Generative AI for Health

IN LOW & MIDDLE INCOME COUNTRIES

STANFORD
CENTER FOR
DIGITAL HEALTH

TABLE OF CONTENTS

KEY DEFINITIONS

Generative artificial intelligence (GenAI):
computational techniques capable of generating seemingly new, meaningful content such as text, images, or audio from training data.[1]

Large language model (LLM):
A a type of GenAI system trained on large amounts of text data, that understands and generates human-like language.[2]

Low- and middle-income countries (LMICs):
As classified by the World Bank Atlas method using gross national income (GNI) per capita.[3]

Health behaviors or health-related behaviors:
Intentional or unintentional actions taken by individuals that affect health or mortality.[4] Examples include smoking, diet, physical activity, sleep, substance use, risky sexual activities, healthcare seeking behaviors, and adherence to prescribed medical treatments and vaccination programmes.

Human-in-the-loop (HITL):
Cause of human interaction or intervention to control or change the outcome of a process.[5]

Retrieval-Augmented Generation (RAG):
A technique for enhancing the accuracy of LLM outputs by retrieving relevant information from specific external sources to supplement the model's training data.[6]

Tokens:
The basic units of text processed by a language model. Depending on tokenization strategy, a token may comprise a phrase, a word, part of a word, or a character. Language models break down text into tokens to analyze and generate responses. The number of tokens used in a query affects processing time, cost, and the amount of information the model can consider at once.[7]

Application Programming Interface (API):
A set of rules and tools that allows different software applications to communicate with each other. In the context of AI, an API enables developers to integrate AI capabilities—such as text generation, speech recognition, or image analysis—into their own applications without needing to build an AI model from scratch.[8]

EXECUTIVE SUMMARY

Generative AI (GenAI) has the potential to improve health and healthcare in low- and middle-income countries (LMICs). Where is GenAI currently being used and what are the greatest successes? How can we realize greater impact and unlock the full potential of GenAI, both for behavior change and broader healthcare applications?

To help answer these questions, from August to December 2024 we conducted an extensive review including two roundtable events, in-depth interviews with dozens of people who are actively working on applications of GenAI to health and healthcare in LMICs (including academics, health system leaders, implementers, and funders), and a quantitative survey with over 100 respondents. Additionally, we reviewed 14 GenAI accelerator programs for health that have collectively supported over 250 projects worldwide.

This white paper has a specific focus on the use of GenAI tools to drive health-related behavior change (HBC). Our scoping analysis, framework and key recommendations are inclusive of a range of health use cases, to contextualise HBC interventions within the wider ecosystem, and facilitate broadly applicable learnings.

Here is what we found.

Where is GenAI being currently used, and what are the greatest successes?

Use cases typically centered around applying large language models (LLMs) to health and healthcare-related tasks related to summarization, classification, extraction, translation, and/or conversation (please see definitions in Table 1). Use cases typically fell in one of three categories: direct-to-consumer, direct-to-provider, or system-level.

Some examples include:

Direct-to-Consumer

Offering personalized counseling on sensitive topics (e.g., HIV testing, sexual and reproductive health) via conversational LLM-based agents. In some cases, using improved voice capabilities of LLMs to better engage consumers, especially in low-literacy settings.

Direct-to-Provider

Providing better support for healthcare worker-to-consumer communication in traditional helpdesk workflows, including triaging and routing incoming questions, providing personalized suggested responses for healthcare workers, and live translation between languages.

System-Level

Generating early-warning alerts for potential emerging pandemics by analyzing large amounts of unstructured data from diverse sources, such as health records, news articles, social media and climate data.

EXECUTIVE SUMMARY

We provide quantitative summaries of projects from the GenAI accelerator programs, encompassing a broad range of health use cases, as well as survey respondents' perspectives on priority use cases and health areas, and key factors and barriers for successful implementation of GenAI health interventions. Additionally, we profile five case studies of GenAI deployments with a specific focus on health-related behavior change in LMICs.

In terms of scale of deployment, we found many projects in the "pilot phase", including some that are deployed to over 10,000 monthly users. As of late 2024, we found only one application reaching scale (to 100,000 or more monthly users) of GenAI in health-related behavior change for LMICs, detailed in our included case studies. Several pilots had promising preliminary data on cost-effective impact, including health worker efficiency gains, and all are planning further evaluation in 2025 with a move to greater scaling. Pilots conducted as part of a broader scaled system (for example, an existing helpdesk workflow with millions of total users that is now testing integrating GenAI for efficiency improvements) have a more predictable path to fast scaling.

Given the nascency of the field, the relatively small scale of existing projects and limited evaluation data is unsurprising, but highlights the need for sustained focus to realize greater health impact.

While this review represents the most comprehensive analysis of GenAI in health-related behavior change to date, it is not exhaustive. Our findings focus on deployments funded by major GenAI accelerator programs and insights from expert interviews. While we believe this provides a representative snapshot of the current landscape, we recognize that some applications may not be captured. Additionally, this paper has not directly explored the perspectives of end users. We welcome input on additional large-scale deployments and user-centered insights to build collective knowledge in this evolving space.

> *"To unlock the full potential of generative AI in healthcare for low- and middle-income countries, we must bridge technical innovation with local realities. This means sharing knowledge, building inclusive infrastructure, and creating systems that learn and evolve with communities. The true measure of success is not just technological advancement, but the lives we improve and the health disparities we reduce through thoughtful, collaborative action."*
>
> - Fei-Fei Li, PhD, Co-Director, Stanford Human-Centered AI Institute (HAI), Professor of Computer Science, Stanford University

HOW CAN WE REALIZE GREATER IMPACT AND UNLOCK THE FULL POTENTIAL OF GENAI?

Share learnings

Stakeholders wanted to learn more from others' experiences; this is especially important given how quickly technology and applications are evolving. Specific needs included: (a) understanding of the types of tasks LLMs are well suited to; their weaknesses; and strategies to address; and (b) summaries of specific successes, with concrete case studies reporting on comparable outcome metrics.

Strategies to address:

a. Produce practical guidance on how to identify LLM applications while mitigating risks and then pilot/validate/scale them. A regular update process will be required given technical capabilities are changing quickly.

b. Utilize consistent outcome metrics to describe scale of projects (such as monthly active users, total users and retention of users) and specificity regarding the type of AI system being used (for example deterministic vs. generative) to facilitate meaningful comparisons and benchmarking.

c. Establish a regular process to identify successes and disseminate learnings, including case studies.

Focus on actionable measurement

Stakeholders wanted better ways of measuring benefits, costs, and risks, in ways that provide rigorous but also timely data. For example, funders cannot wait 3 years for results of a randomized controlled trial to guide annual investment decisions, but we still need scientifically valid ways to measure success to inform implementation decisions in the interim. Establishing a clear evidence base will also be essential for supporting government decisions to implement successful applications at a national scale.

Strategies to address:

a. Establish standards for measurement and best practice, with concrete examples.

b. Identify opportunities for implementer partnership with academics on measurement. Since many projects are in pilot phase (and some are starting to scale), there is a time-sensitive opportunity for accelerative partnerships.

Improve language & localization

Experts noted that the quality of models varies by language, by medium (with voice particularly important for low-literacy settings) and by use case (e.g., health-specific contexts).The fact that large language models are not trained on or fluent in local languages was the most commonly selected barrier to using GenAI in healthcare settings in LMICs in our quantitative survey. We highlight the importance of identifying and closing gaps in quality as a key next step.

Strategies to address:

a. Establish standardized measures to evaluate model performance across different languages and specific health contexts to ensure consistent quality.

b. Curate high-quality datasets for underserved languages, including region-specific dialects, culturally relevant health information, and voice data for low-literacy populations.

Improve technical capacity & shared infrastructure

Experts noted that technical capabilities of GenAI implementers varied dramatically; similarly some funders and health system leaders identified gaps in their own knowledge that, if addressed, would allow them to make more impactful funding and procurement decisions. They also noted that some technical barriers (e.g., language models) likely would be better addressed centrally vs. in a fragmented way.

Strategies to address:

a. Identify elements of technical infrastructure that should be shared, and establish ways to centralize these efforts.

b. Provide technical capacity and consulting expertise to health system leaders, funders, and implementers.

Improve digital & basic health infrastructure

Throughout our research, the risk of inadvertently perpetuating the digital divide emerged as a key concern: no matter how advanced AI models and datasets become, their potential to effect behavior change is wasted if the people who need them most cannot access the necessary digital or physical infrastructure (for example, lack of stable internet connection or access to healthcare facilities recommended by GenAI chatbots).

Strategies to address:

a. Prioritize investment in basic healthcare infrastructure alongside digital interventions.

b. Consider whether GenAI is the highest impact way of addressing your use case, taking into account existing basic healthcare and digital infrastructure.

c. Evaluate an organization’s digital readiness before deploying AI tools to avoid avoidable costly failures, and first focus funds on ensuring digital readiness where needed.

INTRODUCTION

There has been a rapid rise in the use of GenAI since the launch of Chat Generative Pre-trained Transformer (ChatGPT) in November 2022, with an array of emerging use cases in healthcare already in the implementation phase. GenAI has the potential to improve health and healthcare in LMICs on an unprecedented scale. Yet, we are at an early stage, with an urgent need for cross-sectional collaboration to address barriers to realizing maximal impact. Implementing such technologies in health contexts requires specific considerations and nuances, with the risk of potential harms, but also potential benefits, amplified, compared to other sectors.

> *“The big question mark that remains is, are the risks associated with using a GenAI based tool outweighed by the benefits of what you can now achieve?” - Bilal Mateen, Chief AI Officer, PATH*

This white paper has a particular focus on the use of GenAI tools to drive health-related behavior change (HBC), with our selected case studies illustrating use cases within HBC. Well-designed HBC interventions can empower individuals to make informed decisions to improve health. Equally, ill-considered interventions risk exacerbating existing disparities stemming from lack of supporting infrastructure.

> *“The ability of Gen AI to be much more nuanced and talk much more directly to the user's specific question is really exciting ... I think that's going to be a real step change” - Isabelle Amazon-Brown, The MERL Tech Initiative*

> *“One of the other opportunities for Gen AI and health behavior change is that AI doesn't get frustrated or tired ... it will keep having that conversation with that person and answer all their questions, and it will never act as if it's getting bored of the conversation.” - Shawna Cooper, Principal Product Manager, Audere.*

Our scoping analysis, framework and key recommendations are inclusive of a range of health use cases, not limited to behavior change, to contextualise the HBC interventions within the wider ecosystem, and facilitate broadly applicable learnings.

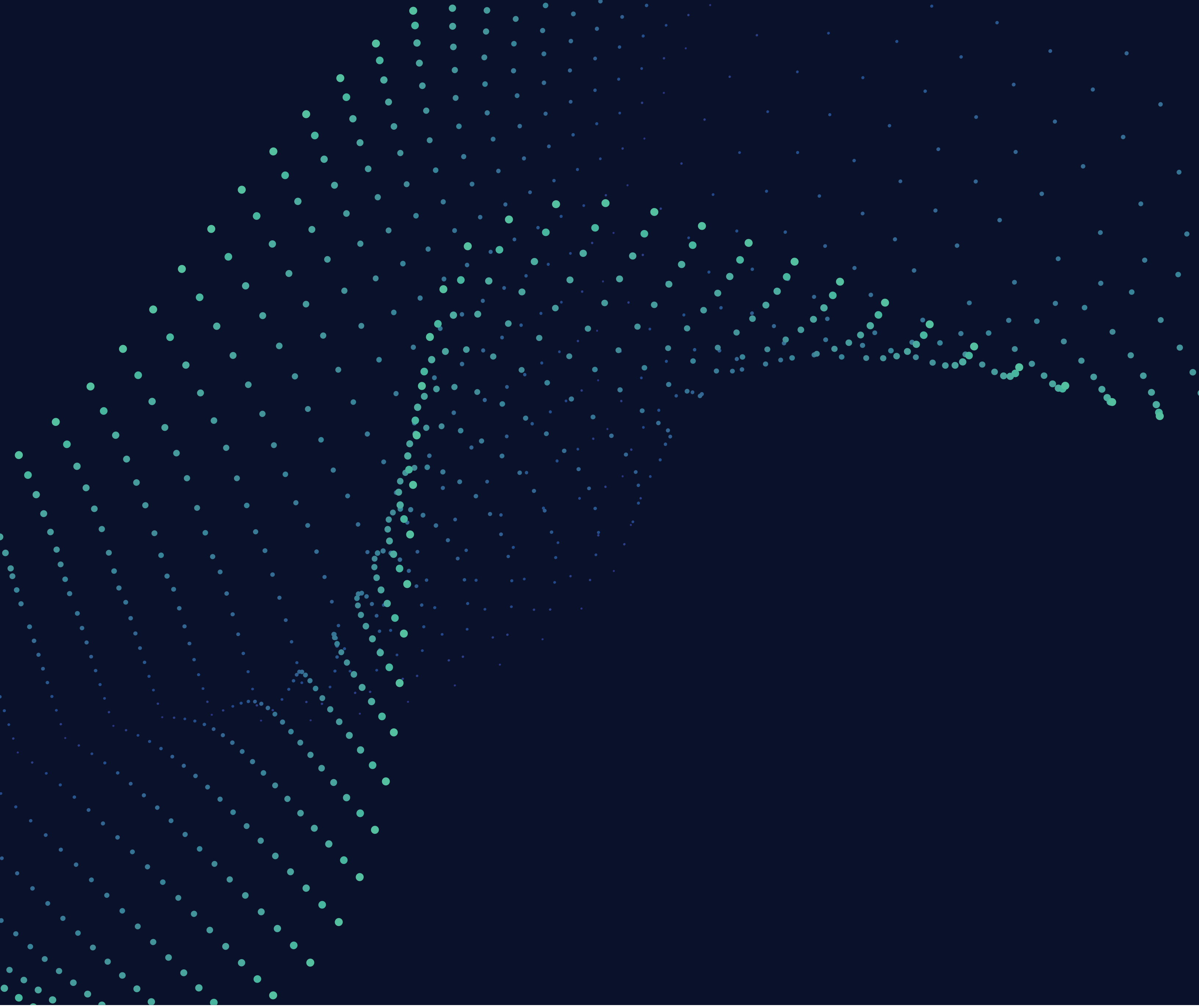
Alongside five case studies of HBC applications, we present our learnings from an analysis of key GenAI accelerator programs; a quantitative survey with 145 respondents; two roundtable events; and 24 in-depth qualitative interviews with experts in Generative AI and digital health, encompassing perspectives from academia, funding bodies, implementers and health system leaders.

