al.*, "Providing 2 Types of mHealth Interventions to Support Self-Management Among People Living With HIV: Randomized Clinical Trial," *JMIR mHealth and uHealth,* vol. 13, pp. e60905-e60905, 2025-05-29 2025, doi: 10.2196/60905.

[83] H. Huang, M. Xie, Z. Yang, and A. Wang, "Enhancing HIV Cognitive Abilities and Self-Management Through Information Technology–Assisted Interventions: Scoping Review," *Journal of Medical Internet Research,* vol. 27, p. e57363, 2025-01-13 2025, doi: 10.2196/57363.

[84] D. Krebs *et al.*, "Telehealth Interventions to Improve HIV Care Continuum Outcomes: A Narrative Review," *AIDS Patient Care and STDs,* vol. 39, no. 4, pp. 129-140, 2025, doi: 10.1089/apc.2024.0237.

[85] E. Proctor *et al.*, "Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda," *Adm Policy Ment Health,* vol. 38, no. 2, pp. 65-76, Mar 2011, doi: 10.1007/s10488-010-0319-7.

[86] L. Weyers, T. Crowley, and L. Tokwe, "Digital technology for HIV self-management in low- and middle-income countries: a scoping review of adolescents' preferences," *AIDS Care,* vol. 36, no. 12, pp. 1805-1814, 2024-12-01 2024, doi: 10.1080/09540121.2024.2383868.

[87] L. B. Whiteley, E. M. Olsen, K. K. Haubrick, E. Odoom, N. Tarantino, and L. K. Brown, "A Review of Interventions to Enhance HIV Medication Adherence," *Current HIV/AIDS Reports,* vol. 18, no. 5, pp. 443-457, 2021-10-01 2021, doi: 10.1007/s11904-021-00568-9.

[88] V. Cooper, J. Clatworthy, J. Whetham, and E. Consortium, "mHealth Interventions To Support Self-Management In HIV: A Systematic Review," *The Open AIDS Journal,* vol. 11, no. 1, pp. 119-132, 2017-11-21 2017, doi: 10.2174/1874613601711010119.

[89] R. C. Li, S. M. Asch, and N. H. Shah, "Developing a delivery science for artificial intelligence in healthcare," *npj Digital Medicine,* vol. 3, no. 1, 2020-08-21 2020, doi: 10.1038/s41746-020-00318-y.

[90] A. Janssen, J. Passlick, D. Rodríguez Cardona, and M. H. Breitner, "Virtual Assistance in Any Context," *Business & Information Systems Engineering,* vol. 62, no. 3, pp. 211-225, 2020, doi: 10.1007/s12599-020-00644-1.

[91] M. Nißen *et al.*, "See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons," *Computers in Human Behavior,* vol. 127, 2022, doi: 10.1016/j.chb.2021.107043.

[92] H. Siala and Y. Wang, "SHIFTing artificial intelligence to be responsible in healthcare: A systematic review," *Soc Sci Med,* vol. 296, p. 114782, Mar 2022, doi: 10.1016/j.socscimed.2022.114782.

[93] J. C. C. Kwong, G. C. Nickel, S. C. Y. Wang, and J. C. Kvedar, "Integrating artificial intelligence into healthcare systems: more than just the algorithm," *npj Digital Medicine,* vol. 7, no. 1, 2024-03-01 2024, doi: 10.1038/s41746-024-01066-z.

[94] L. Bitomsky, M. Nißen, and T. Kowatsch, "Equity by Design Principles for Digital Health Interventions," Springer Science and Business Media LLC, 2025-05-30, 2025.

[95] M. Nuruzzaman and O. K. Hussain, "A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks," presented at the 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), 2018.

[96] J. Grudin and R. Jacques, "Chatbots, Humbots, and the Quest for Artificial General Intelligence," presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[97] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A Survey on Conversational Agents/Chatbots Classification and Design Techniques," in *Web, Artificial Intelligence and Network Applications*, (Advances in Intelligent Systems and Computing, 2019, ch. Chapter 93, pp. 946-956.

[98] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications,* vol. 2, 2020, doi: 10.1016/j.mlwa.2020.100006.

[99] T. W. Bank. "World Bank Country and Lending Groups." https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (accessed July 20, 2025).

[100] M. L. Peng *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,* vol. 6, no. 10, p. e42055, Oct 6 2022, doi: 10.2196/42055.

[101] W. S. Comulada *et al.*, "A necessary conversation to develop chatbots for HIV studies: qualitative findings from research staff, community advisory board members, and study participants," *AIDS Care,* vol. 36, no. 4, pp. 463-471, Apr 2024, doi: 10.1080/09540121.2023.2216926.

[102] M. Laymouna *et al.*, "Needs-Assessment for an Artificial Intelligence-Based Chatbot for Pharmacists in HIV Care: Results from a Knowledge-Attitudes-Practices Survey," *Healthcare (Basel),* vol. 12, no. 16, Aug 20 2024, doi: 10.3390/healthcare12161661.

[103] J. J. N. Ndenkeh *et al.*, "Formative evaluation of the acceptance of HIV prevention Artificial Intelligence chatbots by Black gay, bisexual, and other men who have sex with men in the Southern United States: Focus group study," *PLOS Digital Health,* vol. 4, no. 6, p. e0000891, 2025-06-04 2025, doi: 10.1371/journal.pdig.0000891.

[104] J. C. Moreno, V. Sánchez-Anguix, J. M. Alberola, V. Julián, and V. Botti, "A Conversational Agent for Medical Disclosure of Sexually Transmitted Infections," in *Lecture Notes in Computer Science*: Springer International Publishing, 2022, pp. 431-442.

[105] X. Ntinga, F. Musiello, A. K. Keter, R. Barnabas, and A. Van Heerden, "The Feasibility and Acceptability of an mHealth Conversational Agent Designed to Support HIV Self-testing in South Africa: Cross-sectional Study," *Journal of Medical Internet Research,* vol. 24, no. 12, p. e39816, 2022-12-12 2022, doi: 10.2196/39816.

[106] T. Nadarzynski *et al.*, "The impact of Chatbot-Assisted Self Assessment (CASA) on intentions for sexual health screening in people from minoritised ethnic groups at risk of

sexually transmitted infections," *Sexual Health,* vol. 21, no. 4, 2024-07-25 2024, doi: 10.1071/sh24058.

[107] C. Zhang, M. Wharton, and Y. Liu, "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," (in English), *JMIR Res Protoc,* Protocol vol. 13, p. e59975, 2024, doi: 10.2196/59975.

[108] J. C. Moreno, V. Sánchez-Anguix, J. M. Alberola, V. Julián, and V. Botti, "An intelligent conversational agent for educating the general public about HIV," *Neurocomputing,* vol. 563, 2024, doi: 10.1016/j.neucom.2023.126902.

[109] N. Cerchiari, A. Grangeiro, P. Massa, A. C. Santos, and P. C. De Soárez, "PrEP demand creation strategies for adolescents at increased risk of HIV infection in São Paulo, Brazil: a cost-consequence analysis," *BMC Health Services Research,* vol. 25, no. 1, 2025-02-13 2025, doi: 10.1186/s12913-025-12398-1.

[110] S. Mathur *et al.*, ""Let's chat!" Piloting a digital chatbot for HIV prevention among cisgender women and transgender men in Nigeria," *AIDS Care,* vol. 37, no. 6, pp. 1015-1025, Jun 2025, doi: 10.1080/09540121.2025.2470318.

[111] J. T. Galea *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,* vol. 13, p. e55559, May 7 2024, doi: 10.2196/55559.

[112] N. Rupani *et al.*, ""Like Someone Is Paying Attention to You, Listening to You, and Guiding You": Acceptability of a Mental Health Chatbot Among Caregivers of Adolescents Living With HIV," *J Int Assoc Provid AIDS Care,* vol. 24, p. 23259582251327911, Jan-Dec 2025, doi: 10.1177/23259582251327911.

[113] G. Rouleau *et al.*, "Mapping theories, models and frameworks to implement or evaluate digital health interventions: A scoping review. (Preprint)," *Journal of Medical Internet Research*, 2023-07-26 2023, doi: 10.2196/51098.

[114] P. Nilsen, "Making sense of implementation theories, models and frameworks," *Implementation Science,* vol. 10, no. 1, 2015-12-01 2015, doi: 10.1186/s13012-015-0242-0.

[115] M. Hagglund and I. Scandurra, "Usability of the Swedish Accessible Electronic Health Record: Qualitative Survey Study," *JMIR Hum Factors,* vol. 9, no. 2, p. e37192, Jun 23 2022, doi: 10.2196/37192.

[116] I. O. f. Standardization, "Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts," 2018. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

[117] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion,* vol. 58, pp. 82-115, 2020-06-01 2020, doi: 10.1016/j.inffus.2019.12.012.

[118] R. Tatoud, R. B. Jones, K. Dong, T. Ndung'u, S. Deeks, and C. T. Tiemessen, "Advancing HIV cure research in low- and middle-income countries requires empowerment of the next generation of scientists," *J Virus Erad,* vol. 10, no. 1, p. 100364, Mar 2024, doi: 10.1016/j.jve.2024.100364.

[119] E. S. Kay, D. S. Batey, and M. J. Mugavero, "The HIV treatment cascade and care continuum: updates, goals, and recommendations for the future," *AIDS Research and Therapy,* vol. 13, no. 1, 2016-12-01 2016, doi: 10.1186/s12981-016-0120-0.

[120] T. Schachner, R. Keller, and V. W. F, "Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review," *J Med Internet Res,* vol. 22, no. 9, p. e20701, Sep 14 2020, doi: 10.2196/20701.

[121] I. Leite, C. Martinho, and A. Paiva, "Social Robots for Long-Term Interaction: A Survey," *International Journal of Social Robotics,* vol. 5, no. 2, pp. 291-308, 2013, doi: 10.1007/s12369-013-0178-y.

[122] K. Baraka, P. Alves-Oliveira, and T. Ribeiro, "An extended framework for characterizing social robots," in *Human-robot interaction,* 2020: Springer, pp. 21-64.

[123] M. A. Handley, A. Gorukanti, and A. Cattamanchi, "Strategies for implementing implementation science: a methodological overview," *Emergency Medicine Journal,* vol. 33, no. 9, pp. 660-664, 2016.

[124] C. A. McKim, "The value of mixed methods research: A mixed methods study," *Journal of mixed methods research,* vol. 11, no. 2, pp. 202-222, 2017.

[125] A. Fiske, I. M. Radhuber, T. Willem, A. Buyx, L. A. Celi, and S. McLennan, "Climate change and health: the next challenge of ethical AI," *The Lancet Global Health,* vol. 13, no. 7, pp. e1314-e1320, 2025, doi: 10.1016/s2214-109x(25)00124-x.