STRENGTHS OF LARGE LANGUAGE MODELS

LLMs have been shown to perform significantly better than previous AI approaches in specific task areas. Technologies are advancing rapidly, but some currently validated task domains with healthcare-specific examples are outlined in Table 1. The most successful LLM implementations require identifying domain-specific use cases that map well onto these potential strengths.

LLM Task	Definition	Healthcare-specific example
Summarization	Condensing content into shorter summaries	Summarizing long medical guidelines into succinct summaries for immediate ‘in-clinic’ use.
Classification	Assigning labels or categories to content	Categorizing incoming patient messages in an online healthcare portal into categories such as medical versus administrative queries, to facilitate more efficient handling of queries.
Extraction	Identifying and retrieving information from a larger body of content	Identifying and extracting salient data points such as diagnoses, medications, and test results from patients’ medical records.
Translation	Converting content from one form to another, across languages, formats, or styles	Rewriting content to match different tones or styles, e.g. transforming clinical documents into patient-facing material.
Conversation	Engaging in dynamic, context-aware exchanges	A health chatbot providing real-time, personalized responses to user questions.

Table 1: Key LLM Task Domains



KEY LIMITATIONS OF THIS REPORT

While this review provides what we believe to be the most comprehensive analysis of GenAI applications in health-related behavior change to date, it is important to acknowledge its limitations. Our findings are based primarily on deployments funded by GenAI accelerator programs and insights gathered through snowball sampling interviews. As a result, while we believe this report is representative of the current landscape in global public health, it is possible that some deployments—particularly those outside of these funding networks—have not been captured. Additionally, this paper did not engage directly with end users, meaning that their perspectives on the usability, impact, and challenges of these interventions are not reflected. Further, its findings may not fully reflect commercial use cases, since its primary focus is grantmaking in global health.

We envisage this report as a starting point rather than a definitive catalogue of successful applications, and we encourage stakeholders who are aware of additional large-scale or high-impact deployments, as well as those who can contribute user-centered perspectives, to share their insights. Continued collaboration and knowledge-sharing will be essential in tracking the evolution of GenAI deployment for health behavior change.

SCOPING ANALYSIS

How widespread are GenAI deployments for health currently?

Our scoping analysis encompasses 285 grants from 14 accelerator programs sponsored by 10 funding organizations, funding 279 projects (please see Appendix for the list of included programs). These projects cover a diverse spectrum of health areas and use cases in LMICs across the globe.

Key inclusion criteria:

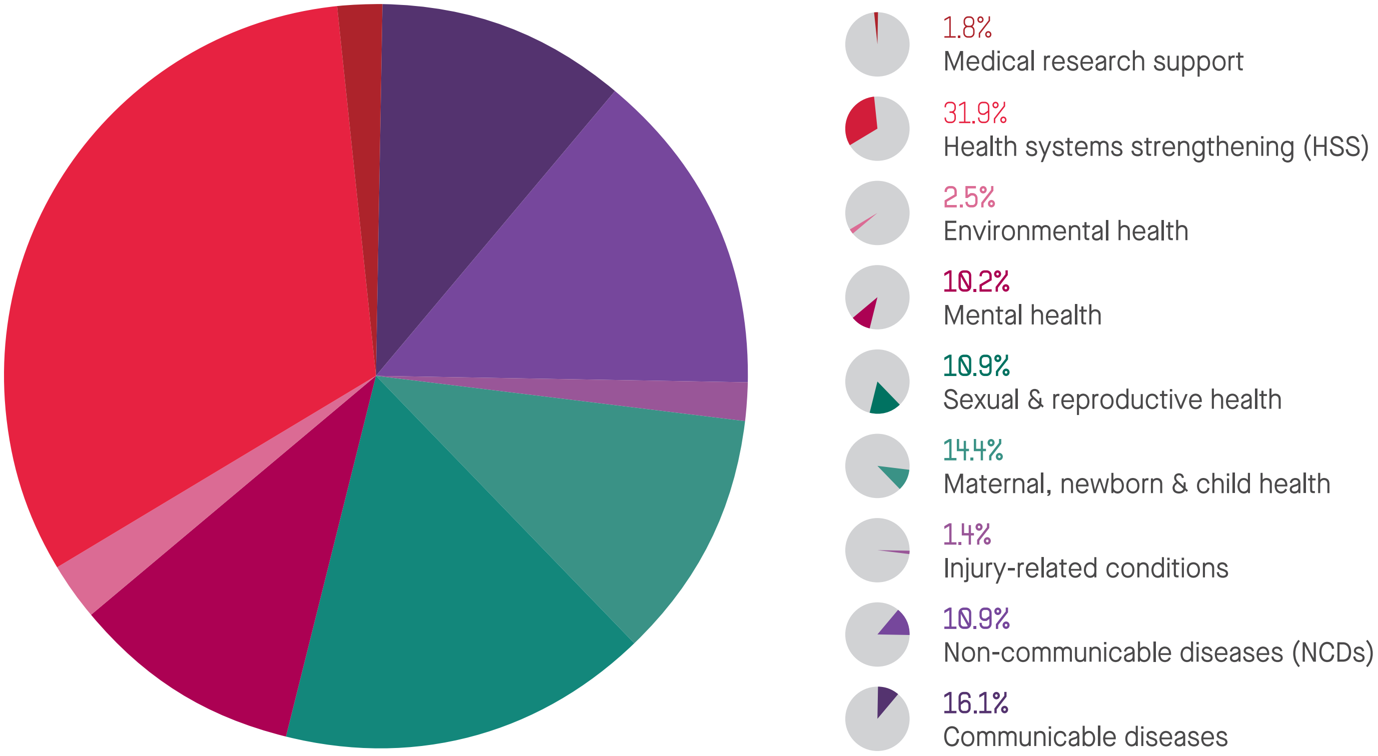
- Accelerator funds GenAI or LLM projects in health for LMICs
  - Accelerators whose projects had broader scope (e.g. AI outside of GenAI/LLMs, or topics beyond health) were included as long as they had at least 1 project that utilized GenAI or an LLM for health in an LMICs.
  - Accelerator programs with sufficient public data available to enable classification were included. Although OpenAI did not have public data available, we worked closely with their team to secure relevant project details given their very substantial contributions to the current funding landscape.
  - Projects which were inclusive of both high and low- and middle- income settings were included.

Classification of projects:

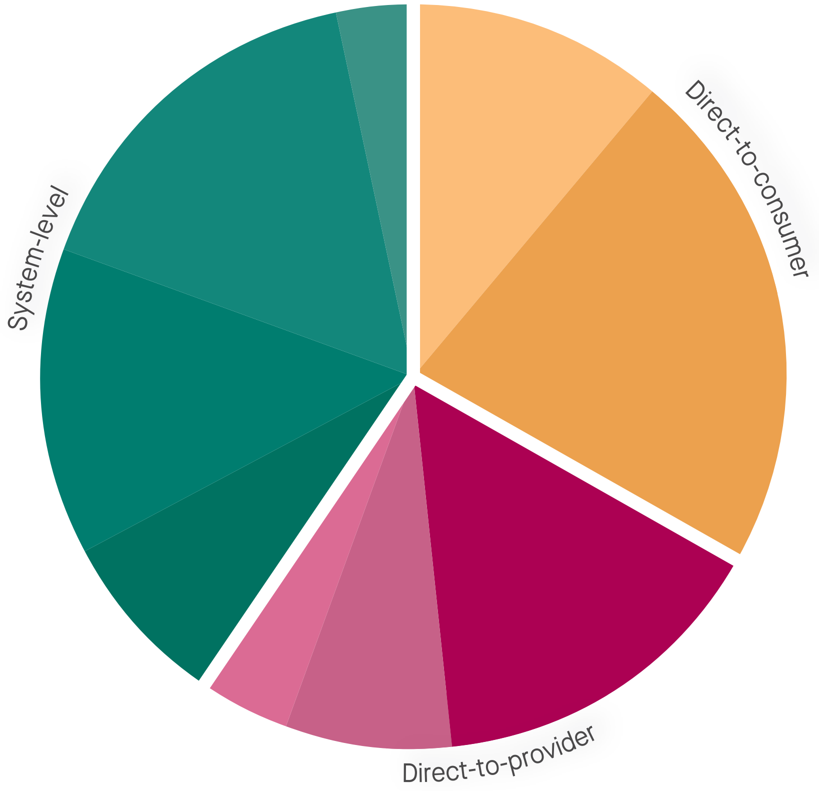
- By reviewing the established global health literature and piloting 50 projects from Gates Global Grand Challenges, we developed a taxonomy to categorize projects by use case and health areas described below. Although in some instances projects encompassed multiple different use cases or health areas, the project was classified according to its primary focus.

Use cases typically fell into one of three categories: direct-to-provider, direct-to-consumer, or system-level. Although direct-to-provider interventions will always involve human review, the degree to which 'human-in-the-loop' safeguards are implemented for direct-to-provider and system-level interventions varies depending on the use case. Despite the lack of current clear regulatory guidance, most of the use cases we reviewed currently incorporate a high level of human supervision.

The most commonly funded health area was health system strengthening, which included a variety of projects aiming to streamline access to or delivery of healthcare services generally. Communicable diseases was the most commonly funded disease-specific health area (16% of funded projects), followed by maternal, newborn and child health (14% of funded projects).



System-level interventions (public health surveillance, remote care, workflow optimization and healthcare system efficiency, and medical research support), were the most commonly funded group in terms of GenAI use case, representing a combined 41% of funded projects. This was followed by **direct-to-consumer interventions** (health education and awareness, and health-related behavior change), which represented a combined 37% of funded projects. **Direct-to-provider interventions** (clinical decision support, diagnostic and triage tools, and disease prediction and risk modeling) made up the remaining 22% of funded projects.