[126] E. Osmanlliu *et al.*, "The Urgency of Environmentally Sustainable and Socially Just Deployment of Artificial Intelligence in Health Care," *Catalyst non-issue content,* vol. 6, no. 4, p. CAT.24.0501, 2025, doi: doi:10.1056/CAT.24.0501.

[127] T. Zack *et al.*, "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study," *Lancet Digit Health,* vol. 6, no. 1, pp. e12-e22, Jan 2024, doi: 10.1016/S2589-7500(23)00225-X.

[128] D. Busch *et al.*, "A blueprint for large language model-augmented telehealth for HIV mitigation in Indonesia: A scoping review of a novel therapeutic modality," *Health Informatics Journal,* vol. 31, no. 1, p. 14604582251315595, 2025, doi: 10.1177/14604582251315595.

[129] M. Senekane, "BERTina: A BERT-powered HIV/AIDS Question Answering Tool," presented at the 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), 2021.

[130] M. C. Y. Koh, J. N. Ngiam, J. Yong, P. A. Tambyah, and S. Archuleta, "The role of an artificial intelligence model in antiretroviral therapy counselling and advice for people living with HIV," *HIV Medicine,* 2024-01-02 2024, doi: 10.1111/hiv.13604.

[131] A. De Vito *et al.*, "Assessing ChatGPT’s Potential in HIV Prevention Communication: A Comprehensive Evaluation of Accuracy, Completeness, and Inclusivity," *AIDS and*

Behavior, vol. 28, no. 8, pp. 2746-2754, 2024-08-01 2024, doi: 10.1007/s10461-024-04391-2.

[132] P. Rust, J. Frings, S. Meister, and L. Fehring, "Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations," *Communications Medicine,* vol. 5, no. 1, 2025-05-29 2025, doi: 10.1038/s43856-025-00927-2.

[133] K. Du *et al.*, "Comparing Artificial Intelligence–Generated and Clinician-Created Personalized Self-Management Guidance for Patients With Knee Osteoarthritis: Blinded Observational Study," *Journal of Medical Internet Research,* vol. 27, p. e67830, 2025-05-07 2025, doi: 10.2196/67830.

[134] C. Li *et al.*, "Unveiling the Potential of Large Language Models in Transforming Chronic Disease Management: Mixed Methods Systematic Review," *Journal of Medical Internet Research,* vol. 27, p. e70535, 2025-04-16 2025, doi: 10.2196/70535.

[135] T. V. Afanasieva, P. V. Platov, A. V. Komolov, and A. V. Kuzlyakin, "Leveraging ChatGPT and Long Short-Term Memory in Recommender Algorithm for Self-Management of Cardiovascular Risk Factors," *Mathematics,* vol. 12, no. 16, p. 2582, 2024-08-21 2024, doi: 10.3390/math12162582.

[136] J. Li *et al.*, "Integrated image-based deep learning and language models for primary diabetes care," *Nature Medicine,* vol. 30, no. 10, pp. 2886-2896, 2024-10-01 2024, doi: 10.1038/s41591-024-03139-8.

[137] C. Spinuzzi, "The methodology of participatory design," *Technical communication,* vol. 52, no. 2, pp. 163-174, 2005.

[138] K. Engler, D. Lessard, I. Toupin, A. Lènàrt, and B. Lebouché, "Engaging stakeholders into an electronic patient-reported outcome development study: On making an HIV-specific e-PRO patient-centered," *Health Policy and Technology,* vol. 6, no. 1, pp. 59-66, 2017, doi: 10.1016/j.hlpt.2016.11.002.

[139] K. Engler, A. Lènàrt, D. Lessard, I. Toupin, and B. Lebouché, "Barriers to antiretroviral therapy adherence in developed countries: a qualitative synthesis to develop a conceptual framework for a new patient-reported outcome measure," *AIDS Care,* vol. 30, no. sup1, pp. 17-28, 2018-12-14 2018, doi: 10.1080/09540121.2018.1469725.

[140] T. Greenhalgh *et al.*, "Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies," *J Med Internet Res,* vol. 19, no. 11, p. e367, Nov 1 2017, doi: 10.2196/jmir.8775.

[141] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," *Management Science,* vol. 46, no. 2, pp. 186-204, 2000, doi: 10.1287/mnsc.46.2.186.11926.

[142] N. L. Bragazzi, A. Crapanzano, M. Converti, R. Zerbetto, and R. Khamisy-Farah, "The Impact of Generative Conversational Artificial Intelligence on the Lesbian, Gay, Bisexual, Transgender, and Queer Community: Scoping Review," *Journal of Medical Internet Research,* vol. 25, p. e52091, 2023-12-06 2023, doi: 10.2196/52091.

[143] G. Sanabria *et al.*, "“A Great Way to Start the Conversation”: Evidence for the Use of an Adolescent Mental Health Chatbot Navigator for Youth at Risk of HIV and Other STIs," *Journal of Technology in Behavioral Science 2023 8:4,* vol. 8, no. 4, 2023-05-11, doi: 10.1007/s41347-023-00315-4.

[144] K. L. Schaecher, "The importance of treatment adherence in HIV," (in eng), *Am J Manag Care,* vol. 19, no. 12 Suppl, pp. s231-7, 2013/09// 2013. [Online]. Available: http://europepmc.org/abstract/MED/24495293.

[145] C. Rauschenberg *et al.*, "Living lab AI4U - artificial intelligence for personalized digital mental health promotion and prevention in youth," *European Journal of Public Health,* vol. 31, no. Supplement_3, 2021-10-20 2021, doi: 10.1093/eurpub/ckab164.746.

[146] M. Zanetti, G. Nollo, and M. De Cecco, "Living Lab, an innovative approach in healthcare," *IEEE Smart Cities Initiative-Trento white paper,* pp. 1-5, 2019.

[147] J. E. Myers *et al.*, "Redefining Prevention and Care: A Status-Neutral Approach to HIV," *Open Forum Infectious Diseases,* vol. 5, no. 6, 2018, doi: 10.1093/ofid/ofy097.

[148] K. N. Althoff *et al.*, "The forecasted prevalence of comorbidities and multimorbidity in people with HIV in the United States through the year 2030: A modeling study," *PLOS Medicine,* vol. 21, no. 1, p. e1004325, 2024-01-12 2024, doi: 10.1371/journal.pmed.1004325.

[149] N. T. T. Phi, V. M. Montori, M. Kunneman, P. Ravaud, and V.-T. Tran, "Cumulative Burden of Digital Health Technologies for Patients With Multimorbidity," *JAMA Network Open,* vol. 8, no. 4, p. e257288, 2025-04-25 2025, doi: 10.1001/jamanetworkopen.2025.7288.

[150] K. Sowa and A. Przegalinska, "Digital coworker: human-AI collaboration in work environment, on the example of virtual assistants for management professions," in *Collaborative innovation networks conference of Digital Transformation of Collaboration,* 2019: Springer, pp. 179-201.

[151] N. Bienefeld *et al.*, "Solving the explainable AI conundrum by bridging clinicians’ needs and developers’ goals," *npj Digital Medicine,* vol. 6, no. 1, 2023-05-22 2023, doi: 10.1038/s41746-023-00837-4.

[152] Y. Hao *et al.*, "MedPAIR: Measuring Physicians and AI Relevance Alignment in Medical Question Answering," *arXiv preprint arXiv:2505.24040,* 2025.

[153] A. M. Salih *et al.*, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Advanced Intelligent Systems,* vol. 7, no. 1, p. 2400304, 2025.

[154] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Autonomous agents and multi-agent systems,* vol. 33, pp. 673-705, 2019.

[155] P. Belcak *et al.*, "Small Language Models are the Future of Agentic AI," *arXiv preprint arXiv:2506.02153,* 2025.

[156] D. S. Villanueva Guzman, Y. Ma, S. Achiche, K. Engler, D. Lessard, and B. Lebouche, "An AI-Powered Preventive Intervention for Stigma and Suicidal Ideation in HIV Self-Management," presented at the Conference on Retroviruses and Opportunistic Infections

(CROI) 2025, San Francisco, CA, March 9-12, 2025, 2025, Oral. [Online]. Available: https://www.croiconference.org/abstract/1635-2025-2/.

[157] O. o. t. P. C. o. Canada. "The Personal Information Protection and Electronic Documents Act (PIPEDA)." https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/ (accessed Oct 07, 2024.

[158] S. Hao *et al.*, "Synthetic data in AI: Challenges, applications, and ethical implications," *arXiv preprint arXiv:2401.01629,* 2024.

[159] M. U. U. Haq, D. Rigoni, and A. Sperduti, "Llms as data annotators: How close are we to human performance," *arXiv preprint arXiv:2504.15022,* 2025.

[160] R. D. Riley *et al.*, "Importance of sample size on the quality and utility of AI-based prediction models for healthcare," *The Lancet Digital Health*, 2025.

[161] R. K. Arora *et al.*, "Healthbench: Evaluating large language models towards improved human health," *arXiv preprint arXiv:2505.08775,* 2025.