System-level

3.5%
Medical research support

16.1%
Workflow optimization & health system efficiency

13.3%
Remote care

7.7%
Public health surveillance

Direct-to-provider

3.9%
Disease prediction & risk modelling

7.4%
Diagnostic & triage tools

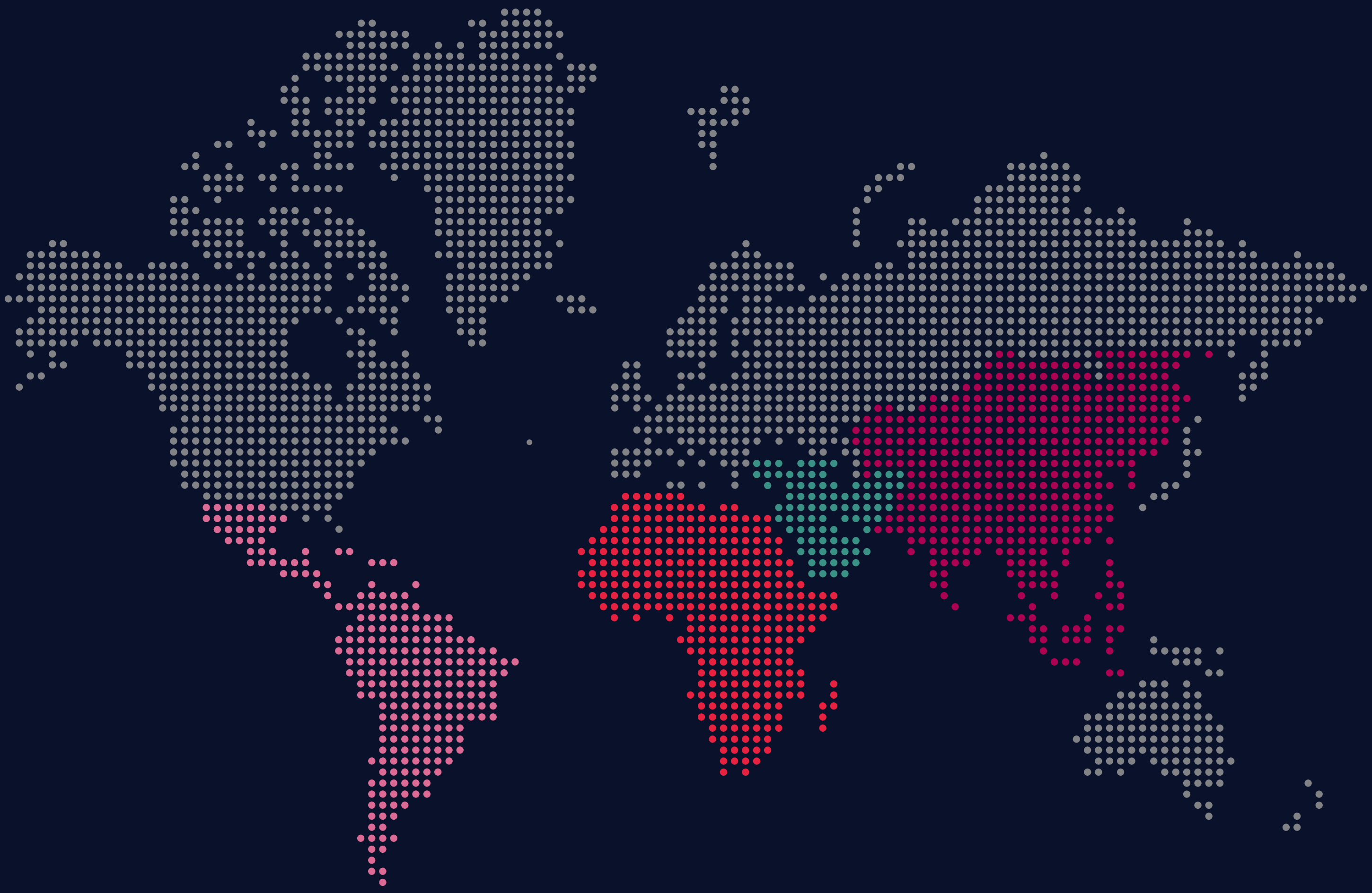
10.9%
Clinical decision support

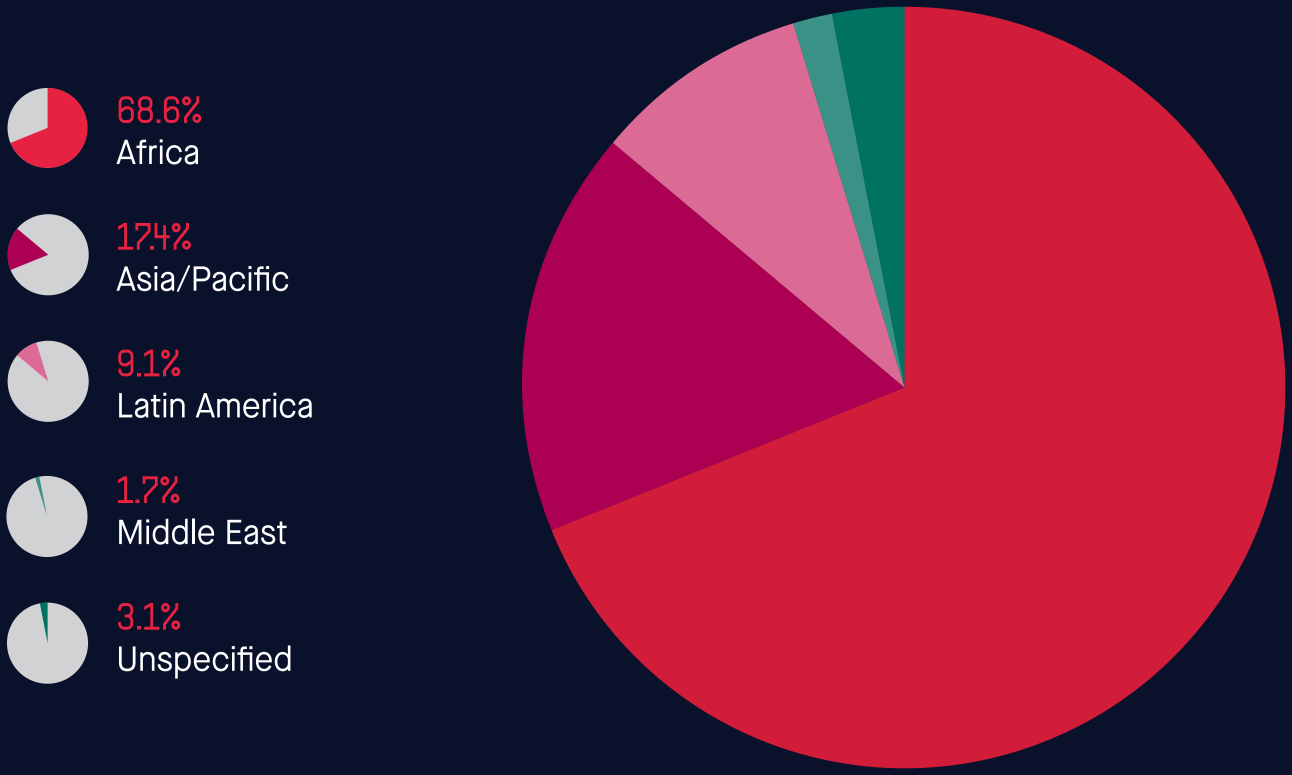
Direct-to-consumer

15.1%
Health-related behavior change

22.1%
Health education & awareness

The majority of funded projects were based primarily in Africa (68% of funded projects). This reflects Global Burden of Disease (GBD) data, with Sub-Saharan Africa having the lowest life expectancy of the super-regions, followed by South Asia,[9] and Sub-Saharan Africa also having the lowest coverage of essential health services.[10]





USE CASE CATEGORIES

Each intervention has been categorized according to the following taxonomies for Use Case and Health Area:

Direct-to-consumer

Health-Related Behavior Change: AI-driven systems that provide personalized recommendations or nudges to encourage healthier behaviors.
Examples: AI chatbots that give personalized advice on smoking cessation.

Health Education and Awareness: AI-based tools designed to educate populations or raise awareness about specific health topics or access to healthcare.
Examples: AI chatbots for reproductive health education.

Direct-to-provider

Clinical Decision Support: AI tools that assist healthcare providers in making better clinical decisions or answering patient queries.
Example: AI system that assists health worker responses to consumer questions via more efficient triaging and routing of questions and suggested answers.

Diagnostic and Triage Tools: AI systems that help in diagnosing diseases, or triaging patients for urgent assessment.
Example: machine learning models to diagnose tuberculosis based on X-rays.

Disease Prediction and Risk Modeling: AI tools that predict individual health risks and outcomes.
Example: predicting an individual patient's risk of developing diabetes.

System-level interventions

Public Health Surveillance: AI tools that predict disease outbreaks or perform public health surveillance.
Examples: early warning systems for infectious disease outbreaks.

Remote Care: AI-enabled platforms that support remote health consultations, monitoring or care delivery.
Example: systems enabling virtual consultations between patients and healthcare providers.

Workflow Optimization and Health System Efficiency: AI tools designed to streamline healthcare operations, optimize workflows, and improve the efficiency of healthcare delivery systems.
Example: tools for improving supply chain logistics for medications.

Medical Research Support: AI tools that optimize medical research such as clinical trials.
Example: a tool that can streamline clinical trial recruitment by identifying suitable participants.

HEALTH AREA CATEGORIES

Communicable Diseases: includes infectious diseases such as HIV/AIDS, tuberculosis, malaria, viral hepatitis and neglected tropical diseases.[11]

Non-Communicable Diseases (NCDs): also known as chronic diseases; the main types of NCD are cardiovascular diseases, cancers, chronic respiratory diseases and diabetes.[12]

Injury-related conditions: covering health issues that result from trauma, violence, road traffic accidents and occupational hazards.

Maternal, Newborn, and Child Health (MNCH): addresses pregnancy, childbirth, neonatal care, child nutrition, and the prevention of maternal and child mortality.

Sexual and Reproductive Health (SRH): includes access to contraception, sexual education, prevention and treatment of sexually transmitted infections, and protection from gender-based violence (GBV).

Mental Health: Mental health conditions include depression, anxiety and psychosis, as well as neurological and substance use disorders.[11]

Environmental Health: includes health risks related to environmental factors such as air pollution, climate instability, access to clean water, sanitation and hygiene, and exposure to hazardous chemicals.[13]

Health Systems Strengthening (HSS): includes the strategies, responses, and activities that are designed to sustainably improve the performance of a health system,[14] including enhancing healthcare infrastructure, policy implementation, and workforce training. Note that while HSS interventions are often system-level use cases, this is not always the case. For instance, a chatbot delivering information about available healthcare services would be classified by use case as direct-to-consumer, and by health area as HSS. Conversely, a public health surveillance tool focusing on cervical cancer would be classified by use case as system-level, but by health area as non-communicable diseases (not HSS).

Medical Research Support: Tools that optimize medical research, such as facilitating clinical trial recruitment or streamlining ethical review board processes.

Photo Source: Girl Effect

SURVEY RESULTS

To support our scoping analysis and qualitative interviews, our research partner, the Bay Area Global Health Alliance, sent out a survey to 572 individuals on its 2024 AI and Global Health Discussion Series email list, encompassing perspectives from research, funding, and implementation. We received 145 responses (25% response rate). Respondents selected their primary professional role in relation to the use of GenAI as follows:

Health Implementer
Part of organization responsible for creating or implementing health programs (37%)

Tech Facilitator
Part of technology platforms or business solutions providers that support technical implementation (19%)

Academic/Researcher

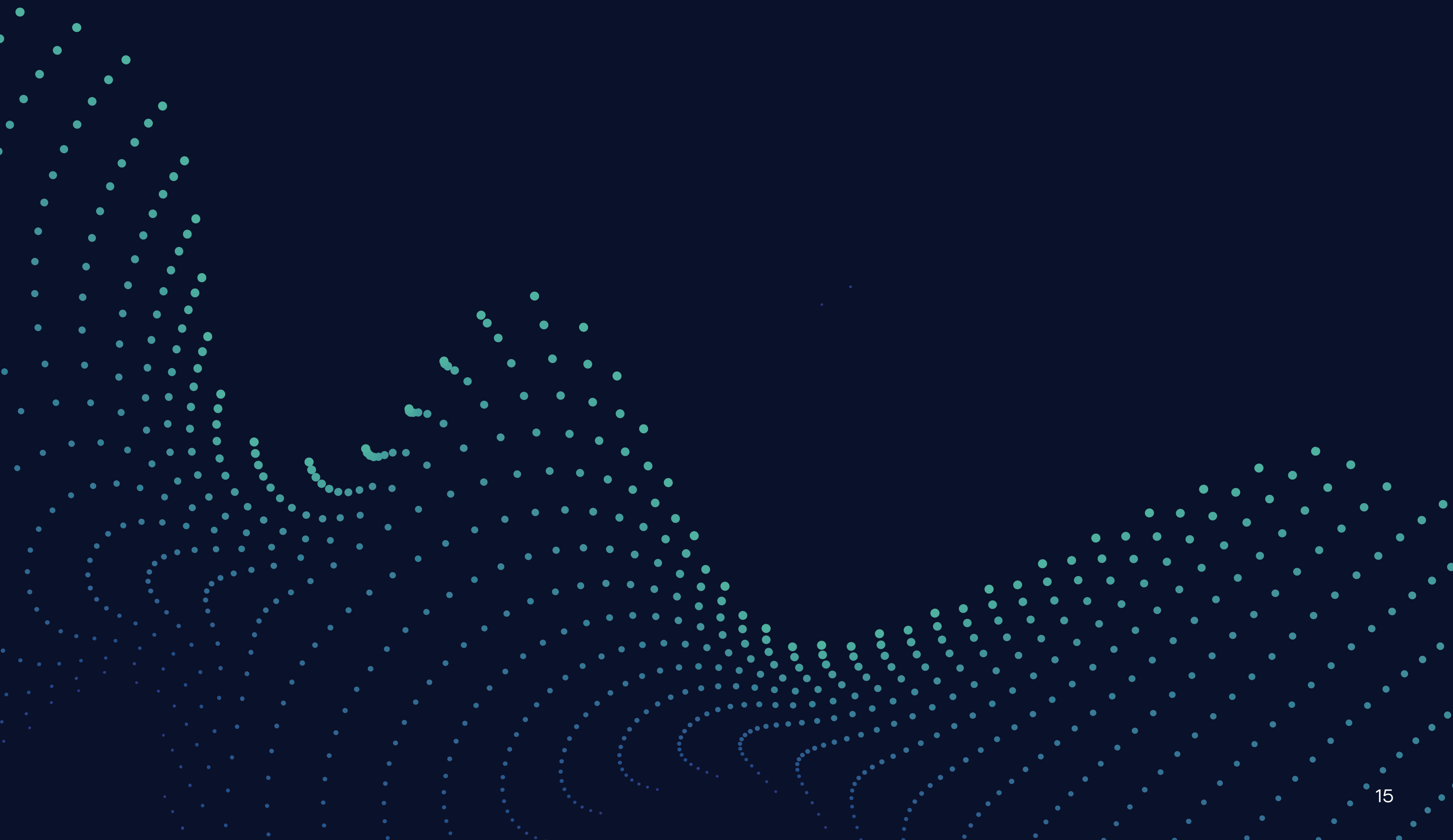
Part of organization that conducts research on health programs (18%)

Health Funder
Part of organization responsible for funding the implementation of health programs; can include NGOs or government (12%)

Health System Expert

Part of organization that delivers health care services, including running hospitals and clinics (5%)

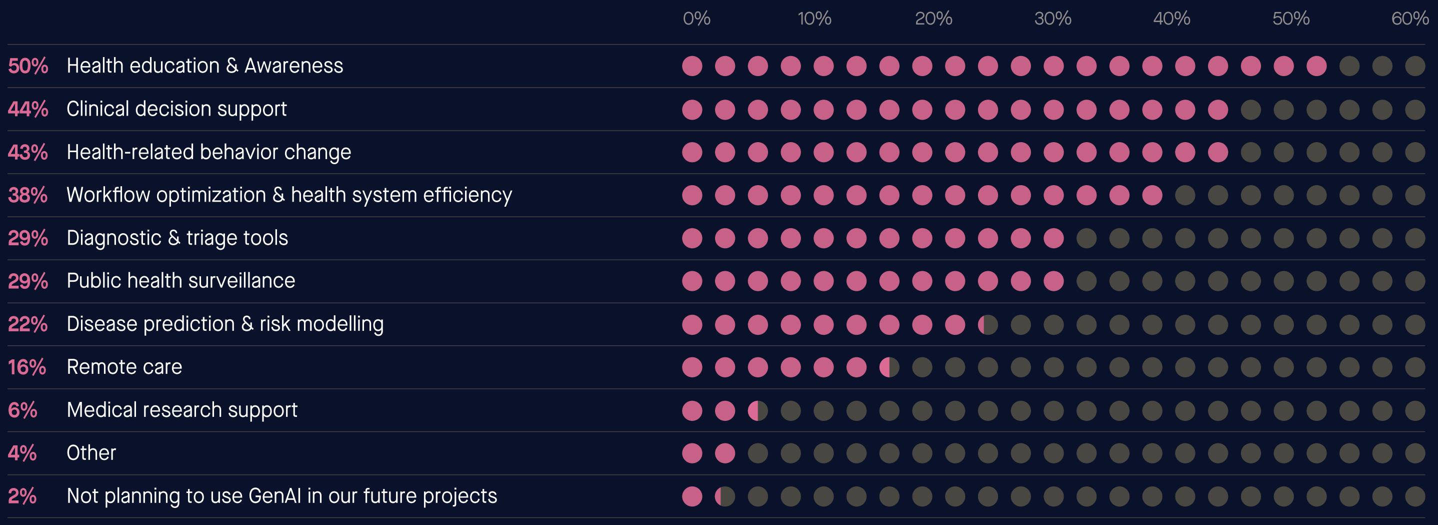
Fifty-one percent of respondents were actively involved in projects or initiatives using GenAI for health in LMICs, with 91% of these projects including work in Africa, and 30% including work in Asia/Pacific (note that many projects include multiple regions of implementation).