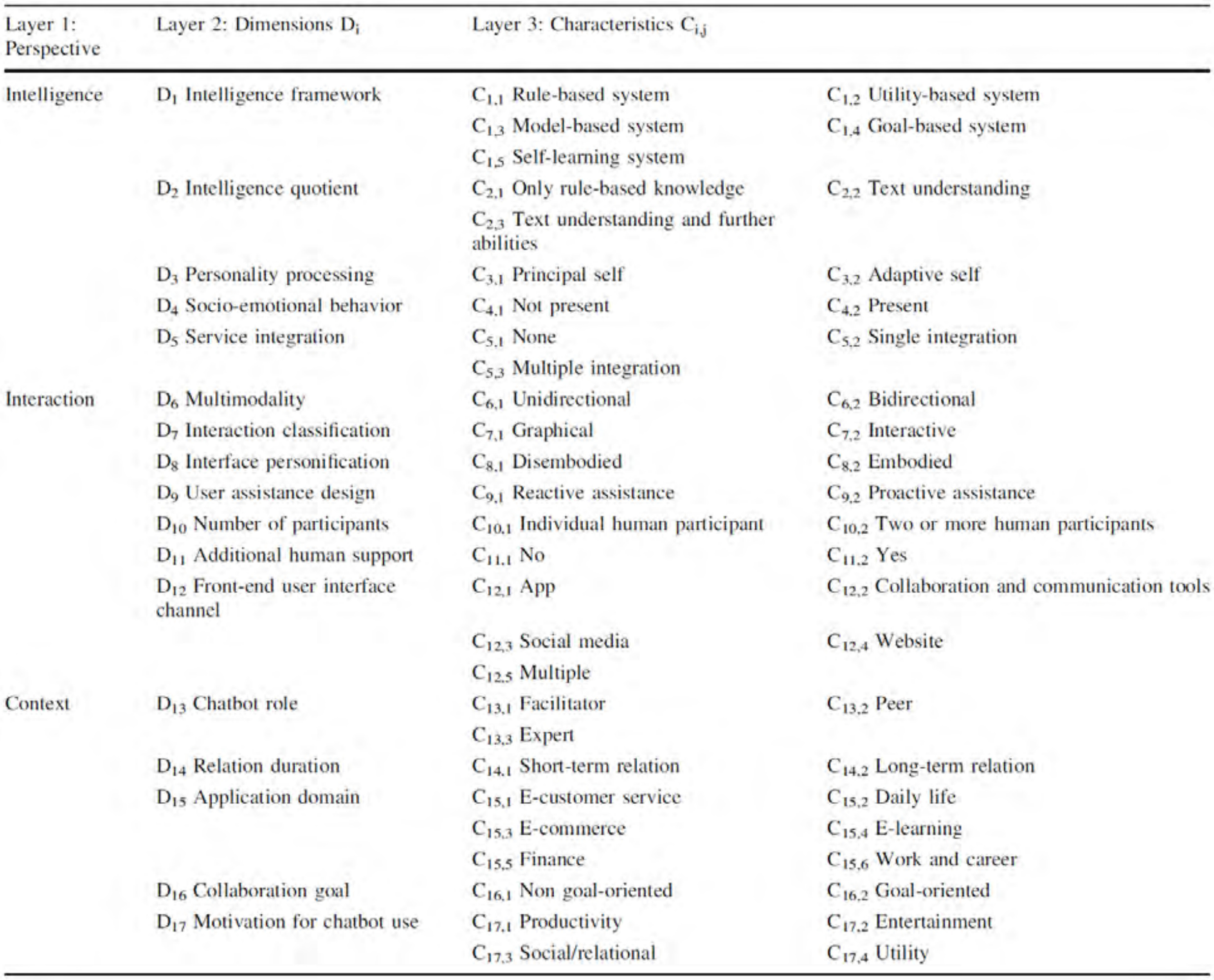
[162] R. Korom *et al.*, "AI-based Clinical Decision Support for Primary Care: A Real-World Study," *arXiv preprint arXiv:2507.16947,* 2025.

[163] A. Sellergren *et al.*, "MedGemma Technical Report," *arXiv preprint arXiv:2507.05201,* 2025.

[164] M. Hossain, S. Leminen, and M. Westerlund, "A systematic review of living lab literature," *Journal of Cleaner Production*, vol. 213, pp. 976-988, 2019-03-01 2019, doi: 10.1016/j.jclepro.2018.12.257.

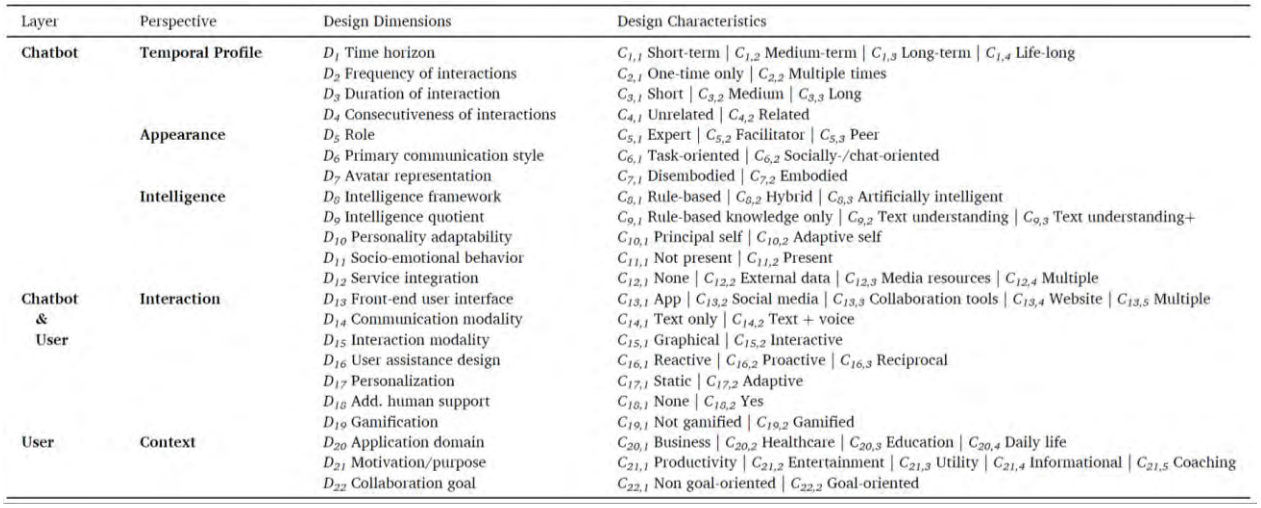
APPENDIX A SUPPORTING MATERIALS FOR CHAPTER 2

Table 8.1 Final taxonomy of design elements for chatbots, extracted from [90].



Layer 1: Perspective	Layer 2: Dimensions D_i	Layer 3: Characteristics $C_{i,j}$	
Intelligence	D_1 Intelligence framework	$C_{1,1}$ Rule-based system	$C_{1,2}$ Utility-based system
		$C_{1,3}$ Model-based system	$C_{1,4}$ Goal-based system
		$C_{1,5}$ Self-learning system	
	D_2 Intelligence quotient	$C_{2,1}$ Only rule-based knowledge	$C_{2,2}$ Text understanding
		$C_{2,3}$ Text understanding and further abilities	
	D_3 Personality processing	$C_{3,1}$ Principal self	$C_{3,2}$ Adaptive self
	D_4 Socio-emotional behavior	$C_{4,1}$ Not present	$C_{4,2}$ Present
	D_5 Service integration	$C_{5,1}$ None	$C_{5,2}$ Single integration
		$C_{5,3}$ Multiple integration	
Interaction	D_6 Multimodality	$C_{6,1}$ Unidirectional	$C_{6,2}$ Bidirectional
	D_7 Interaction classification	$C_{7,1}$ Graphical	$C_{7,2}$ Interactive
	D_8 Interface personification	$C_{8,1}$ Disembodied	$C_{8,2}$ Embodied
	D_9 User assistance design	$C_{9,1}$ Reactive assistance	$C_{9,2}$ Proactive assistance
	D_{10} Number of participants	$C_{10,1}$ Individual human participant	$C_{10,2}$ Two or more human participants
	D_{11} Additional human support	$C_{11,1}$ No	$C_{11,2}$ Yes
	D_{12} Front-end user interface channel	$C_{12,1}$ App	$C_{12,2}$ Collaboration and communication tools
		$C_{12,3}$ Social media	$C_{12,4}$ Website
		$C_{12,5}$ Multiple	
Context	D_{13} Chatbot role	$C_{13,1}$ Facilitator	$C_{13,2}$ Peer
		$C_{13,3}$ Expert	
	D_{14} Relation duration	$C_{14,1}$ Short-term relation	$C_{14,2}$ Long-term relation
	D_{15} Application domain	$C_{15,1}$ E-customer service	$C_{15,2}$ Daily life
		$C_{15,3}$ E-commerce	$C_{15,4}$ E-learning
		$C_{15,5}$ Finance	$C_{15,6}$ Work and career
	D_{16} Collaboration goal	$C_{16,1}$ Non goal-oriented	$C_{16,2}$ Goal-oriented
	D_{17} Motivation for chatbot use	$C_{17,1}$ Productivity	$C_{17,2}$ Entertainment
		$C_{17,3}$ Social/relational	$C_{17,4}$ Utility

Table 8.2 Design taxonomy for chatbots with different temporal profiles, extracted from [91].



Layer	Perspective	Design Dimensions	Design Characteristics
Chatbot	**Temporal Profile**	D_1 Time horizon	$C_{1,1}$ Short-term | $C_{1,2}$ Medium-term | $C_{1,3}$ Long-term | $C_{1,4}$ Life-long
		D_2 Frequency of interactions	$C_{2,1}$ One-time only | $C_{2,2}$ Multiple times
		D_3 Duration of interaction	$C_{3,1}$ Short | $C_{3,2}$ Medium | $C_{3,3}$ Long
		D_4 Consecutiveness of interactions	$C_{4,1}$ Unrelated | $C_{4,2}$ Related
	Appearance	D_5 Role	$C_{5,1}$ Expert | $C_{5,2}$ Facilitator | $C_{5,3}$ Peer
		D_6 Primary communication style	$C_{6,1}$ Task-oriented | $C_{6,2}$ Socially-/chat-oriented
		D_7 Avatar representation	$C_{7,1}$ Disembodied | $C_{7,2}$ Embodied
	Intelligence	D_8 Intelligence framework	$C_{8,1}$ Rule-based | $C_{8,2}$ Hybrid | $C_{8,3}$ Artificially intelligent
		D_9 Intelligence quotient	$C_{9,1}$ Rule-based knowledge only | $C_{9,2}$ Text understanding | $C_{9,3}$ Text understanding+
		D_{10} Personality adaptability	$C_{10,1}$ Principal self | $C_{10,2}$ Adaptive self
		D_{11} Socio-emotional behavior	$C_{11,1}$ Not present | $C_{11,2}$ Present
		D_{12} Service integration	$C_{12,1}$ None | $C_{12,2}$ External data | $C_{12,3}$ Media resources | $C_{12,4}$ Multiple
Chatbot & User	**Interaction**	D_{13} Front-end user interface	$C_{13,1}$ App | $C_{13,2}$ Social media | $C_{13,3}$ Collaboration tools | $C_{13,4}$ Website | $C_{13,5}$ Multiple
		D_{14} Communication modality	$C_{14,1}$ Text only | $C_{14,2}$ Text + voice
		D_{15} Interaction modality	$C_{15,1}$ Graphical | $C_{15,2}$ Interactive
		D_{16} User assistance design	$C_{16,1}$ Reactive | $C_{16,2}$ Proactive | $C_{16,3}$ Reciprocal
		D_{17} Personalization	$C_{17,1}$ Static | $C_{17,2}$ Adaptive
		D_{18} Add. human support	$C_{18,1}$ None | $C_{18,2}$ Yes
		D_{19} Gamification	$C_{19,1}$ Not gamified | $C_{19,2}$ Gamified
User	**Context**	D_{20} Application domain	$C_{20,1}$ Business | $C_{20,2}$ Healthcare | $C_{20,3}$ Education | $C_{20,4}$ Daily life
		D_{21} Motivation/purpose	$C_{21,1}$ Productivity | $C_{21,2}$ Entertainment | $C_{21,3}$ Utility | $C_{21,4}$ Informational | $C_{21,5}$ Coaching
		D_{22} Collaboration goal	$C_{22,1}$ Non goal-oriented | $C_{22,2}$ Goal-oriented

APPENDIX B SUPPORTING MATERIALS FOR CHAPTER 4

Multimedia Appendix 1: CONSORT (Consolidated Standards of Reporting Trials) extension for pilot and feasibility trials checklist.