WE ASKED THE FOLLOWING QUESTIONS ABOUT THE USE OF GENAI FOR LMIC HEALTHCARE:

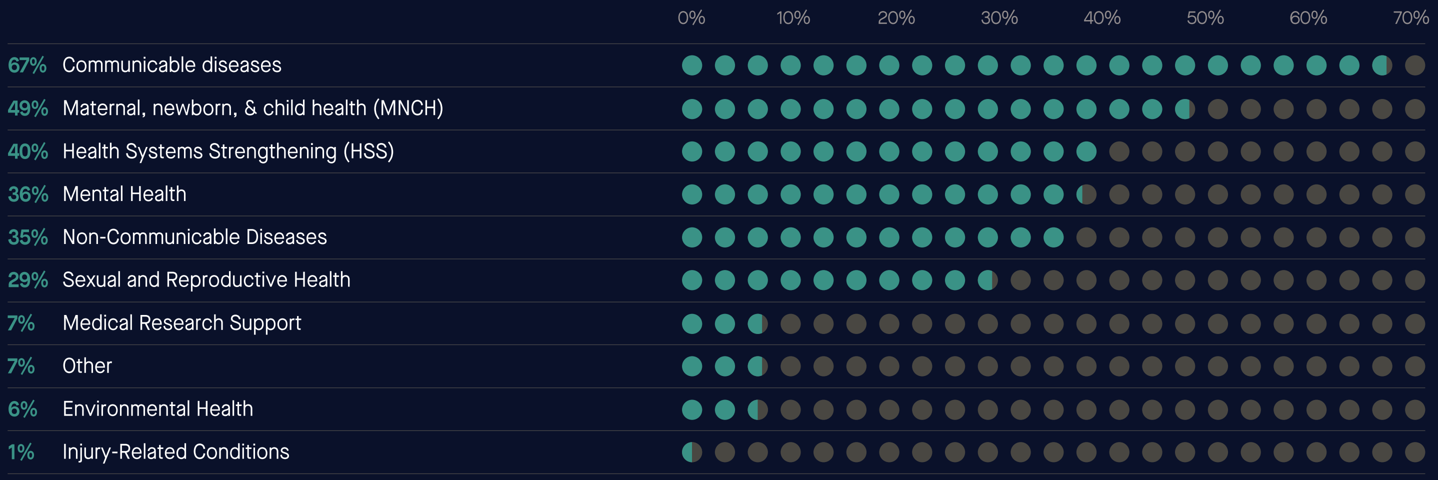
What are the primary use cases you would like GenAI to be used for in a healthcare setting in LMICs?

We asked respondents to select their top 3 priority use cases. The most popular response was Health education and awareness (selected by 50% of participants), closely followed by Clinical decision support (44%) and Health-related behavior change (43%).



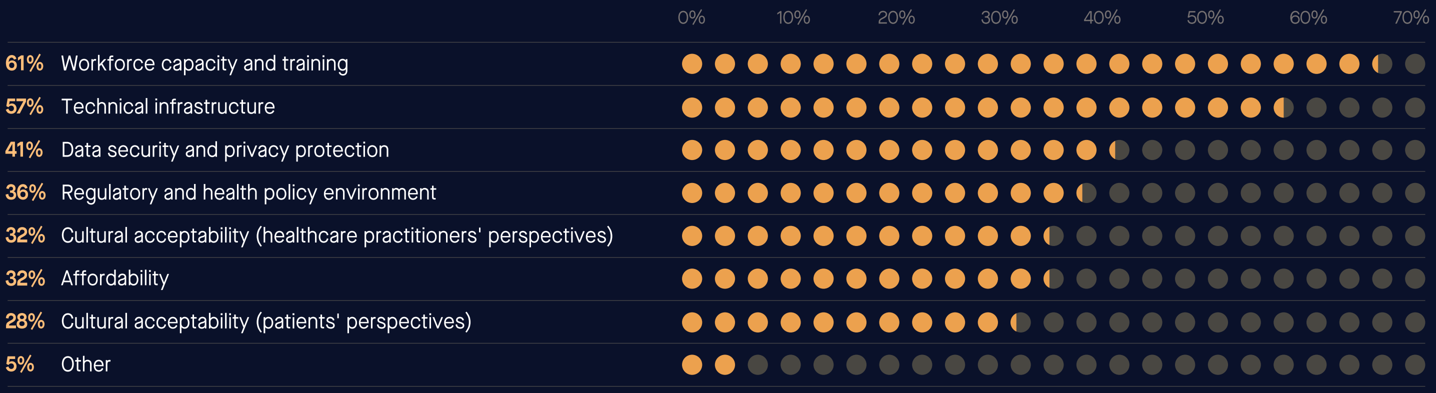
Which health areas do you think should be prioritized for the use of GenAI in healthcare settings in LMICs?

We asked respondents to select their top 3 priority health areas. Communicable diseases was the most popular response (67% of respondents), followed by Maternal, newborn and child health (49%), Health systems strengthening (40%), Mental health (36%), Non-communicable diseases (35%) and Sexual and reproductive health (29%).



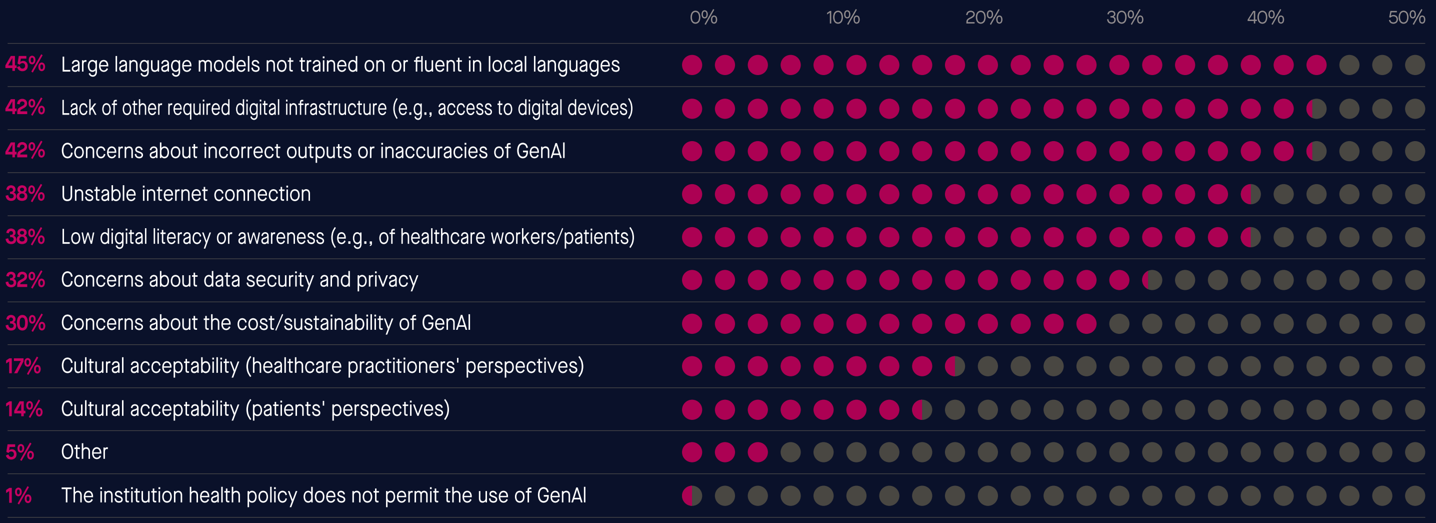
What do you think are the most important factors for ensuring successful use of GenAI for health in LMICs?

In terms of the most important factors for ensuring successful use of GenAI, the majority of participants selected ‘Workforce capacity and training’ (61%), followed by ‘Technical infrastructure’ (57%). ‘Data security and privacy protection', and ‘Regulatory and health policy environment’ were selected by 41% and 36% of respondents respectively. Slightly more respondents felt that cultural acceptability from healthcare practitioners (32%) was an important factor than cultural acceptability from patients (28%).



What are the key barriers to using GenAI in healthcare settings in LMICs?

In terms of key barriers to use, ‘Large language models not trained on or fluent in local languages’ was the most commonly selected response (45%), followed by ‘Lack of other required digital infrastructure (e.g., access to digital devices)’ and ‘Concerns about incorrect outputs or inaccuracies of GenAI’ (both 42%). ‘Unstable internet connection’ and ‘Low digital literacy or awareness’ were both selected by 38% of respondents as key barriers. Fewer participants were concerned about cultural acceptability as a potential barrier to use.



FRAMEWORK TO GUIDE THE USE OF GENERATIVE AI FOR HEALTHCARE IN LOW- AND MIDDLE-INCOME COUNTRIES

Aiming to synthesize a diverse range of perspectives from research, policy, funding and implementation, across the public and private sectors, we highlight four key principles and four key risks of GenAI health interventions.The recent FUTURE-AI framework delineates broad ethical principles and considerations for the development and deployment of trustworthy AI tools in healthcare, covering the lifecycle of healthcare AI.[15] Our recommendations complement this by offering deeper implementation insights and actionable strategies, supported by real-world case studies.

KEY PRINCIPLES

Prioritize user-centered design

GenAI solutions must be tailored to the needs of their end-users—whether healthcare workers or patients. Partnership with community organizations is crucial to ensuring the needs and concerns of target end-users are heard and integrated throughout the design and implementation process. Interventions should play to the particular strengths of LLMs, map onto local priorities, and take into account the level of contextual risk given the use case.

> *"One of the best ways we can mitigate some of the risk is to authentically co-develop solutions with the communities where we want them ... sitting with the people who are going to be using these solutions." - Stanford Workshop*

Define and implement evaluation frameworks

Although existing health-specific outcome measures are broadly applicable across digital health interventions, with morbidity and mortality remaining key, we currently lack established standards for measurement and benchmarking specific to GenAI tools. Measuring success starts with clearly defining goals and outcome metrics upfront. Going forward, utilizing consistent outcome metrics to describe scale of projects (such as monthly active users, total users and retention of users) and specificity regarding the type of AI system being used (for example deterministic vs. generative) will facilitate meaningful comparisons and benchmarking. For tools using LLMs to generate responses to queries (whether from a healthcare worker or user), metrics of interest are likely to include correctness and completeness of LLM responses and time- and cost-savings compared to previous non-GenAI methods, as well as qualitative factors such as understandability, empathy, and appropriateness of tone and style. Unlike a pharmaceutical compound, GenAI outputs are constantly evolving, so evaluation and review processes must be continuous to ensure tools remain accurate, relevant, and aligned with local health guidelines.

> *"It's changing under us constantly, every minute of the day ... So we need to make sure that if we're using it for healthcare ... we are putting it through its paces on a regular basis" - Stanford Workshop*

Cost-effectiveness is a key evaluation metric in resource-constrained contexts where healthcare systems face significant financial and operational challenges. Evaluations should consider not only the upfront costs of implementing GenAI tools but also the cost-benefit analysis involved in improving healthcare access and outcomes.

> *"The moment you start introducing services that give more access to patients who need healthcare, your costs go up. So the first thing that we saw when we started using digital tools to increase patient loyalty in their care journey, the cost went up. However, the cost went up a little bit, and the quality went up big time. So there's a cost efficiency in the improvement of health outcomes." - Nicole Spieker, Chief Executive Officer, PharmAccess*

Balance safety and potential benefits

The principle of "do no harm" is deeply embedded in healthcare. Yet while safety remains paramount, it is important to consider the opportunity cost of not using GenAI to address unmet health needs, particularly in LMICs. Pursuing perfection with GenAI tools is futile, and disproportionate given existing human error rates in healthcare.