[Link]

Multimedia Appendix 2: CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) V 1.6.1 checklist.

[Link]

Multimedia Appendix 3: CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) checklist.

[Link]

Multimedia Appendix 4: Demonstration video—the MARVIN chatbots study.

[Link]

Multimedia Appendix 5: Privacy policy—the MARVIN chatbots study.

[Link]

Multimedia Appendix 6: Information and consent form.

[Link]

Multimedia Appendix 7: Summary of study procedures for participants.

[Link]

Multimedia Appendix 8: Study questionnaires and semistructured interview guide for the usability study.

[Link]

APPENDIX C SUPPORTING MATERIALS FOR CHAPTER 5

Multimedia Appendix 1: Complete list of conversation topics.

[Link]

Multimedia Appendix 2: CONSORT artificial intelligence (AI) guidelines.

[Link]

Multimedia Appendix 3: Sociodemographic and study questionnaires.

[Link]

Multimedia Appendix 4: Focus group interview guide.

The MARVIN Chatbot usability study

Qualitative focus group/interview script

1. What led you to enroll in this study? (icebreaker/general interest/motivation)

Prompt: What makes you interested in participating this study?

What is your motivation?

2. How easy or hard was it for you to learn to use MARVIN? (perceived ease of use)

Prompt: What made MARVIN easy/hard to use?

Why is it hard/easy?

3. How helpful did you find your conversations with MARVIN? (perceived usefulness)

Prompt: What was most/least helpful about MARVIN?

When do you feel is most helpful?

4. How satisfied are you overall with MARVIN? (attitude towards use)

Prompt: Why do you feel satisfied? What makes you feel satisfied?

5. If you could use MARVIN in the future, to what extent would you? (intent to use)

Prompt: What will make you use MARVIN in the future?

In what situation, or what context?

When will you want to use it?

6. How would you improve MARVIN? (future directions)

Prompt: How could MARVIN be more helpful?

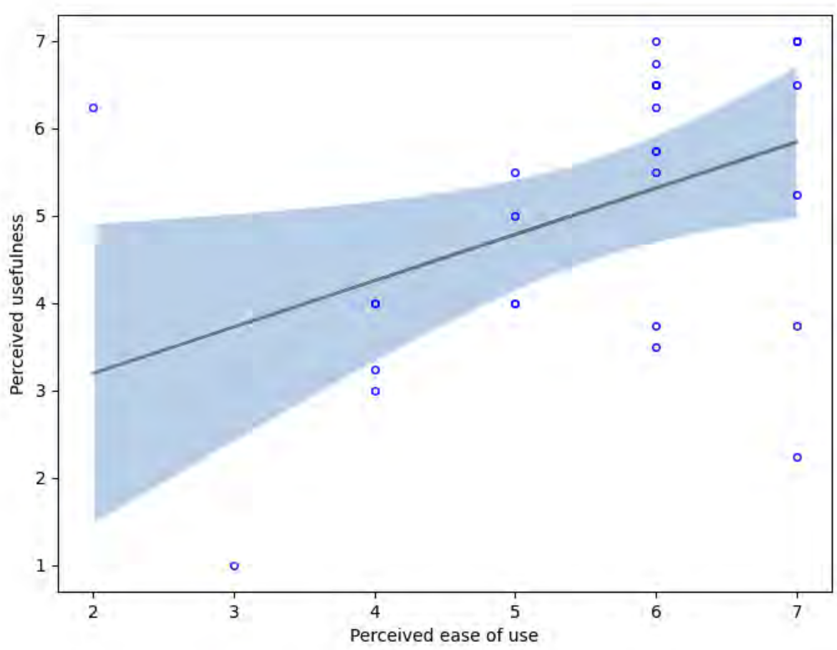
How could MARVIN be more user-friendly?

What topics could be added to MARVIN?

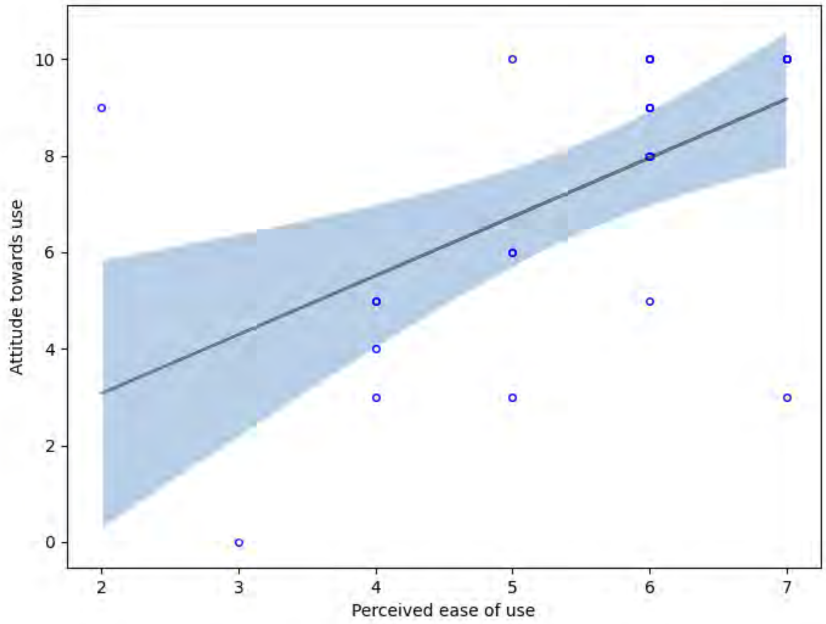
What functions could be added to MARVIN?

Multimedia Appendix 5: Scatterplots – linear regression models of inter-subconstruct relationships.

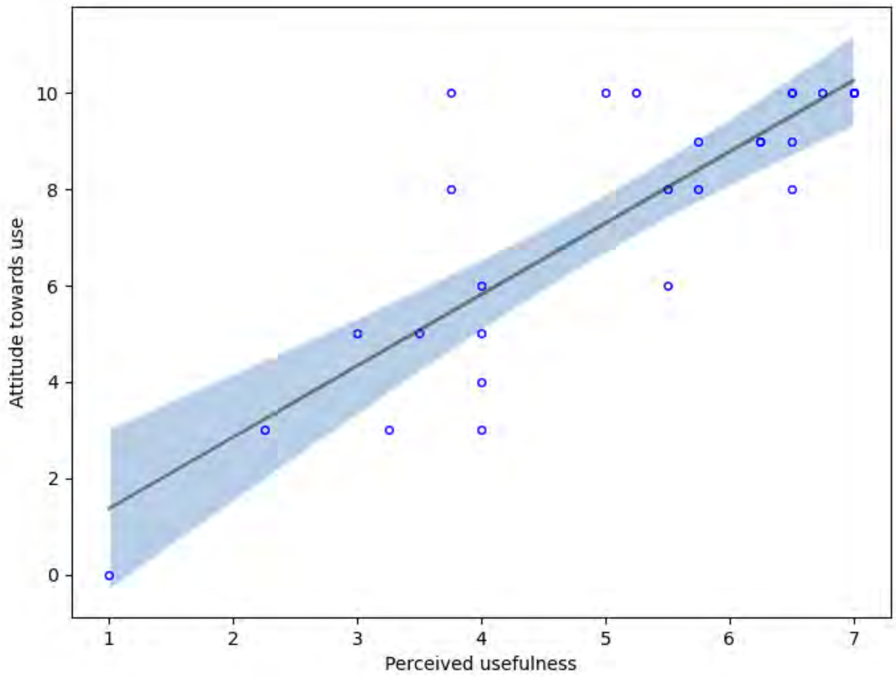
H1: PU=f(PEU)



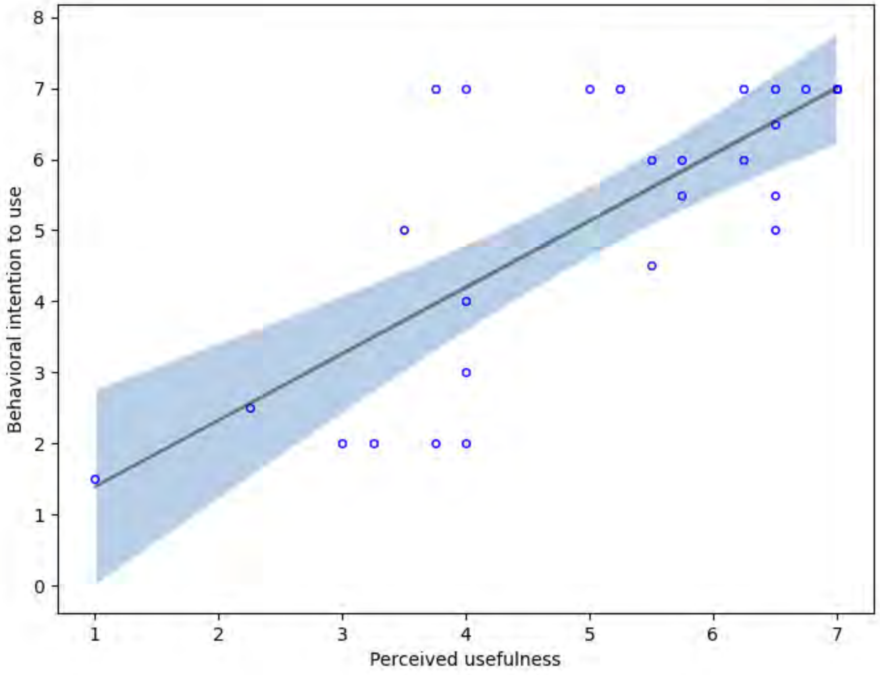
H2: ATU=f(PEU)