> *"Perfection is the enemy here ... because even when talking to large companies that are based here [in LMICs], I don't think they visibly understand how bad the next best alternative is for many people" - Stanford Workshop*

THE EVOLVING LANDSCAPE OF AI REGULATION

The EU Artificial Intelligence Act is the first comprehensive regulatory framework for AI globally, coming into force on August 1, 2024.[16] It regulates AI based on risk levels: unacceptable-risk systems, like social scoring and manipulative AI, are prohibited. High-risk systems (which typically include healthcare applications) face strict requirements including risk-mitigation systems, high-quality data sets, clear user information and human oversight. Limited-risk systems, such as non-health-focussed chatbots, must meet transparency requirements. Minimal-risk systems, like AI in video games, are mostly unregulated, though this may change with advances in GenAI.

While the EU AI Act sets a high bar with binding, risk-based regulations, AI governance frameworks are rapidly emerging in LMICs: a growing number of countries in Africa, South Asia and South America have released national AI strategies, with others in development. However, these often focus more on strategic development and ethical guidelines rather than enforceable legal requirements. Examples include:

Nigeria's National Artificial Intelligence Strategy, released in 2024, outlines ethical principles and governance priorities, including data protection and responsible AI use.[17] However, it stops short of establishing binding regulations or specific compliance mechanisms, focusing on guiding principles for AI development in sectors including healthcare while placing reliance on existing legislation not specific to AI, such as the Nigerian Data Protection Act (NDPA).

Kenya's Draft National AI Strategy, developed in collaboration with German and EU partners, similarly promotes responsible AI through ethical guidelines, data governance frameworks, and capacity-building initiatives.[18] Whilst it acknowledges the need for regulatory oversight in sensitive areas such as health, it currently lacks enforceable regulatory provisions, instead emphasizing a roadmap for future legislative development.

India's National Strategy for Artificial Intelligence (#AIforAll) aims to foster AI innovation across key sectors, including healthcare, through ethical guidelines and policy recommendations. However, India does not yet have a dedicated AI regulatory framework; governance relies on existing data protection and IT laws, with AI-specific regulatory discussions still ongoing.[19]

Brazil's National AI Strategy (EBIA) provides broad principles for AI ethics and responsible innovation.[20] Recent proposals include investments to support AI governance, but like many LMIC strategies, Brazil's approach focuses on strategic direction and sector-specific guidelines rather than comprehensive AI-specific regulations.

Striking this balance can be a challenge given there is not a clear regulatory landscape internationally, and regulators in many LMICs lack the resources to govern digital health tools effectively. Ideally forthcoming regulation will foster experimentation and implementation in lower-risk applications to maximize benefits, while putting in appropriate controls in higher-risk settings.

> *"The most important thing we can do amidst all the hype and flashy tools, is ensure we are leading with evidence. It's true that generating such evidence is not straightforward and we are largely in uncharted territory but this is no excuse for not ensuring we do that hard work. This is also likely to challenge our current models of regulatory approval and post launch surveillance, but this too, shouldn't slow down the enormous upside the tools have for communities in the Global South. The challenge is worth taking on!" - Zameer Brey, Deputy Director, Technology Diffusion, Gates Foundation*

Ensure collaboration, transparency & knowledge-sharing

The need for cross-organizational collaboration and sharing of experiences and learnings was stressed consistently throughout workshop discussions and interviews alike. Even for those working at the forefront of innovation in this arena, there is much that is not yet well understood, and the speed at which the field is evolving entails considerable uncertainties along the way. Failures are inevitable, but are also valuable opportunities to learn—transparency and knowledge sharing can help us avoid repeating the same mistakes.

> *"We have to waste time, but let's do it quick. Let's learn, let's iterate. There's going to be a ton of failures, and that's okay, but if ... there's no cross-organizational learning ... that is one of my biggest fears of AI" - Stanford Workshop*

> *"A lot of the lessons learned on the digital health side were hard, hard-won lessons that we're still working to try to implement. And by breaking the AI out independent of the rest of the digital health space, they're going to relearn all of those lessons again. So rather, you encapsulate it within it and use the same sorts of approaches and metrics that we would for any sort of digital intervention—that's the biggest approach we want to take." - Merrick Schaefer, Director of the Center for Innovation and Impact, USAID's Global Health Bureau*

Government buy-in is crucial in successfully scaling innovations, particularly in public health systems that serve diverse and widespread populations.The complexity of decentralized systems—where federal, state, and local authorities often have separate but overlapping roles—means any scaling effort requires engagement across multiple layers of governance. Achieving alignment across such diverse entities takes time, but lays the groundwork for sustainable and equitable implementation.

> *"The government ownership was critical. I think when you're looking at scale, you're talking of each and every public health facility, which means the first mile to the last mile, and in any public health system in any middle-, lower middle-income country, typically, 90% of these health facilities are sub-district level, and focus usually ends up being at the national or the state level hospitals, but really it's the primary and secondary health centers which are the first contact for these communities that they serve, where the system really works." - Manish Pant, Policy Specialist, Digital Health, UNDP [On his work digitizing vaccine supply chains in India in 2015]*

RAISE HEALTH INITIATIVE

Stanford University's RAISE Health (Responsible AI for Safe and Equitable Health) initiative is dedicated to advancing responsible AI innovation in biomedicine by fostering collaboration across organizations and sectors. Convening stakeholders from academia, government, and industry, RAISE Health promotes knowledge-sharing on best practices and challenges in AI development and implementation. Through its forums and collaborations, RAISE Health aims to ensure that AI technologies are designed and deployed with safety, accountability, and inclusivity in minds.[21]

> *"We launched RAISE Health because we recognize no single organization can chart the future of AI in biomedicine alone. To build responsible AI solutions, it is imperative that we collaborate across sectors, ensuring that every voice and interest is part of the process." -Lloyd Minor, MD Dean of the Stanford School of Medicine and Vice President for Medical Affairs, Stanford University*

Photo Source: Jacaranda Health

KEY RISKS

Even when embracing the above key principles, risks associated with the use of GenAI tools for health will remain and must be considered throughout design, implementation and evaluation phases.

Model-based risks

Inaccuracies

Inaccuracies or "hallucinations" (outputs that sound plausible but are incorrect) are a universal flaw in current GenAI algorithms. To mitigate potential harms associated with inaccurate outputs, it is important to establish an acceptable error rate for a given use case, integrate human oversight, and employ strategies like Retrieval Augmented Generation (RAG) to improve contextually accurate responses.

> *"Any sort of machine learning tool is never going to be 100% correct ... Different applications have different error rates that are acceptable. And in some applications, it's more severe if you get things wrong than others" - Stanford Workshop*

Cost and environmental impact

Large Language Models (LLMs) are resource-intensive. Costs vary by language, with non-Western languages generally requiring more processing power due to higher token density. Additionally, GenAI systems are energy-consumptive, raising environmental concerns.

Data security and privacy

While this issue goes beyond the intrinsic risks associated with GenAI, there is a specific model-based concern regarding the potential of GenAI chatbots to elicit large amounts of personal information more readily than in current clinical settings. This raises concerns about storage practices and the risk of misuse, particularly in contexts where a given health-related behavior (such as same-sex sexual activity) is stigmatized or criminalized.

Limitations of the training data

GenAI models are only as good as the data they are trained on. Existing datasets are heavily skewed towards European languages and cultural contexts. Training LLMs requires vast amounts of written text in the target language, and for many languages globally this may be challenging to acquire. Other challenges arise from differences between cultural contexts, which may lead to LLMs misinterpreting local slang expressions. This can have profound consequences in the context of medical communication, from damaging trust and rapport in provider-patient relations to missing key health information. Additionally, the available data for a given language may have inadequate coverage for particular health areas—for example, in places which lack gender parity, women's health concerns may be overlooked or trivialized in existing language datasets.

> *"If you ask a chat model in Swahili, 'My baby has this high fever, what should I do?' Chances are the model will tell you to go to your local pediatrician, which is just not a thing, right? Nobody has a local pediatrician ... you've offered essentially garbage information to someone who maybe just needs to call a nurse, or go to the local clinic. But either we'll spend a lot of money to go to a pediatrician, or we'll just disregard the message and say, 'Well, I can't do that' ... That's the appropriate message to give to someone who lives in North America ... if you have access to care, but it is not the appropriate message to give to someone else." - Sathy Rajasekharan, Co-Executive Director of Jacaranda Health*

> *"A user asked about an empty chest ... GPT-4 interpreted that one as depression, and it turned out that the person was having chest pains ... And that's a common way ... people [in Nigeria] might say chest pain" - Stephen Meyer, Director of Partnerships, Viamo*

We welcome the announcement in February 2025 by the AI for Development Funders Collaborative (a global partnership including FCDO, IDRC, BMZ, and the Gates Foundation) of $10 million towards the development of AI models that are inclusive of African languages.[22]

Digital divide & lack of basic health infrastructure

No matter how advanced AI models and datasets become, their potential to effect behavior change is wasted if the people who need them most cannot access the necessary digital or physical infrastructure. The issue of digital access is particularly urgent: in 2022, the World Bank reported that only 36% of people in Africa had broadband internet access.[23] Many low-income settings also lack the computing capacity needed to train and deploy advanced AI models.[21] This forces reliance on external infrastructure, which can reduce local control and ownership of AI initiatives. Additionally, basic healthcare infrastructure often falls short. For instance, a chatbot that encourages young women to get the HPV vaccine will be ineffective if local clinics do not stock it. Even in regions with adequate digital and healthcare infrastructure, a lack of education, familiarity with digital tools, and trust in their use can lead to failed implementation.

> *“We too often just look at the intervention in isolation … What does infrastructure look like at the last mile? Are there power and connectivity challenges? Are the health workforce digitally literate? Are the tools well designed based on their needs, their user experience? There are all these factors that are outside the individual interventions that tend to have an outsize impact on whether these interventions are going to be successful or not.” - Sean Blaschke, Senior Health Specialist, UNICEF*

> *“Let's just take the elephant in the room: we assume that the end user's got some kind of device and we can reach them. And obviously, disproportionately, people without devices will be less capable of accessing health service anyway. And we need to keep that in mind, because it's very easy to just say, ‘Well, everybody's got mobile phones’, but that's clearly not the case.” - Gustav Praekelt, Co-Founder, Turn.io*

It is important to recognize how digital divide challenges interact with other societal dynamics and inequalities. For instance, while the gap is narrowing, women in LMICs are still 15% less likely than men to use mobile internet, with a more pronounced disparity in Sub-Saharan Africa (32%) and South Asia (31%).[24]

PATH INITIATIVES TO TACKLE DATASET LIMITATIONS IN SUB-SAHARAN AFRICA

Digital Square at PATH is spearheading initiatives to accelerate the effective use of LLMs for improving primary healthcare in sub-Saharan Africa through three targeted workstreams in 2025:[25]

1. **Developing localized datasets to address biases inherent in LLMs trained on data from high-income countries.** In collaboration with partners in Kenya, Nigeria, and Rwanda, PATH is gathering medical questions and answers to create datasets that reflect local disease burdens and medical practices. They will use the datasets to test the performance of existing LLMs, where an expert panel will compare LLM responses to responses provided by relevant medical experts. The datasets will be made publicly available for others to train and fine-tune LLMs. PATH is also supporting the development of the AfriMed-QA dataset: a multi-institution, open-source dataset of 25,000 Africa-focused medical question-answer pairs created to represent the disease burden and medical practices common across Africa.[26]

2. **Evaluating the accuracy, safety, and effectiveness of LLM-enabled clinical decision support systems for frontline workers in primary health care settings.** Clinical trials in Kenya, Nigeria, and Rwanda will assess tools ranging from voice-based call centers for community health workers to electronic medical record-integrated consult features for clinicians. These trials aim to measure impact on care quality, patient outcomes, and the tools’ appropriateness for diverse healthcare environments.

3. **Establishing a community of practice** to bring together stakeholders from technology companies, academia, donor organizations, and implementing partners to share lessons learned and prevent duplication of efforts. The community of practice will engage through monthly virtual meetings.

Reflecting societal biases and problems

AI systems have a powerful ability to reflect and amplify existing societal biases. These biases are not intrinsic to AI but arise from issues like skewed training data or the assumptions of those designing the systems. For example, even a well-representative dataset may inadvertently embed the biases of its developers. In contexts with entrenched cultural biases such as gender discrimination, the need for language- and context-appropriate LLM training data could result in local gender stereotypes being reinforced.

There is also a serious lack of consistency in standards for responsible AI reporting: the Stanford Institute for Human-centered AI (HAI) has found that leading developers including OpenAI, Google, and Anthropic test their models against different benchmarks.[27] A move towards standardization will enable easier interpretation of risks and optimal solutions.

Further, AI technologies can be exploited by bad actors, who are not bound by ethical guidelines or Responsible AI frameworks.

> *“A number of the things we're concerned about are systemic social problems that that you need to find different routes of actually addressing” - Stanford Workshop*

KEY RECOMMENDATIONS

Prioritize user-centered design

GenAI solutions must be designed with a clear focus on the needs and priorities of their end-users, whether healthcare workers or patients.

For Funders:

- Support co-design approaches by funding projects that actively engage local stakeholders and potential end-users in the design process.
- Ensure that funding priorities reflect the realities of on-the-ground healthcare workers.

For Implementers:

- Engage local partners such as healthcare workers, community leaders, and policymakers to co-design solutions that address local priorities.
- Ensure interventions are realistic and sustainable in resource-limited settings.
- Tailor safeguards to the deployment environment, considering varying levels of contextual risk for clinician-facing and consumer-facing use cases.

> *"Typically, our grants focus on supporting institutions that are very close to the issues that are at hand and that put people and communities at the center of the interventions" - Topaz Mukulu, Strategy Analyst, Patrick J McGovern Foundation grantmaking team*

Ensure robust error safeguards

While no system—human or AI—is error-free, careful planning and safeguards can help minimize these mistakes and their potential impact.