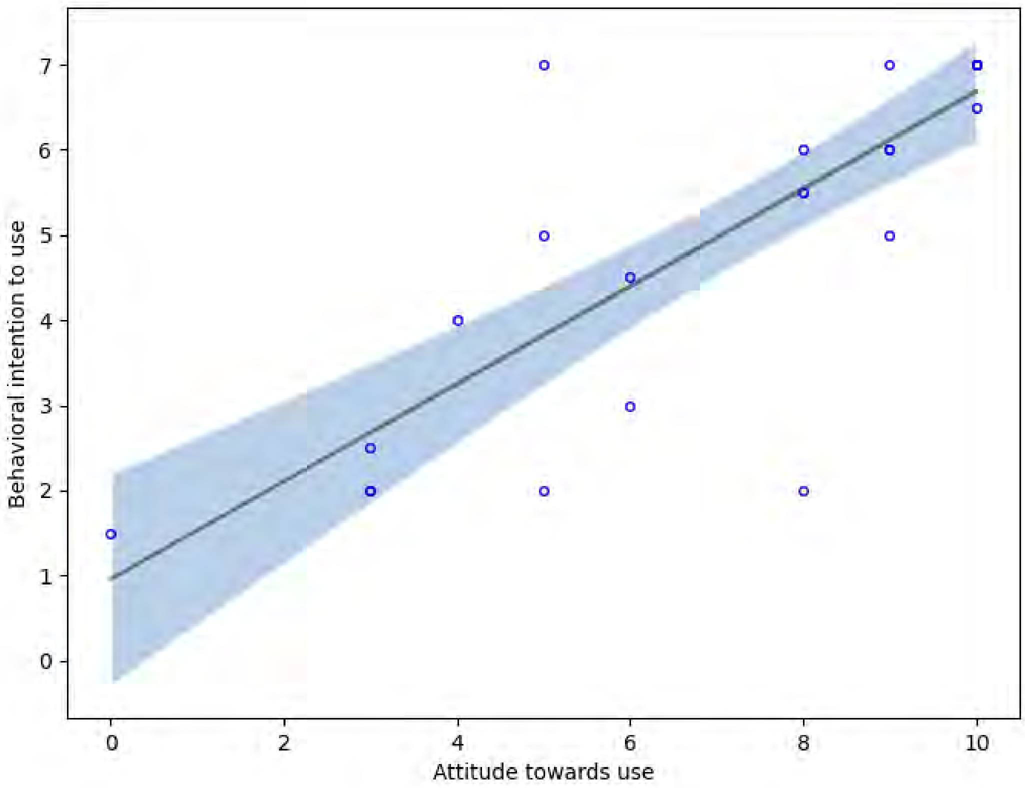
H3: ATU=f(PU)



H4: BIU=f(PU)



H5: BIU=f(ATU)



Multimedia Appendix 6: Complete quotes – qualitative analysis.

[Link]

APPENDIX D SUPPORTING MATERIALS FOR CHAPTER 6

Supplementary file 1. Annotation guideline

Medication Adherence Barriers
&
Risk levels

for People with HIV

Annotation guidelines

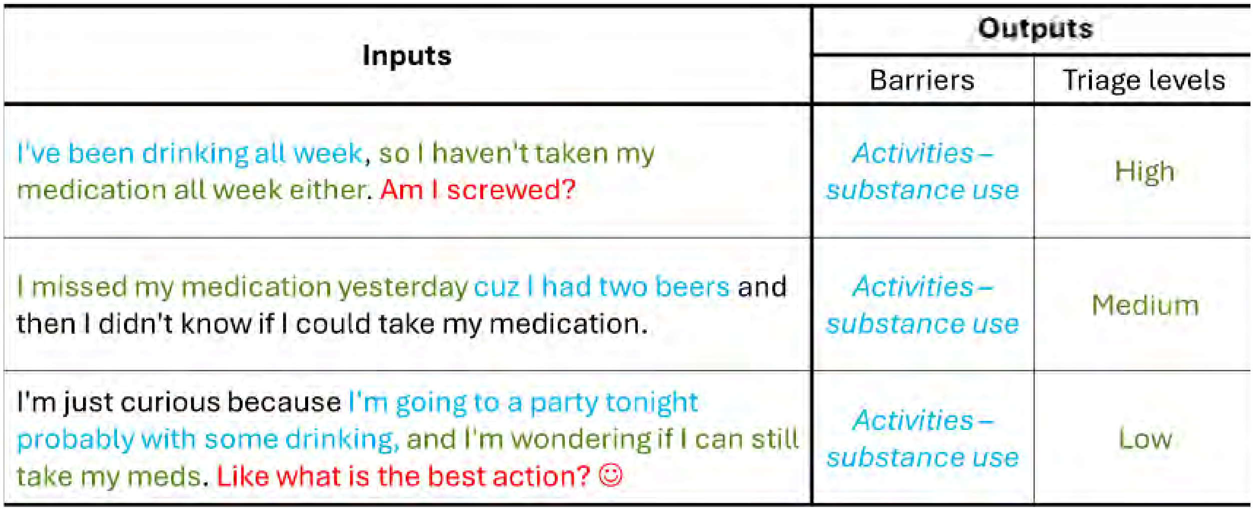
Written by Yuanchao Ma & The MARVIN Chatbot Triage co-construction committee

Introduction and purpose statement

Non-adherence to antiretroviral therapy (ART) is a widespread and multifactorial problem associated with poorer health outcomes that can increase the cost of care and the burden on health systems. However, it has been a challenge to effectively identify factors that affect adherence to medication. This project aims to empower self-management of people with HIV develop a set of deep learning-based classification algorithms that can identify, and extract data related to patients' medication adherence barriers from messages received by a chatbot and triage them according to the level of risk.

These guidelines describe specific ART adherence barriers, based on the I-Score 7-item study, we wish to capture in the messages of patients sent to/discussed with various resources, and the methodology for annotation. In addition to identifying the presence of factors that affect adherence to ART, each barrier will have certain risk levels to further enrich our datasets.

The annotation developed in this project will subsequently be used to train natural language processing systems that can "read" patient messages, identify related ART adherence barriers they may currently be suffering from, and extract them from the conversations.



Inputs	**Outputs**	
	Barriers	Triage levels
I've been drinking all week, so I haven't taken my medication all week either. Am I screwed?	*Activities – substance use*	High
I missed my medication yesterday cuz I had two beers and then I didn't know if I could take my medication.	*Activities – substance use*	Medium
I'm just curious because I'm going to a party tonight probably with some drinking, and I'm wondering if I can still take my meds. Like what is the best action? ☺	*Activities – substance use*	Low

Annotators should carefully read these guidelines in their entirety prior to starting annotations.

Annotation Task

Each message/conversation may have sentences that represent one or more of the barriers that we have defined below. Annotators are tasked with identifying the sentences that describe these ART adherence barriers concepts. If a sentence describes a barrier concept, annotators then assign a risk level of that barrier as further detailed in the specific subsections below. Assigning the label is interchangeably referred to as "annotating" or "tagging" the sentence in these guidelines.

During the annotation process, each message/conversation is pre-split into sentences, where each annotated data point will be a sentence within a document. Most annotations will not require annotators to take in any context other than what is in the boundaries of each sentence. For instances where context outside of the sentence boundaries is needed to make an annotation judgment, we ask annotators to use their best judgment, and to defer to surface level annotation without the use of textual context.

Moreover, annotators should take in the context of the patient-chatbot/physician conversation event, as well as common sense logic. Within reason, an annotator should consider the intent of the patient for a particular sentence.

We first assume in this project that all messages come from people with HIV. If it can be ascertained that it is not from a user with HIV:

- If it is one's partner, we will still categorize it.
- If it is someone with negative diagnostic, we will classify only PrEP-related messages.

The thought process that should be considered is, "If it is not important for the patient to say what he thinks, why did he say it?" More examples of implicit barriers are detailed below, and we ask annotators to consider common sense (not just explicit textual evidence) when labeling sentences.

Finally, a sentence can describe more than one different category of medication adherence barriers, and in some cases a sentence can describe more than one barrier within a single category. **We ask annotators to please label for the moment the most relevant barrier with the highest potential risk per relevant sentence.**

Annotators should annotate strictly according to these guidelines. We acknowledge that there are many different ways to define ART adherence barriers that our guidelines may not always capture. However, the goal of this project is to develop **consistent** annotations for the specific definitions of ART adherence barriers that are detailed below. Sentences may describe a barrier and/or risk level that may seem at first glance to align with one of our barrier categories, but do not meet our specific definitions. **In these cases, please adhere only to the definitions below.** Consistency is key because consistent labels are paramount to being able to train models that can identify and extract barrier.

ART adherence barriers: Definitions

The definition of ART adherence barriers is some aspect of life that may make it difficult for patients to take their HIV medications as prescribed. By "difficult", we mean skipping doses, delaying doses, or being very unpleasant while taking the medication.

In terms of barriers, we followed the original I-Score definition of seven categories and added one new category to refine the prediction task. Each category was defined as follows:

Thoughts and feelings

This includes acceptance of having HIV, emotions (feeling sad, anxious, etc.), medication-related knowledge and beliefs or motivation to take medication.

Example intentions include:

Ask if he/she can stop taking pills

1MT_0741 I don ’ t want to take my medication anymore !

Suicide intention

4MU_2540 i am tired of my HIV condition , just wanna end my life

Always difficult taking meds - thoughts

1MT_2763 I often feel unmotivated to take my meds

Importance of taking medication daily

1MT_0747 Can i skip my complera occasionally?

Habits and activities

This includes daily life (one’s schedule, priorities, etc.) or substance use (alcohol, drugs, or other substances).

Example intentions include:

Missing meds long time

1MT_1534 I haven ’ t taken my stribild for more than a week , what should I do ?

4MU_0846 I didnt take my medication during many days , what should I do

Stop meds during traveling

4MU_0452 Can I take a drug holiday from medication ?

4MU_3036 should i delay my meds if i go to san fransisco

Forgetfulness

4MU_3949 I ' d like some tips on how to stop forgetting to take my medication .

1MT_3083 Often , I forget to take my atripla

Late for meds

4MU_1013 If I miss my medication one day , can I have a double dose in the next ?

4MU_1030 Is any problem if I miss take my medication just one day ?

Social situation

This includes personal relationships or the experience of stigma.

Example intentions include:

didn't take pill because of social stigma

3ISCORE_0604 All that I did say is that the pill I have is the flu medicine, and I decided that day I did n't take my pills, still I did n't leave because of my HIV, but it ' s just, someone who has it there .

didn't take pill because of a lot of people

3ISCORE_0595 When I travel , sometimes I ' m surrounded by several people . Sometimes , like on those days , I did n't take it .

HIV status disclosure

4MU_3683 Should I tell my sexual partner that I am HIV positive ?

4MU_4376 Do I have to tell the customs department that I have HIV ?

3ISCORE_0127 On the face of it , you ca n't tell him , well , look , I ' m , I ' m HIV positive .

Meds claiming during travel

4MU_3704 How do I hide my medication when I travel ?

relationship problem

3ISCORE_5863 Is there anything other than HIV pills , with HIV treatment that ’ s , there was a part of that decision to sort of have you say stay at home and be quiet ?