> *"We cannot guarantee that we have eliminated hallucinations from a particular project. So the goal then becomes to minimize those, or build some infrastructure around containing them". - Brian DeRenzi, VP, Research and AI, Dimagi*

For Funders:

- Support efforts to establish acceptable error thresholds and need for human oversight mechanisms for different healthcare applications.

For Implementers:

- Improve accuracy with design techniques such as prompt engineering and retrieval-augmented generation.
- Define appropriate error tolerances for the tool's purpose. For example, for a tool using GenAI to categorize incoming user queries by intent, false negatives (i.e. health queries misclassified as non-health) should be minimized at the cost of a higher false positive rate.
- Ensure appropriate human oversight: 'human-in-the-loop' is a practice increasingly emphasized in the field of AI, and a key feature of many current use cases where errors would be detrimental to patient care.
- Conduct regular monitoring of error rates, mindful of the need for continual review processes given the ever-changing nature of GenAI algorithms.

Share Learnings

Funders have a vital role to play in creating the incentives for knowledge sharing between organizations across the public and private sectors, whilst implementers have access to the most cutting-edge knowledge regarding successes and failures.

For Funders:

- Provide funding streams for collaborative approaches, and ongoing roundtable and workshop events to accelerate problem-solving around key challenges.
- Support regular processes for sharing and disseminating case studies, as demonstrated in this document.
- Support the production of practical guidance on how to identify LLM applications while mitigating risks and then pilot/validate/scale them. A regular update process will be required given technical capabilities are changing quickly.

> *"How do we create the right incentives for organizations to collaborate? ... Especially where things aren't working well ... which is not intuitive. It's not what they get rewarded for" - Stanford Workshop*

For Implementers:

- Define and use consistent outcome metrics to describe the scale of projects (such as monthly active users, total users, and retention of users) and specificity regarding the type of AI system being used (for example, deterministic vs. generative) to facilitate meaningful comparisons and benchmarking.
- Share implementation insights and lessons learned, particularly regarding challenges and barriers encountered, to inform future efforts for the field.

> *"Being able to capture the lessons that our partners have identified, whether that's challenges that they encountered, barriers or just lessons. And so I think success also means generating the evidence and learning that can inform future efforts, whether it's for your organization or just the field at large." - Topaz Mukulu, Strategy Analyst, Patrick J McGovern Foundation grantmaking team*

Define and enable actionable measurement

People wanted better ways of measuring benefits, costs, and risks, in ways that provide rigorous but also timely data to inform implementation decisions. Traditional evaluation methods such as randomized controlled trials (RCTs) can take years to provide actionable results, meaning we also need better ways to measure success to inform time-critical implementation and funding decisions in interim periods. Establishing a clear evidence base will also be essential for supporting government decisions to implement successful applications at a national scale.

For Funders:

- Identify opportunities and establish funding streams for implementer and academic partnerships to develop robust measurement and evaluation frameworks, leveraging implementers' access to data and academics' expertise on measurement.
- Require grantees to adopt standardized metrics for evaluating GenAI health interventions.

For Implementers:

- Develop continuous monitoring mechanisms to account for the evolving nature of LLM responses.
- Partner with academics to identify appropriate evaluation approaches to enable agile, real-time assessment.
- Prioritize transparency in reporting methodology, data sources, and key assumptions in evaluation.

> *"How can we do research studies going forward, where they're a bit more agile? Because we have a bunch of different system-defined prompts in our clinical decision support system, and we want to be able to iterate on those in real time. Sometimes it's a super small thing: we want to add an additional example to the prompt for how it should behave. And I think the current research paradigm wants us to set up the intervention, exactly as it is, and then freeze it like that for two, three, four months, while we run the trial. And in the meantime, we know we could improve this prompt. So I think being able to have more flexible research paradigms for the LLM age is important." - Robert Korom, Chief Medical Officer, Penda Health*

Improve language and localization

The effectiveness of GenAI tools in health behavior change hinges on their ability to communicate clearly and accurately across diverse languages and cultural contexts. However, the quality of models varies considerably by language, by medium (with voice particularly important for low-literacy settings) and by use case (e.g. health-specific contexts). There is a pressing need for ways to identify and close gaps in quality.

For Funders:

- Invest in the development of high-quality datasets for underserved languages, including region-specific dialects, culturally relevant health information, and voice data for low-literacy populations.
- Fund efforts to establish standardized measures to evaluate model performance across different languages and specific health contexts to ensure consistent quality.

For Implementers:

- Ensure LLM-generated content is accurate and culturally appropriate by involving local experts in the testing and training process.
- Test and train models on data relevant to the intended healthcare setting.

> *"There is potential for Gen AI to bridge that spoken language gap ... the danger, though, is that most of AI has been trained on English—and is there enough written material in some of these other spoken languages to really exploit the possibility?" - Stanford Workshop*

> *"I would hope more funders understand the need for building capacity locally to be able to tune and train models appropriately." - Sathy Rajasekharan, Co-Executive Director of Jacaranda Health*

Improve digital & basic health infrastructure

There is a risk that implementing GenAI tools could further exacerbate the digital divide in low-resource contexts where digital infrastructure is unevenly distributed. While GenAI tools can increase demand for health services by empowering users with better information and decision-making support, this must be accompanied by corresponding investment in the supply side of service delivery.

> *"LMICs need to invest in digitizing care. Otherwise, we're not going to be able to take advantage of this" - Robert Korom, Chief Medical Officer, Penda Health*

For Funders:

- Prioritize sustained investment in foundational healthcare systems alongside digital initiatives—both are essential for meaningful, equitable progress.
- Prioritize GenAI investments that complement existing healthcare systems.

For Implementers:

- Evaluate whether GenAI is the highest-impact way to address a given use case, considering existing healthcare and digital infrastructure.
- Assess organizational digital readiness before implementing AI tools, using tools such as the Global Digital Health Monitor.[28]
- Design with maximal inclusivity in mind, considering how to reach individuals without smartphones and internet connectivity.

> *"Just because we have a hammer, we don't want to go out and think that everything's an AI nail" - Stanford Workshop*

Design for scale and consider shared infrastructure.

Too often, promising digital health interventions stall after the pilot phase due to insufficient funding, inadequate infrastructure, or a lack of strategic planning for expansion.

For Funders:

- Structure funding to support pilots in achieving outcome data that will be needed to take an intervention to scale.
- Remain open to funding pilots initially supported by other funding bodies to enable sustained investment required for scale.
- Identify opportunities for centralized investment in shared infrastructure that multiple organizations can access and adapt.
- Encourage licensing and development of open-source models to maximize collective impact.

For Implementers:

- Design solutions for long-term sustainability from the outset.
- Prioritize establishing a clear evidence base for your intervention to support government decisions to implement at a national scale.
- Partner with governments and healthcare systems to embed GenAI tools into national strategies and service delivery models.

> *"There's also been a shift in funding, where initially there was a lot of small scale, innovation-based funding for pilots. The problem that is ever-present in the sector is that the whole purpose of a pilot is so that you can then ideally scale something that works well, but without the funding to do that, you're just left with a pilot. And so it was just a lot of, 'Look at how we were able to use AI in the small use case,' and then they kind of fizzle, or you can't maintain the systems because, of course, you need the money and so on. And we are seeing more scale funding." - Elizabeth Shaughnessy, Director of Digital Programming & Co-Lead of AI Working Group, NetHope*

TURN.IO: EXEMPLIFYING SHARED INFRASTRUCTURE SUPPORTING MULTIPLE ORGANIZATIONS

> *"We find the best implementing organizations, of which there are thousands in the world that are trying to have an impact in the Global South, and we try and provide them with the resources, the technology or the platforms and advice in order to deliver and to scale evidence based solutions in healthcare" - Gustav Praekelt, Co-Founder, Turn.io*

Turn.io addresses the challenges of health service scalability through its GenAI-powered helpdesk, enabling organisations in LMICs to deploy chat-based solutions across the public health and low-cost private healthcare sectors. The system functions as a centralized platform that various health organizations can use and customize. This shared infrastructure approach reduces redundant technology development, contributing to the development of a shared "health commons" and enabling multiple organizations to scale their interventions efficiently. Their target market encompasses healthcare providers, NGOs, and government health services across the Global South seeking to scale their digital health engagement.

After successful pilots, the platform is now deployed by over 200 organizations, with 50 million users across deployments including 12 million new users in 2024. Multiple organizations using the platform are running RCTs with results expected in 2025.

Successful deployments include:

- Penda Health (Kenya): Penda Health is a primary healthcare provider, using the Turn.io helpdesk to enable remote care and increase access. Using the platform, they have scaled telemedicine delivery from 20–50 to over 1,500 interactions per month, with a 50% reduction in response time.
- MomConnect (South Africa): MomConnect is a flagship initiative of the Department of Health in South Africa, providing interactive maternal health support. Through the Turn.io platform, they are answering an average of 40,000 maternal health questions per month.
- Noora Health (India): Noora Health's digital companion powered by Turn.io supported 700,000 users in 2024 Please see our in-depth Noora Health case study for further information.

DIMAGI'S OPEN CHAT STUDIO

> *"So our approach to getting involved in large language models was to build a platform, initially for ourselves, just as a way of being able to quickly spin up different chatbots and try to understand what it looked like to do different prompting and understand kind of the safety and controls of everything. So we put that together and then realized it might be useful for other people. So open sourced it and put it out as Open Chat Studio" - Brian DeRenzi, VP, Research and AI, Dimagi*

Dimagi is a global social enterprise working to build and scale sustainable, high-impact digital solutions that amplify frontline work in healthcare and other sectors. Their Open Chat Studio (OCS) is an open-source platform to facilitate the rapid prototyping, testing and deployment of LLM-based chatbots, democratizing access to GenAI technology and helping to ensure that its benefits are realised equitably by enabling developers to tailor interventions to local needs. The platform works with any LLM with an API, such as GPT-4, and can be deployed over the web as well as via mobile messaging apps including WhatsApp and Telegram.

There are approximately 50 organizations currently onboarded with Open Chat Studio. Innovative recent deployments (currently in early pilot testing) include collaborations with two established multimedia behavior change organizations: Shujaaz in Kenya and Réseau Africain de l'Éducation pour la Santé (RAES) in Senegal. Both organizations seek to enhance sexual and reproductive health (SRH) education and behavior change among adolescents through role-playing conversations, with chatbots emulating known characters from comic books and TV series.

Confront societal biases and problems

GenAI tools have the potential to amplify and perpetuate societal biases embedded in their training data or influenced by the assumptions of developers, which must be accounted for throughout the design and implementation phase.

> *"When we talk about bias and algorithms, we're talking about very specific technical bias that is very difficult to mitigate in practice, because if you're using a package or tool off the shelf, it might already have the encoded bias. So it's important to acknowledge that we won't be able to fully understand the scope of what that bias might look like, but maybe we can see what the impacts will look like, and how to mitigate it from there, which also means human review is really important." - Elizabeth Shaughnessy, Director of Digital Programming & Co-Lead of AI Working Group, NetHope*

For Funders:

- Fund research into ethical AI frameworks tailored to LMIC contexts.

For Implementers:

- Develop algorithms that include checks against bias and discrimination.[29]
- Use datasets that reflect diverse populations, geographies, and healthcare contexts to minimize bias. Collaborate with local experts to include underrepresented perspectives, including gender-specific health needs. Regularly audit training datasets for bias and inaccuracies.
- Ensure transparency in development by clearly communicating methodologies, data sources, and bias mitigation strategies.
- Comply with legal governance frameworks and Responsible AI guidelines where they are in place, and remain alert to the potential for rapid changes in the regulatory landscape.
- Understand and monitor how adversarial actors are using GenAI to accomplish their goals, and establish appropriate mitigations.

STANDING Together (STANdards for data Diversity, INclusivity and Generalisability), a partnership of over 30 academic, regulatory, policy, industry, and charitable organisations worldwide, has published recommendations to support transparency regarding limitations of health datasets and proactive evaluation of their effect across population groups, with the aim of reducing the risk of perpetuating existing biases and health inequalities when using AI technologies.[30]

Build local stakeholder and government buy-in

For GenAI health interventions to succeed, gaining the trust and engagement of local and national stakeholders—including patients, healthcare workers and policymakers—is essential, not only for ensuring the relevance and acceptability of these tools but also enabling smoother implementation, scalability and long-term sustainability.

> *"I think scaling in low and middle income countries is challenging because success requires more than just a great product, but you're also looking at local buy-in. So that's one of the criteria that we're trying to understand when we talk to folks: who are they connected to? Are they working with governments? Are they working with Ministries of Health? Do they have those networks already?" - Topaz Mukulu, Strategy Analyst, Patrick J McGovern Foundation grantmaking team*

> *"Ultimately, any digital health solution, be it AI enabled or non-AI enabled… At the most basic level, it relies on data generation, which happens by the people in the field, the Last Mile Health Worker, and if they are convinced that this technology is of no use to me in my daily work, they will not use it. And the best of technologies will fail… So the buy-in of the Last Mile Health Worker is critical, because that's where the real health outcome data health comes in" - Manish Pant, Policy Specialist, Digital Health, UNDP*

For Funders:

- Support alignment with national health strategies by funding projects that actively engage local policymakers and health system leaders.
- Provide grants for community engagement efforts needed to develop successful interventions.

For Implementers:

- Partner with local organizations and government agencies to ensure alignment with regional and national health priorities.
- Demonstrate relevance and impact by ensuring evaluation metrics address specific local challenges.
- Clearly communicate the capabilities, limitations, and safeguards of GenAI health tools.
- Offer ongoing training and support for end-users to ensure sustained adoption.
- Create channels for stakeholders to provide feedback and voice concerns during implementation.

Photo Source: Noora Health

CASE STUDIES IN HEALTH-RELATED BEHAVIOR CHANGE

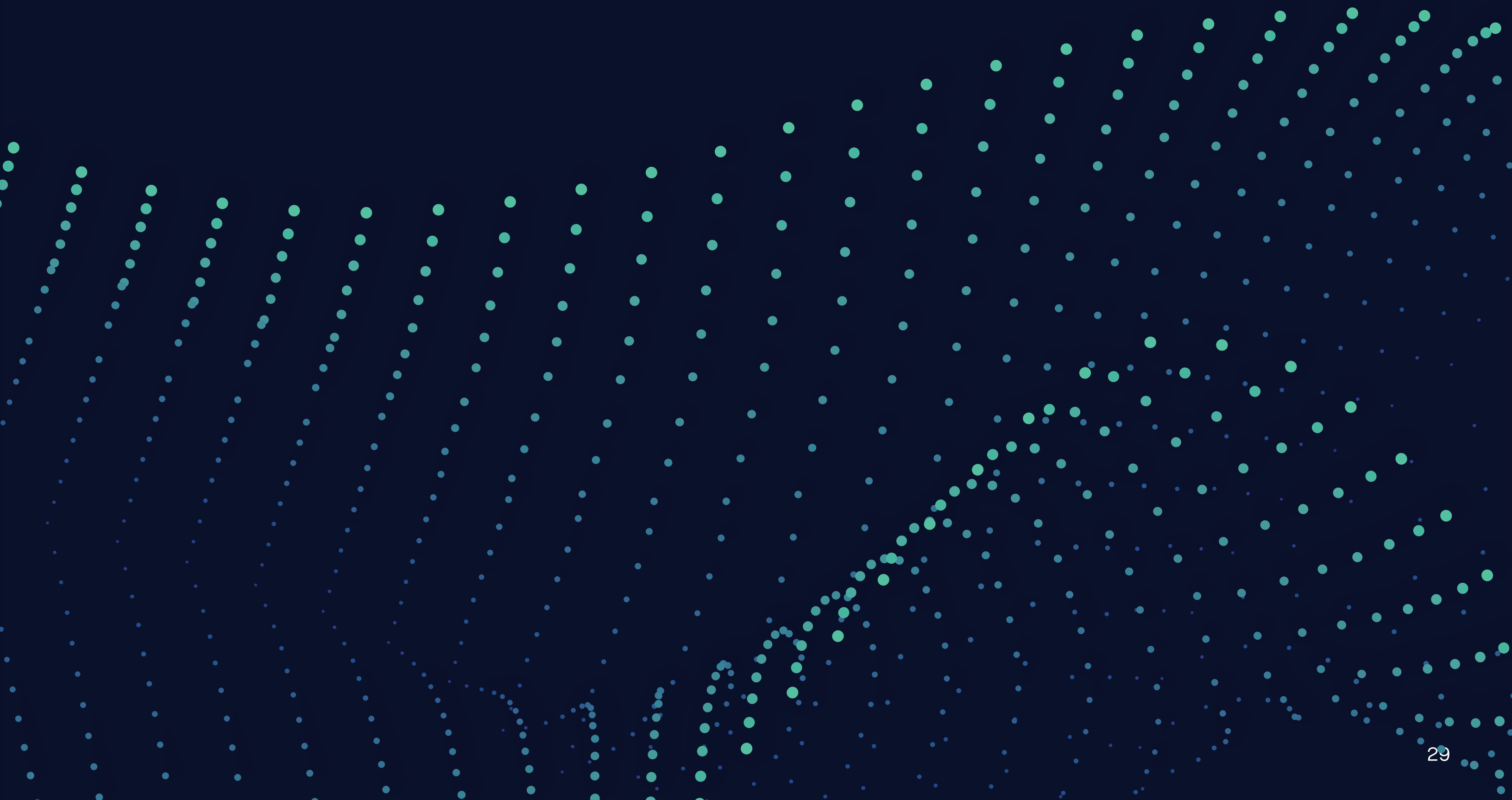
Through our two roundtable events, in-depth qualitative interviews, analysis of key GenAI accelerator programs and survey, we have identified many promising projects utilizing GenAI for health-related behavior change in the 'pilot phase', including some that are already deployed to 10,000+ monthly users. As of late 2024, the only widely scaled application (to 100,000 or more monthly users) of GenAI in health-related behavior change for LMICs we were able to find is Jacaranda Health's PROMPTS. However, several others have imminent scaling plans, and a predictable path to fast scaling is apparent for GenAI pilots conducted as part of an established broader scaled system: for example, an existing helpdesk workflow with millions of total users that is now testing integrating GenAI for efficiency improvements.

We have identified sharing of learnings, including case studies demonstrating implementation principles and evaluation processes, as an important process to accelerate progress in the field. We therefore present five exemplar case studies which have promising preliminary impact data, all of which are planning further evaluation in 2025 with a move to greater scaling. These projects also demonstrate various key principles and risk mitigation approaches detailed in our recommended framework.

Given the nascency of the field, most projects are still in early phases regarding outcome data specific to GenAI integration. Where possible, we have presented outcome data on:

- Scale of deployment
- Evaluation of LLM performance, for example accuracy and completeness of LLM responses, as well as qualitative factors such as understandability, empathy, and appropriateness of tone and style
- Health impact: intended or actual health-related behavior change (largely not yet available)
- Cost-effectiveness analysis (largely not yet available).

We found that in terms of describing scale, there is currently not a consistent set of metrics being used by implementers, and it is often difficult to separate out the impact related to non-AI, deterministic AI, and generative AI elements of a given use case. As highlighted in our key recommendations section, a move towards consistency with this going forward, whilst acknowledging unique applications and individualized requirements, would facilitate more meaningful comparisons between projects. Evaluation of cost-effectiveness will be an important focus in upcoming evaluation frameworks.



CASE STUDIES IN HEALTH-RELATED BEHAVIOR CHANGE

CASE STUDY	USE CASE	DEPLOYMENT STATUS	TARGET AUDIENCE
Jacaranda Health: PROMPTS	Direct-to-consumer SMS messaging for maternal and newborn health.	Scaled: 526,000 users of GenAI-enabled platform in Kenya in 2024. 3 million cumulative users in Kenya inclusive of pre- GenAI models.	Pregnant and postpartum mothers in Sub-Saharan Africa.
Viamo: Ask Viamo Anything (AVA)	Direct-to-consumer, voice-based query responses accessed by basic phone call, for populations without internet access.	Late pilot: 32,000 users in Zambia, DR Congo, Nigeria, Botswana, Tanzania and Pakistan.	No-/low-literacy, underserved populations using non-internet phones in Sub-Saharan Africa and Asia.
Girl Effect: Big Sis & Bol Behen	Direct-to-consumer chatbots for youth SRH education.	Over 75,000 users in India routed to content via LLMs. Successful early pilot of GenAI chatbot in South Africa (4,000 users) has led to scale-up phase. Deterministic chatbots active at scale, with over a million conversations initiated since 2018.	Adolescent girls and young women in Sub-Saharan Africa and South Asia.
Audere: Self-Care from Anywhere	Direct-to-consumer conversational AI for HIV, SRH, GBV education, counseling, and linkage to care.	Early pilot; scaling studies planned for 2025 in South Africa and Zimbabwe.	High-stigma populations at risk for HIV.
Noora Health: Remote Engagement Service query classifier	Direct-to-provider query classification for clinical teams.	Classification system in use by a 20-person clinical team processing 10,000 messages/day.	Caregivers in India, Bangladesh, and Indonesia.

BENEFITS	GENAI OUTCOME DATA	FUTURE PLANS
Improved knowledge of pregnancy/postpartum danger signs and uptake of maternal health services; improved government accountability to improve services.	Significant improvement in average response times for users (~10-15 mins vs. 2-4 days); capacity to respond to 10,000+ incoming questions a day from mothers.	Improved personalization at scale; clinical efficiency and cost-effectiveness analyses; further platform expansion in Sub-Saharan Africa.
Increased access to healthcare services; stigma-free information delivery; reaching underserved populations.	94% response listening rate; 59% female users; high engagement in rural areas; high self-reported behavior change.	Scale to Viamo's 14 other countries and 27 million existing non-AI users; increased localization; traceable referrals to health product and service providers.
Improved user knowledge and agency; increased SRH service uptake.	AB testing of GenAI service demonstrated a significant increase in key message consumption and service access; 104% 'deep' content consumption increase in India.	Further A/B testing to assess impact of GenAI integration on user engagement, retention and behavior change; potential RCT evaluation of chatbots.
Improved sexual and reproductive health education; stigma-free access to HIV testing, prevention or confirmatory care options; improving efficiency and efficacy of clinical follow-up.	Greater than 90% usability, acceptability and appropriateness of the AI Companion; qualitative data demonstrated success in building user trust and comfort in addressing sensitive topics, and gathering more honest risk information.	Three scaling studies planned for 2025, targeting vulnerable populations for HIV counseling, prevention and treatment support.
Operational efficiency.	80% reduction in nurse-reviewed queries; low false negative rate.	Partnership with local organizations to improve vernacular language coverage and reduce misclassifications.

CASE STUDIES

CASE STUDY

1 JACARDANDA HEALTH: PROMPTS

> *"We realized very early on, quite accidentally, that if you send messages and there's a free way to respond, moms start asking questions seeded by the messages you're sending, but they have a ton of questions, and then we realized very quickly that we need a way to answer those questions in an efficient way." - Sathy Rajasekharan, Co-Executive Director of Jacaranda Health*

What is the GenAI use case?

Use Case Category: Direct-to-consumer (human-in-the-loop)
Health Area: Maternal, Newborn and Child Health (MNCH)

PROMPTS is a two-way SMS service designed to promote positive care-seeking behaviors amongst new and expectant mothers through timely health information and support throughout the pregnancy and postpartum journey. Responses are generated by Jacaranda's customized LLM, UlizaLlama, which is based on Meta's Llama2 and fine-tuned for use in Swahili and English. Launched in October 2023, UlizaLlama is the first free-to-use Swahili LLM, and since August 2024 has been further fine-tuned for other African languages. As of October 2024, GenAI is integrated as standard on the platform.

What tasks are LLMs being used for?

LLM tasks:

- **Summarization:** Condensing health guidance into easy-to-understand SMS messages.
- **Classification:** Identifying high-risk users based on message content and flagging them for escalation.
- **Extraction:** Extracting key information from user conversational history and clinical information to enable personalized risk profiles.
- **Translation:** Handling mixed-language queries, local vernacular and slang; presenting technical medical information in accessible, patient-friendly language.
- **Conversation:** Personalized, two-way communication with users in Swahili and English.

Designing for Inclusivity:

- **Offline accessibility:** SMS-based model ensures accessibility for users without smartphones or internet connectivity.
- **Local adaptation:** UlizaLlama supports multiple African languages, ensuring responses are culturally and linguistically relevant.
- **Community trust-building:** Strong government partnerships and local tech and helpdesk teams enhance credibility and engagement.

Mitigating risks:

- All messages flagged as high-risk are escalated straight to the helpdesk for a human response (from a trained clinical nurse on the PROMPTS helpdesk).
- For routine queries, a second LLM audits UlizaLlama-generated responses for correct grammar, clarity and coherence, and medical accuracy.
- Responses scoring 85% or more are sent directly to the user (this threshold was derived from a previous system assessing responses of human helpdesk agents).

- Responses that fail the audit process are sent to human agents for review.
- Users who flag query answers as unsatisfactory are connected to the human helpdesk.

What is the current deployment status?

- Approximately 526,000 unique users in Kenya engaged with the GenAI-enabled platform in 2024.
- Around 10,000 questions are answered per day, with 70% directly answered by GenAI (generated response passed audit), and the remainder referred to human agents for review (generated response failed audit) or answered directly by humans (high-risk cases bypassed automated response).

> *"The generative tool does not answer questions related to the miscarriage, or significant trauma; anything like that goes straight to a human being, but they need time to be able to answer that. So by taking away all the other stuff, they're now able to focus on that… [Staff are] much happier not answering backlog questions and focusing on significant problems." - Sathy Rajasekharan, Co-Executive Director of Jacaranda Health*

How widely deployed could it be over time?

- Target market is new and expectant mothers in Kenya and potentially across Sub-Saharan Africa. Sub-Saharan Africa is home to over 250 million women of reproductive age.
- In total, over 3 million people have used the PROMPTS platform since its launch in 2017.
- Work has begun on expansion into Ghana, Nigeria, Eswatini, and Nepal.
- Work is underway to make the multilingual UlizaLlama LLM specific to the maternal and newborn health domain for other African languages, including Hausa, Yoruba, Xhosa, and Zulu.

How are they measuring success?

LLM Performance:

- Evaluation of LLM responses for medical accuracy and appropriateness, personability, and simplicity: UlizaLlama outperformed the top-rated 'off-the-shelf' LLMs by approximately 14% on overall scores. For example, off-the-shelf models use more complex words and medical jargon, which is not appropriate for the PROMPTS audience (as per reading level estimates).
- Response time to user queries: Average response times decreased from approximately 5 hours in September 2024 to less than 15 minutes in December 2024. With GenAI, 70% of queries receive an instant response.
- Language-specific outcomes, assessed using established NLP evaluation tools such as BLEU (Bi-Lingual Evaluation Understudy).

Health impact:

- (Data specific to GenAI integration forthcoming; health impact data prior to GenAI integration):
- 20% increase in mothers attending 4+ prenatal care visits.
- 100% increase in uptake of postpartum family planning services.
- 89% of users exclusively breastfed for the first six months post-delivery.