3ISCORE_1396 There is a lot of trusting, I don ’ t trust .

Economic situation

This includes financial or housing issues.

Example intentions include:

financial constraints

4MU_1210 One of the problem is financial constraints

3ISCORE_1579 Ah , you do n't have any food at home .

3ISCORE_1675 I ask for amounts of money every week

medication costs

4MU_0780 How much it costs medication in Canada ?

4MU_1080 It is cheap to taka hiv treatment ?

support group

4MU_1138 Montreal HIV support group

3ISCORE_1586 There are always organizations to help me

prep costs

4MU_2418 how much does prep cost

Jobless

3ISCORE_0214 Right now , I ' m off work too .

Food access-related problems are a barrier in the ***economic situation***.

Medication

This includes the HIV medication’s side effects, instructions or physical features (for example, pill size or taste).

Example intentions include:

Best time for taking meds?

1MT_1124 Should I take my ART before going to bed instead of the next morning ?

4MU_3873 I take my medication in the evening , but at what time is it recommended ?

change treatment

4MU_0853 I do n't like my medication , can I switch to another one ?

4MU_3871 I can take another antiretroviral treatment without being mine .

Drug interaction

4MU_1391 The medication have some interactions with protein or another supplement ?

4MU_4555 Is ginseng compatible with Juluca ?

side effects

4MU_4570 do my medications increase the risk of developing certain cancers ?

Double meds

4MU_4479 i took my medication yesterday at 20:00 , and when i woke up i did n't know if i had taken it , so i took another one , and then i remembered that i had really taken one the day before . what should i do ?

Care

This includes issues with healthcare professionals, the clinic, the pharmacy, or payment of medication or care.

Example intentions include:

no more treatment

4MU_0364 Can I change my medication by another medication for one time if I don ’ t have any left .

4MU_0922 I only have 3 of my 4 antiretroviral medications while travelling for two weeks .

3ISCORE_0200 There , it was particular that I forgot , that I did n't have enough pills for four days .

where to get the treatment

4MU_4017 Where can I buy my pills ?

alternatives to prep

1MT_3891 What are the other preventive measures in addition to PrEP ?

1MT_3889 What other things can I do after taking prep that will help me lower my risk of hiv infection ?

care access

4MU_3376 where can i find an HIV health provider in Montreal ?

3ISCORE_1003 Because right now , we seem to be , it ' s as if I never had a family doctor .

Health

This includes lab test results, HIV symptoms or overall health.

Example intentions include:

general health - emergency

4MU_0280 Am not able to get the stuff like first add kit

problematic - health

4MU_0880 I have a fever , can I take the medication ?

4MU_0829 I am feeling sick , where should I go ?

Reproductive health

4MU_0357 Can I breastfeed my baby if I ' m HIV positive ?

4MU_3308 what kind of sex can i have if i ' m HIV positive ?

How to know if undetectable

4MU_4179 If I start taking my medication , how long will it take for my viral load to become undetectable ?

U=U Type of sex

4MU_4666 sexual relations if I am undetectable

4MU_0412 Can I have oral sex with my partner if I have an undetectable viral load ?

None

This includes all sentences that do not mention any barrier-related information.

Risk levels: Definition

We defined adherence level based on the number of nonadherent days within a 30-day period:

Perfect adherence: No missed doses

Optimal adherence: ≥95% and <100% (i.e., ≤1 missed dose)

Near-optimal adherence: <95% and ≥80% (i.e., 2–5 missed doses, not occurring consecutively)

Suboptimal adherence: <80% (i.e., ≥6 missed doses)

Thus, we further define the following risk level classification scheme to assess the severity and frequency of adherence-related challenges in the message:

High: Explicit mention of an adherence barrier that leads, or may lead, to suboptimal adherence.
Medium: Explicit mention of adherence barrier that leads, or may lead, to near-optimal adherence.

Low: Explicit mention of an adherence barrier that leads, or may lead, to optimal but not perfect adherence.

None: No mention of adherence barriers, or explicit mention of a barrier with an explicit statement of perfect adherence.

See Table 1 for illustrative examples of adherence risk levels.

Table 1. Illustrative examples of adherence risk levels.

Example	**Barrier**	**Risk level**
And that might mean another 5 days of not taking my medications, cause I will not go back to the pharmacy until my next day off.	Care	High
So, you're going out and sometimes it's going to make you forget about your treatment.	Habits & Activities	Medium
I can't think of anything yet that could stop me, maybe if I was really sick.	Health	Low
I didn't see any difference, because between taking a small tablet and a bigger one at the same time, or taking just one, it doesn't make much difference.	Medication	None

Supplementary file 2

Hyper-parameters for model training

During the model development, the fine-tuning process of the best-performing models follows specific hyper-parameters using one A100 GPU from Google Colab.

A training batch size of 32 was used for BERT family, T5-small and -base.

For T5-large and -xl, auto_find_batch_size was set to “True” for training batch size.

Learning rate was equal to 3e-5 for BERT family, and models were trained for a total of 7 epochs.

Learning rate was equal to 1.1e-4 for T5-small and -base, and models were trained for a total of 10 epochs.

Learning rate was equal to 4e-4 for T5-large and -xl, and models were trained for a total of 10k steps.

Additionally, the LoRA configuration was employed with a rank 'r' of 16 and a 'lora_alpha' value of 32. This configuration targeted the "q" and "v" modules in the transformer layers and incorporates a dropout rate of 0.1.

```
lora_config = LoraConfig(
    task_type = "SEQ_CLS",
    r=16,
    lora_alpha=32,
    target_modules=["q", "v"],
    lora_dropout=0.1,
    bias= 'none',
  )
```

Data augmentation – LLMs candidates and prompting strategies:

```
model_list = {
  "gpt_4_1": "gpt-4.1",
  "llama_3_2": "llama3.2:3b",
  "gemma3": "gemma3:12b",
  "mistral": "mistral:latest",
  "phi4": "phi4:latest",
  "mixtral": "mixtral:8x7b",
}
```

prompt_generation = '''

Your goal is to create diverse and realistic samples for the indicated class while adhering to the task description.

Task description: [We are developing a healthcare chatbot to converse with people with HIV or those at risk. The goal here is to develop a triage algorithm to identify medication adherence problems and the associated risk level. Here's the definition of each category and risk level.

We define eight categories:

1. Thoughts and feelings. This includes acceptance of having HIV, emotions (feeling sad, anxious, etc.), medication-related knowledge and beliefs, or motivation to take medication.

2. Habits and activities. This includes daily life (one's schedule, priorities, etc.) or substance use (alcohol, drugs, or other substances).

3. Social situation. This includes personal relationships or the experience of stigma.

4. Economic situation. This includes financial or housing issues.

5. Medication. This includes the HIV medication's side effects, instructions, or physical features (for example, pill size or taste).

6. Care. This includes issues with healthcare professionals, the clinic, the pharmacy, or payment of medication or care.

7. Health. This includes lab test results, HIV symptoms, or overall health.

8. None. This includes all sentences that do not mention any barrier-related information.

Regarding the risk level, we first define adherence level based on the number of nonadherent days within a 30-day period:

- Perfect adherence: No missed doses

- Optimal adherence: ⩾95% and <100% (i.e., ⩽1 missed dose)

- Near-optimal adherence: <95% and ⩾ 80% (i.e., 2 – 5 missed doses, not occurring consecutively)

- Suboptimal adherence: <80% (i.e., ⩾6 missed doses)

Then we define four levels: high, medium, low, and none.

1) High: Explicit mention of an adherence barrier that leads, or may lead, to suboptimal adherence.

2) Medium: Explicit mention of adherence barrier that leads, or may lead, to near-optimal adherence.

3) Low: Explicit mention of an adherence barrier that leads, or may lead, to optimal but not perfect adherence.

4) None: No mention of adherence barriers, or explicit mention of a barrier with an explicit statement of perfect adherence.

Instructions:

- Analyze the task description and identify key attributes or features for each class.

- Use reasoning to ensure the synthetic data aligns with the characteristics of the task and its real-world relevance.

- Generate 50 synthetic samples for **[Specific barrier]**, in JSON format, following this:

Format the output as a JSON array, not a list, and output nothing more. The objects have the keys:

'data', 'barrier', 'risklvl'. \

The 'data' should be a realistic text sample that represents the class. Ensure the sample is diverse in terms of text length and content.

The 'barrier' should be one of the nine categories defined above.

The 'risklvl' should be one of the four levels defined above.

- Ensure the samples are diverse, balanced, and representative of each class. Also, ensure that the examples are sufficiently life-like and that the wording is sufficiently colloquial, with enough details.

Give me only a nested JSON object, like {'synthetic_data': [{...}]}, where ... are the keys defined above.

'''

Classification – LLMs prompting strategies:

```
input = [
   #        {"role": "system", "content": "You are an annotator, please provide labels for the given sentences considering the following definition."},
   #        {"role": "user", "content": prompt + text}
   #    ],
```

For zero-shot:

prompt = prompt_classification_zero_shot

prompt_classification_zero_shot = '''

Regarding adherence barrier, we define eight categories:

1. Thoughts and feelings. This includes acceptance of having HIV, emotions (feeling sad, anxious, etc.), medication-related knowledge and beliefs, or motivation to take medication.

2. Habits and activities. This includes daily life (one’s schedule, priorities, etc.) or substance use (alcohol, drugs, or other substances).

3. Social situation. This includes personal relationships or the experience of stigma.

4. Economic situation. This includes financial or housing issues.

5. Medication. This includes the HIV medication’s side effects, instructions, or physical features (for example, pill size or taste).

6. Care. This includes issues with healthcare professionals, the clinic, the pharmacy, or payment of medication or care.

7. Health. This includes lab test results, HIV symptoms, or overall health.

8. None. This includes all sentences that do not mention any barrier-related information.

Regarding the risk level, we first define adherence level based on the number of nonadherent days within a 30-day period:

- Perfect adherence: No missed doses
- Optimal adherence: ⩾95% and <100% (i.e., ⩽1 missed dose)
- Near-optimal adherence: <95% and ⩾ 80% (i.e., 2 – 5 missed doses, not occurring consecutively)
- Suboptimal adherence: <80% (i.e., ⩾6 missed doses)

Then we define four levels: high, medium, low, and none.

1) High: Explicit mention of an adherence barrier that leads, or may lead, to suboptimal adherence.

2) Medium: Explicit mention of adherence barrier that leads, or may lead, to near-optimal adherence.

3) Low: Explicit mention of an adherence barrier that leads, or may lead, to optimal but not perfect adherence.

4) None: No mention of adherence barriers, or explicit mention of a barrier with an explicit statement of perfect adherence.

Instructions:

- Analyze classification definition descriptions to identify key attributes or features of each category.

- Use reasoning to ensure that the predicted labels match the categorical definitions.

- Predict the corresponding barrier category first and then predict the risk level for the input, in JSON format, following this:

Format the output as a JSON array, not a list, and output nothing more. The objects have the keys:

"barrier", "risklvl".

The "barrier" should be your prediction from one of the eight categories defined above.

The "risklvl" should be your prediction from one of the four levels defined above.

Don't give me ANY explanation. Give me ONLY a nested JSON object, like {"prediction": [{...}]}, where ... are the keys defined above.

Now let's think step by step, generate the predictions, and output the results in JSON format.

Here is the input: {input}

'''

For five-shots:

prompt = prompt_classification_five_shot

prompt_classification_five_shot = '''

Regarding adherence barrier, we define eight categories:

1. Thoughts and feelings. This includes acceptance of having HIV, emotions (feeling sad, anxious, etc.), medication-related knowledge and beliefs, or motivation to take medication.

2. Habits and activities. This includes daily life (one's schedule, priorities, etc.) or substance use (alcohol, drugs, or other substances).

3. Social situation. This includes personal relationships or the experience of stigma.

4. Economic situation. This includes financial or housing issues.

5. Medication. This includes the HIV medication's side effects, instructions, or physical features (for example, pill size or taste).

6. Care. This includes issues with healthcare professionals, the clinic, the pharmacy, or payment of medication or care.

7. Health. This includes lab test results, HIV symptoms, or overall health.

8. None. This includes all sentences that do not mention any barrier-related information.

Regarding the risk level, we first define adherence level based on the number of nonadherent days within a 30-day period:

- Perfect adherence: No missed doses
- Optimal adherence: ⩾95% and <100% (i.e., ⩽1 missed dose)
- Near-optimal adherence: <95% and ⩾ 80% (i.e., 2 – 5 missed doses, not occurring consecutively)
- Suboptimal adherence: <80% (i.e., ⩾6 missed doses)

Then we define four levels: high, medium, low, and none.

1) High: Explicit mention of an adherence barrier that leads, or may lead, to suboptimal adherence.

2) Medium: Explicit mention of adherence barrier that leads, or may lead, to near-optimal adherence.

3) Low: Explicit mention of an adherence barrier that leads, or may lead, to optimal but not perfect adherence.

4) None: No mention of adherence barriers, or explicit mention of a barrier with an explicit statement of perfect adherence.

Instructions:

- Analyze classification definition descriptions to identify key attributes or features of each category.

- Use reasoning to ensure that the predicted label match the categorical definition.

- Predict the corresponding barrier category first and then predict the risk level for the input, in JSON format, following this:

Format the output as a JSON array, not a list, and output nothing more. The objects have the keys:

"barrier", "risklvl".

The "barrier" should be your prediction from one of the eight categories defined above.

The "risklvl" should be your prediction from one of the four levels defined above.

Just give me the closest set of barriers and risk levels. If there are two sets that apply, give me the one with the highest risk level. Don't give me your explanation. Give me ONLY a nested JSON object, like {"prediction": [{...}]}, where ... are the keys defined above.

Examples:

Here are several classification examples:

1. "I've been drinking all week, so I haven't taken my medication all week either. Am I screwed?"

- The output should be:

{"prediction": [{"barrier": "Habits and activities", "risklvl": "High"}]}

2. "I sometimes feel unmotivated to take my meds, can i skip my complera occasionally? "

- The output should be:

{"prediction": [{"barrier": "Thoughts and feelings", "risklvl": "Medium"}]}

3. "At the very beginning , the constraint was that we had to see the doctor quite often , because of research protocols."

- The output should be:

{"prediction": [{"barrier": "Care", "risklvl": "Low"}]}

4. "I'm undetectable now, so everything is fine for me."

- The output should be:

{"prediction": [{"barrier": "Health", "risklvl": "None"}]}

5. "To get the medication , you do n't necessarily have the money."

- The output should be:

{"prediction": [{"barrier": "Economic situation", "risklvl": "High"}]}

Now let's think step by step, generate the predictions, and output the results in JSON format.

Here is the input: {input}

'''

Demographic descriptor injection – LLMs prompting strategies:

prompt_socio_adding = """

Randomly add one of race/ethnicity [asian, black, white, latino, indigenous] and gender [woman, man, trans] description into the given sentences as natural as you can, and put the modified demographic description in bracket at the beginning, like these examples:

1. If I'm HIV positive , what is the risk of contamination ? =>
   As a [latino male] latino man, if I'm HIV positive, what is the risk of contamination ?

2. Is this a symptom of hiv ? =>
   Is this a symptom of hiv ? I am an indigenous trans [indigenous tran].

3. I ' m very afraid of hiv. =>
   I ' m a white man [white male] and I'm very afraid of hiv.

4. Can you enlighten me I ' m very afraid . =>
   Can you enlighten me, as a black woman [black female] I ' m very afraid

Now, please think step by step, and generate me just the new sentences with the indicators:
"""

Table 1. Most common discrepancies between ground-truth and best-performing model prediction for each task (Count>=10)

Task	Ground truth	Model prediction	Count
Barrier prediction	None	Thoughts & Feelings	20
	None	Habits & Activities	15
	Thoughts & Feelings	Habits & Activities	11
	Thoughts & Feelings	None	10
Risk level prediction	None	Low	38
	None	Medium	37
	Medium	None	29
	Medium	Low	27
	Low	None	22
	Low	Medium	21
	Medium	High	12
	High	Medium	11

APPENDIX E SUPPORTING MATERIALS FOR CHAPTER 7

Key Scientific Contributions - the MARVIN Project

Peer-reviewed journals

1. **Ma Y**, Achiche S, Tu G, Vicente S, Lessard D, Engler K, Lemire B, MARVIN Chatbots Patient Expert Committee, Laymouna M, de Pokomandy A, Cox J, Lebouché B. (2024) The first AI-based Chatbot to promote HIV self-management: A mixed methods usability study. HIV medicine; 1-23.
2. **Ma Y**, Achiche S, Pomey MP, Paquette J, Adjtoutah N, Vicente S, Engler K, Laymouna M, Lessard D, Lemire B, Asselah J, Therrien R, Osmanlliu E, Zawati MH, Joly Y, Lebouché B. (2024) Adapting and evaluating an AI-Based Chatbot through patient and stakeholder engagement to provide information for different health conditions: Master Protocol for an Adaptive Platform Trial (the MARVIN Chatbots Study). JMIR Research Protocols; 13(1):e54668.
3. Jaiswal, N., **Ma, Y.**, Lebouché, B., Poenaru, D., & Osmanlliu, E. (2025). PedMedQA: Comparing Large Language Model Accuracy in Pediatric and Adult Medicine. Pediatrics Open Science, 1(2), 1-3.
4. Laymouna M, **Ma Y**, Lessard D, Engler K, Therrien R, Schuster T, Vicente S, Achiche S, Haj MNE, Lemire B, Kawaiah A, Lebouché B. (2024) Needs-Assessment for an Artificial Intelligence-Based Chatbot for Pharmacists in HIV Care: Results from a Knowledge–Attitudes–Practices Survey. Healthcare.
5. Laymouna M, **Ma Y**, Lessard D, Schuster T, Engler K, Lebouché B. (2024) Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. Journal of Medical Internet Research; 26:e56930.
6. **Ma Y**, Tu G, Lessard D, Vicente S, Engler K, Achiche S, Laymouna M, de Pokomandy A, Lebouché B. (2023) An Artificial Intelligence-Based Chatbot to Promote HIV Primary Care Self-Management: a Mixed Method Usability Study. *The Annals of Family Medicine*, *21(*S2*).*
7. **Ma Y**, Engler K, Vicente S, Achiche S, Lemire B, Cruz AR, Thériault L, Soussou S, Regazzoni B, Tu G, Haj MNE. (2021) Usability of an artificial intelligence chatbot to facilitate self-management of antiretroviral therapy in HIV patients. *HIV Medicine*, *22*(S3), pp.240-242.

Peer-reviewed conference contributions – Oral presentation

1. Frija-Gruman NM, Villanueva S, **Ma Y**, Asselah J, Lambert S, Achiche S, Osmanlliu E, Pomey M-P, Hijal T, Lessard D, Engler K, Lin J, & Lebouché B. Development of an AI-based emotional tone classifier to support psychosocial triage in digital cancer nursing tools. EONS18 at ESMO 2025, Berlin, Germany. October 17-21, 2025
2. Villanueva Guzman DS, **Ma Y**, Achiche S, Cadri A, Lebouché B. (2025) An AI-Powered Preventive Intervention for Stigma and Suicidal Ideation in HIV Self-Management: Development, Evaluation, and User Testing of the MARVIN Chatbot's integrate mental health management module. Oral presentation at to the 13th IAS conference on HIV Science. Kigali, Rwanda. July 13-17, 2025.
3. Villanueva S, **Ma Y**, Achiche S, Engler K, Lessard D, Lebouché B. An AI-Powered Preventive Intervention for Stigma and Suicidal Ideation in HIV Self-Management. Oral presentation at the Conference on Retroviruses and Opportunistic Infections 2025. San Francisco, US. Mar 9-

12, 2025

4. Lin J, Lebouché B, **Ma Y**, Engler K, Lessard D & Osmanlliu E. Hope or hype? How do Canadian English and French language news articles report on artificial intelligence chatbots in health and medicine? Oral Presentation at the Department of Family Medicine Academic Day, McGill University, Montreal, Canada. May 30, 2025
5. **Ma Y**, Achiche S, Lessard D, Engler K, Vicente S, Tu G, Lemire B, Osmanlliu E, Lebouché B. Development of an AI Chatbot-Based Triage Solution to Support ART adherence for People with HIV. Artificial Intelligence in Infectious Disease Workshop 2024. Amsterdam, Netherlands, Virtual meeting, Dec 5-6, 2024
6. Villanueva S, **Ma Y**, Achiche S, Lessard D, Engler K, Lebouché B. Development of an AI Chatbot-Based Triage Solution to Support ART adherence for People with HIV. Artificial Intelligence in Infectious Disease Workshop 2024. Amsterdam, Netherlands, Virtual meeting, Dec 5-6, 2024
7. **Ma Y**, Tu G, Lessard D, Vicente S, Engler K, Achiche S, Laymouna M, de Pokomandy A, Pomey MP, Lebouché B. An Artificial Intelligence-Based Chatbot to Promote HIV Primary Care Self-Management: a Mixed Method Usability Study. North American Primary Care Research Group (NAPCRG) 51st Annual Meeting. San Francisco, US, Oct 30-Nov 4, 2023
8. Jaiswal N, **Ma Y**, Engler K, Lessard D, Lebouché B, Osmanlliu E. Patient and Stakeholder Engagement in the Integration of Large Language Models in Healthcare Chatbots. T-CAIREM conference: AI in Medicine, Toronto, ON, Canada, Oct 12-13, 2023
9. **Ma Y**, Achiche S, Lebouché B. Development of a Chatbot for HIV patients using Participatory Design. E-Health 2022 Virtual Conference & Tradeshow, Canada, Jun 1-2, 2022.
10. **Ma Y**, Tu G, Lymouna M.A., Achiche S, Frih I, Chehab M, Thériault L, Lemire B, Haj MNE,Lebouché B. Développement d'un Chatbot pour soutenir l'autogestion chez les personnes vivant avec le VIH en temps de pandémie. Symposium on the Challenges of using health information technology in times of COVID-19, ACFAS2022, Canada, Virtual meeting, May 9-13, 2022.
11. **Ma Y**, Engler K, Vicente S, Achiche S, Lemire B, Rodriguez Cruz A, Thériault L, Soussou S, Régazzoni B, Tu G, Haj MNE, de Pokomandy A, Cox J, Zahedi Niaki N, Lebouché B. Usability and Acceptability of an Artificial Intelligence-based Chatbot to Facilitate Antiretroviral Self-management in People Living with HIV. 31st Annual Canadian Conference on HIV/AIDS Research. London, ON, Canada, Virtual meeting, Apr 27-29, 2022.
12. **Ma Y**, Tu G, Frih I, Haj MNE, Lymouna M.A., Chehab M, Thériault L, Lemire B, Achiche S, Lebouché B.(2022, March 16). Enchanté, MARVIN : Un chat-robot pour aider les personnes vivant avec le VIH à mieux autogérer leur santé. Les Journées Québécoises du VIH, Montréal, Québec, Canada, Feb 17, 2022
13. **Ma Y**, Lessard D, Engler K, Vicente S, Achiche S, Lemire B, Rodriguez Cruz A, Thériault L, Soussou S, Régazzoni B, Tu G, Haj MNE, Frih I, de Pokomandy A, Cox J, Zahedi Niaki N, Lebouché B, Usability of the MARVIN Chatbot: Empowering People Living with HIV to Improve Self-Management of Antiretroviral Treatment. Workshop on Healthy Living with HIV 2021, Virtual meeting, Oct 1-2, 2021.
14. **Ma Y**, Bock G, Cherni G, Theriault L, Lessard D, Achiche S, Lebouché B, Marvin: Intelligent Conversational Agent for HIV Patient self-management. Canada Pavilion at AIDS 2020, July 6-10,2020.
15. **Ma Y**, Bock G, Cherni G, Lemire B, Therrien R, Thériault L, Lessard D, Engler K, Achiche S, Lebouché B, Meet Marvin, the Chatbot: Using Artificial Intelligence to Engage HIV Patients in their Antiretroviral Therapy. Workshop on Healthy Living with HIV 2020, Barcelona, Spain,

Nov 19-20, 2020.

Peer-reviewed conference contributions – Poster presentation

16. Frija-Gruman, N. M., Villanueva, S., Ma, Y., Osmanlliu, E., Lambert, S., Pomey, M.-P., Hijal, T., Lessard, D., Engler, K., Lin, J., Asselah, J., & Lebouché, B. (2025, July). Emotional tone classification as a tool for psychosocial risk detection in oncology: An AI-powered chatbot. AACR Special Conference in Cancer Research: Artificial Intelligence and Machine Learning, Montreal, QC, Canada. July 10-12, 2025
17. Villanueva S, **Ma Y**, Achiche S, Engler K, Lessard D, Lebouché B. An AI-Powered Preventive Intervention for Stigma and Suicidal Ideation in HIV Self-Management. Conference on Retroviruses and Opportunistic Infections (CROI) 2025, San Francisco, US, Mar 9-12, 2025
18. **Ma Y**, Achiche S, Lessard D, Engler K, Vicente S, Lemire B, Laymouna M, Lebouché B. Developing of an AI Chatbot-Based Triage Solution for Enhancing Antiretroviral Treatment Adherence in People with HIV. North American Primary Care Research Group (NAPCRG) 52nd Annual Meeting, Quebec City, Canada, Nov 20-24, 2024
19. Villanueva S, **Ma Y,** Achiche S, Lessard D, Lebouché B. Development of a ChatGPT-Powered Crisis and Suicidal Ideation Management Module for an HIV Self-Management Chatbot. North American Primary Care Research Group (NAPCRG) 52nd Annual Meeting, Quebec City, Canada, Nov 20-24, 2024
20. Laymouna M, **Ma Y**, Therrien R, Lessard D, Engler K, Vicente S, Schuster T, Achiche S, Haj MNE, Kawaiah A, Lemire B, Lebouché B. An Artificial Intelligence-based Chatbot for Pharmacists in HIV Care: Results from a Knowledge-Attitudes-Practices Needs-Assessment Questionnaire. E-Health 2024 Conference, Vancouver, BC, Canada, May 26-28, 2024
21. Laymouna M, **Ma Y**, Therrien R, Lessard D, Engler K, Vicente S, Schuster T, Achiche S, Haj MNE, Kawaiah A, Lemire B, Lebouché B. An Artificial Intelligence-based Chatbot for Pharmacists in HIV Care: Results from a Knowledge-Attitudes-Practices Needs-Assessment Questionnaire. 33rd Canadian Conference on HIV/AIDS Research, London, ON, Canada, April 25-28, 2024
22. **Ma Y**, Vicente S, Engler K, Achiche S, Theriault L, Lemire B, Tu G, Frih I, Haj MNE, Laymouna M, de Pokomandy A, Cox J, Lebouche B. An artificial intelligence-based chatbot aimed at promoting HIV self-management: quantitative results of a mixed-method usability study. Fast-track cities 2022, Sevilla, Spain, Oct 11-13, 2022
23. Frih I, Tu G, **Ma Y**, Achiche S, Laymouna MA, Chehab M, Thériault L, Lemire B, Haj MNE EHM, Lebouché B. Lessons learned from the development and implementation of MARVIN: a bilingual artificial intelligence Chatbot for people living with HIV. AIDS2022, Montreal, QC, Canada, Jul 29-Aug 2, 2022.
24. **Ma Y**, Engler K, Vicente S, Achiche S, Lemire B, Rodriguez Cruz A, Thériault L, Soussou S, Régazzoni B, Tu G, Haj MNE, de Pokomandy A, Cox J, Zahedi Niaki N, Lebouché B. Utilisabilité et acceptabilité d'un Chatbot basé sur l'intelligence artificielle pour faciliter l'autogestion des antirétroviraux chez les personnes vivant avec le VIH. 11th AFRAVIH Conference, Marseilles, France, Apr 6-9, 2022.
25. Yan Y, Engler K, Lessard D, **Ma Y**, Pomey M-P, Lebouché B. Solutions de télésanté mises en œuvre par les hôpitaux pour les soins de la COVID-19 : un examen rapide. 11th AFRAVIH Conference, Marseilles, France, Apr 6-9, 2022.
26. **Ma Y**, Engler K, Vicente S, Achiche S, Lemire B, Rodriguez-Cruz A, Thériault L, Soussou S, Régazzoni B, Tu G, Haj MNE, Frih I, de Pokomandy A, Cox J, Zahedi NN, Lebouché B.

Usability of an Artificial Intelligence Chatbot to Facilitate Self-Management of Antiretroviral Therapy in HIV Patients. 18th European AIDS Conference (EACS 2021), London, UK, Virtual meeting, Oct 27-30, 2021

27. **Ma Y**, Sanmiguel D, Lessard D, Theriault L, Achiche S, Arora A, Mate K, Lemire B, Hijal T, Schuster T, Kildea J, de Pokomandy A, Cox J, Kronfli N, Klein M, Lebouché B, Chatbot MARVIN: Development study of an Intelligent Conversational Agent to Promote HIV Patients' Engagement in Care and Management of ART Adherence Barriers. 29th Annual Canadian Conference on HIV/AIDS Research, Quebec City, Canada, May 1-2, 2020
28. **Ma Y**, Sanmiguel D, Lessard D, Theriault L, Achiche S, Lacombe K, Routy JP, Arora A, Mate K, Lemire B, Schuster T, Hijal T, de Pokomandy A, Kronfli N, Cox J, Lebouché B. (2020). Chatbot MARVIN: Étude de développement d'un agent conversationnel basé sur l'Intelligence artificielle pour mieux engager les patients dans leur gestion des barrières à une bonne observance aux antirétroviraux. 10th AFRAVIH Conference, Virtual meeting, Nov 8-11, 2020

Seminars and Media Engagement – the MARVIN Project

Academic seminar

1. **Ma Y** and Lebouché B. (2025) AI for Person-Centred Care: The MARVIN Chatbot in HIV and Beyond. Scientific Series, St. Mary's Research Centre, Montreal, CA
2. **Ma Y** and Lebouché B. (2025) Human-centeredness, Healthcare AI & Patient Partnership. Café Klatsch on Artificial intelligence, McGill University, Montreal, CA
3. Lebouché B and **Ma Y**. (2024) When AI Meets Healthcare. Scientific Conference, Santa Marcelina Hospital at Sao Paolo, Brazil
4. **Ma Y**. (2024) Comment les patients peuvent bénéficier de l'intelligence artificielle: exploration de l’implémentation du Chatbot MARVIN pour l'autogestion des patients. Scientific Conference, Bureau de Consultation en Statistique, ETS.
5. **Ma Y**. (2023) Enchanté, MARVIN: Artificial Intelligence-based Chatbots to Facilitate Patient Self-management For Different Health conditions. Seminar presentation, IDlGH Trainee Seminar Series, RI-MUHC.
6. Achiche S, **Ma Y**. (2021) When AI meets Primary Care, Invited lecture for FMED702 Advanced Doctoral Primary Care Research Seminars, McGill University
7. **Ma Y** et al. (2020) Marvin: Intelligent Conversational Agent for Patient self-management. CVIS academic rounds, MUHC
8. **Ma Y**. (2020) Marvin: Intelligent Conversational Agent for HIV Patient self-management. PolyAI, Polytechnique Montreal

Interview

1. Lafontaine Y. *Marvin, l’Intelligence artificielle pour un meilleur engagement des patients VIH.* Fugues Magazine. 01/11/2020
2. Rassy S. *HIV cases surge in Quebec, experts call for increased access to prevention and support*. CTV news. 01/12/2024

APPENDIX F DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the author used GPT-4o (version 2025-03-27) to enhance the clarity and conciseness of the language of their own text. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the thesis.