Cost-Effectiveness:

- GenAI adds an estimated $0.10 to cost per user ($0.74 per user for duration of use prior to GenAI integration).
- The principal cost driver is SMS costs, at around $0.40 per user. Enrollment costs per user are $0.12–0.20, depending on rural or urban location, with remaining costs operational (e.g. helpdesk team, field agents).

Future measurement plans

Developing a framework for efficiency of maternal care, to evaluate the cost-effectiveness of PROMPTS in terms of appropriate care-seeking.

What's been instrumental for PROMPTS to enable scaling?

"This is a topic that's come up a lot in the last couple of weeks—what's the secret sauce? The reality is this series of years of just battling. We started off as a small pilot, so I never have a problem with pilots; there's a thing people say: 'Oh, we're tired of pilots'. And the problem with that is it assumes you can innovate without having pilots, which is not possible. All good things emerge from some idea that needs iteration. I think enablers for our scale have been:

- ***A relatively lean platform.*** *Right from the start, we have always focused on the most efficient way to deliver the service for impact. So continuously, we've looked at the impact and then said, 'what can we pare down to maintain that impact?'*
- ***Partnership with the government*** *has been key—we're partnering directly with the Ministry of Health in Kenya that builds trust with the enrolled at facilities. This probably wouldn't have worked as well if we did some sort of big mass media campaign to enroll people, because there's a lot of fluff out there, and you probably wouldn't reach the right audience.*
- ***Continuous learning*** *has been helpful, because we're able to pivot to things like using AI as part of a solution. As we scale from 1,000 moms to 100,000 to a million moms, very different prospect in terms of technology and operations that you need.*
- ***Team is a huge piece*** *of the scaling thing that I think gets discounted a lot because everyone wants a model that they could say, oh, here's how you replicate. But I think the people piece is probably the biggest one. Our entire tech team is Kenyan. We've benefited a lot from smart partnerships where we've had experts come in and guide the team, but by and large, they're the ones who've done and built everything, and I think that's been instrumental as well. So this isn't a project. This is our business. This is what we do, and it's healthier for us to have a local team building solutions for a local problem ... Finding the right people for your organization makes or breaks it."*

  - Sathy Rajasekharan, Co-Executive Director of Jacaranda Health

Photo Source: Jacaranda Health

> *If you want to reach economically disadvantaged people in low- and middle-income countries, most don't have smartphones, or don't have access to the internet, and literacy is an issue. You need to engage disadvantaged groups on the device they already have in their pocket." - Stephen Meyer, Director of Partnerships at Viamo*

What is the GenAI use case?

Viamo has two GenAI products which have been deployed in 9 countries and 5 languages in Africa and Asia: 'Ask Viamo Anything' (AVA) and 'Ask An Expert' (AAE). We focus on AVA as an example of a tool with exciting potential to drive health-related behavior change.

> *"In short, it's ChatGPT for offline audiences on a phone call." - David McAfee, CEO, Viamo*

> *Use Case Category: Direct-to-consumer (fully autonomous)*
>
> *Health Area: Health Systems Strengthening (HSS)*

Ask Viamo Anything (AVA) is a voice-based GenAI system enabling conversational interactions, designed for users with basic, non-internet-enabled phones. The platform converts user calls into text files to input to the LLM (GPT-4 or others depending on language), and converts the LLM text responses to voice files to return to users. This approach tackles two key barriers: 1.) low literacy, which hinders use of text-based solutions; and 2.) digital divide challenges in LMIC contexts, by enabling offline access to reliable, context-specific information. During a 2024 pilot in Zambia, approximately 30% of a total of 570,000 questions from 32,000 users were health-related, addressing topics including HIV prevention, maternal health, and mental health. Responses to health-related queries aim to effect meaningful behavior change, such as attending a healthcare clinic to access essential care.

> *"People are asking questions strongly around highly stigmatized topics ... We suspect that our users have maybe never had an honest and direct and open conversation about HIV before, and now they have the opportunity to have this conversation from the privacy of their own phone, where they don't have to discuss with their auntie or their health worker or their partner, and they just ask all these nitty-gritty questions and get answers." - Stephen Meyer, Director of Partnerships at Viamo*

What tasks are LLMs being used for?

- **Summarization:** Condensing complex health information into accessible query responses.
- **Classification:** Categorizing user queries by topic to map to relevant responses.
- **Extraction:** Extracting relevant details from user queries to provide tailored responses.
- **Translation:** Supporting conversations in multiple core languages including English, French, Swahili, Portuguese, Urdu, Arabic; adapting to local vernacular and cultural nuances; converting between voice and text formats.
- **Conversation:** Facilitating tailored, empathetic interactions on sensitive and stigmatized health topics.

Designing for inclusivity:

A key focus of AVA is to reach underserved populations (e.g. women in rural areas; non-literate users), achieved by:

- **Voice-based interface:** Responses delivered by phone call to tackle challenges of low literacy.
- **Offline accessibility:** Designed for non-smartphone users without access to internet connectivity.
- **Localized content:** Responses are tailored to users' cultural norms.

Mitigating risks:

- Early auditing of a large sample of health-related questions with Harvard Global Health found that all responses given were accurate, with room for improvement in localization and referrals.
- Resulting modifications to prompt engineering, in-country partnership strategy, and improved use of ChatGPT's moderations API. HITL processes are no longer in place.

What is the current deployment status?

- Late pilot: pilot testing with 32,000 users in Zambia and smaller pilots in DR Congo, Nigeria, Botswana, Tanzania and Pakistan conducted in 2024. On average, users called in 7.7 times per month.

How widely deployed could it be over time?

- Target audiences are underserved populations including those who lack internet access (2.7 billion people globally), use non-smartphone devices, and/or have low literacy.
- Viamo now has 27 million users on the basic, non-GenAI platform, in 68 languages across 19 countries.

How are they measuring success?

LLM performance:

- Evaluation of LLM outputs for accuracy, cultural appropriateness, and empathy.
- User engagement metrics:
  - » 94% of users listened to complete responses in pilot testing;
  - » engagement across different demographics: 59% of users were female; 74% were aged 24 or under; high engagement in rural areas.

Health outcomes:

- Key health outcomes of interest—formal evaluation data forthcoming:
- Self-reported behavior change
- Verified behavior change (such as redeeming a coupon, connecting to a partner call center, attending an appointment).

Cost-effectiveness:

- Viamo has negotiated zero-rating agreements with telecommunications companies (Airtel, MTN, Vodafone, Orange, and more) to eliminate the airtime costs. Telecoms contributed over $200M in airtime in 2024.
- Viamo has so far secured free product credits for all GenAI tech, and will explore either zero-rating or open-source options at scale.
- The resulting cost-per-engagement is a few cents, and will continue to decrease with scaling.

Future measurement plans

- The non-GenAI version of the Viamo platform has traceably connected users directly to health services (family planning and appointments, health call centres, and many non-health services). Viamo is interested in understanding if AVA is a more powerful tool to create these traceable outcomes.

Photo Source: Viamo

CASE STUDY

3

GIRL EFFECT: BIG SIS AND BOL BEHEN

> *"The role of our AI-enhanced chatbots is largely around building space for young people to be able to ask questions, and they're in a contemplation phase in making a choice about their sexual health ... we have discovered in every market that they lack a judgment-free space to contemplate these decisions." - Karina Rios Michel, Chief Creative and Technology Officer, Girl Effect*

What is the GenAI use case?

Use Case Category: Direct-to-consumer (human-in-the-loop)
Health Areas: Sexual and Reproductive Health (SRH); Mental Health

Girl Effect is a nonprofit organization dedicated to empowering adolescent girls globally by providing them with the tools, information, and support they need to overcome societal barriers, including those relating to accessing healthcare, focussing within the domains of sexual health, economic empowerment, education and mental health. They use social behavior change methods through multi-channel, multi-product programs delivered via radio, TV, social media, and community activations to promote agency of girls and young women.

Central to Girl Effect's strategy are chatbots designed to motivate and inspire young people to take charge of their health and access relevant health services. These chatbots have previously used deterministic, classification-based BERT (Bidirectional Encoder Representations from Transformers) AI models to map user queries to pre-curated responses. In 2024, Girl Effect began integrating GenAI to deliver more personalized, dynamic responses, piloting a GenAI version of their South African chatbot 'Big Sis', and enhancing their Indian chatbot 'Bol Behen' with LLM-powered content classification. GenAI integration is supported by HITL safeguards, enabling escalation of high-risk cases.

> *"The advantages of GenAI ... we're using it for a much more complex understanding of what users want. So we're kind of seeing it as a way to understand our users better, and how we can use that power to deliver our services better." - Soma Mitra-Behura, Lead AI Researcher, Girl Effect*

What tasks are LLMs being used for?

- **Summarization:** Condensing health guidance into easy-to-understand SMS messages.
- **Classification:** Classifying user inputs according to topic to provide relevant responses; identifying high-risk cases for escalation to human supervisors.
- **Extraction:** Retrieving relevant details from user queries to provide tailored responses.
- **Translation:** Handling mixed-code languages such as Hinglish (Hindi-English) and Sheng (Swahili-English), and adapting to local/youth vernaculars and slang.
- **Conversation:** Natural, human-like interactions with users on sensitive topics, emulating a trusted 'big sister' persona.

Designing for inclusivity:

- Youth-friendly design: chatbots emulate a trusted 'big sister' persona to encourage rapport.
- Cultural relevance: chatbots adapt to local slang and code-mixed languages, ensuring contextual authenticity.
- Community co-creation: target users are engaged throughout, ensuring content delivery reflects their needs and concerns.

Mitigating risks:

- LLM classification flags high-risk disclosures for human review, and automatically directs users to emergency or professional support services.
- LLM outputs are limited to predefined and vetted content boundaries.
- Bespoke LLM evaluation framework has been built for Girl Effect, measuring safety, accuracy, relevance and tone. Unsupervised user engagement could only commence once the framework demonstrated a sufficiently high pass rate.

What is the current deployment status?

- Early pilot of generative chatbot ('Big Sis') in South Africa:
  - Supervised co-creation and alpha testing of the GenAI prototype were conducted in South Africa in August 2024 to assess user trust and barriers to GenAI engagement.
  - An unsupervised beta pilot ran from December 2024 to January 2025 in South Africa, reaching 4,000 users and generating over 11,000 responses.
  - A/B testing compared the GenAI-enabled chatbot to the deterministic model, with users rating the GenAI chatbot as significantly more supportive and trustworthy.
  - Results showed users were significantly more likely to engage with key messaging content and access service information: GenAI implementation will now be scaled.
- Generative classification system is active in India:
  - Over 75,000 users have had their questions routed to content by LLMs.
  - 88,000 user submissions were successfully routed to key messaging.

Photo Source: Girl Effect

How widely deployed could it be over time?

- Target users are 18-24 year-olds with high unmet needs in sexual and reproductive health and mental health support.
- Across South Africa, Kenya, and India, Girl Effect's deterministic AI chatbots engage approximately 1.4 million users, with 110,000 monthly active users and 80,000 daily messages received.
- Potential market reach is estimated at 110 million users across these markets, with expansion plans in Nigeria.

How are they measuring success?

LLM evaluation:

- User engagement: high levels of engagement in supervised testing of generative chatbot in South Africa.
- User experience: qualitative feedback captures positive feedback on chatbot tone and relevance of responses.
- Evaluation framework: LLM tests the safety, accuracy, reliability and tone of the GenAI answers.
- A/B testing comparing the efficacy of GenAI and non-GenAI experiences.

Health outcomes:

- Increased consumption of content previously demonstrated to be successful in achieving behavior change: the generative classification system in India has achieved a 104% increase in repeat engagement with key messaging, and increased breadth of content consumed.
- A/B testing in South Africa found a 300% increase in engagement among users of the GenAI model compared to the legacy model. GenAI users were 12% more likely to access service information in the chatbot and 11% more likely to engage with key messaging compared to the control group.

Key health outcomes of interest: demonstrated for pre-GenAI models, formal GenAI evaluation data forthcoming:

- Increase in awareness and knowledge of contraceptive methods, STI prevention and mental health coping mechanisms.
- Increase in intention to use and actual usage of contraceptives among adolescent girls and young women (AGYW).
- Increase in AGYW intending to get tested and regularly getting tested for HIV.

Cost-effectiveness:

- Girl Effect is assessing the cost-effectiveness of their Kenyan chatbot by analyzing costs in relation to contraceptive needs and service uptake conversions.
- Research by the Guttmacher Institute shows each dollar spent on contraceptive services for adolescents in Kenya saves $2.71 in maternal and newborn healthcare costs; fully addressing the contraceptive needs of adolescents in Kenya could reduce pregnancy-related healthcare expenses by $46 million.
- For Girl Effect's non-AI-enabled chatbot WAZZII, the estimated cost to encourage a young person to uptake a health service was $28 per user.
- Four-week beta testing of 'Big Sis' in South Africa cost approximately $80 for the duration of the test ($0.007 per question), indicating that using GenAI could remain cost-efficient.

Future measurement plans

- Analysis of A/B testing results to determine the impact of GenAI integration on:
  - » User uptake of topic recommendations
  - » User engagement and retention
  - » User behavior change
- Additional A/B testing will explore impact of other system variables:
  - » Allowing users to have a conversation history of three or more messages
  - » Different editorial approaches
  - » Additional layers of natural language understanding before generating answers
- RCT planned in Kenya in partnership with the World Bank; exploring potential for additional RCT in South Africa.

4 CASE STUDY

AUDERE: 'SELF-CARE FROM ANYWHERE'

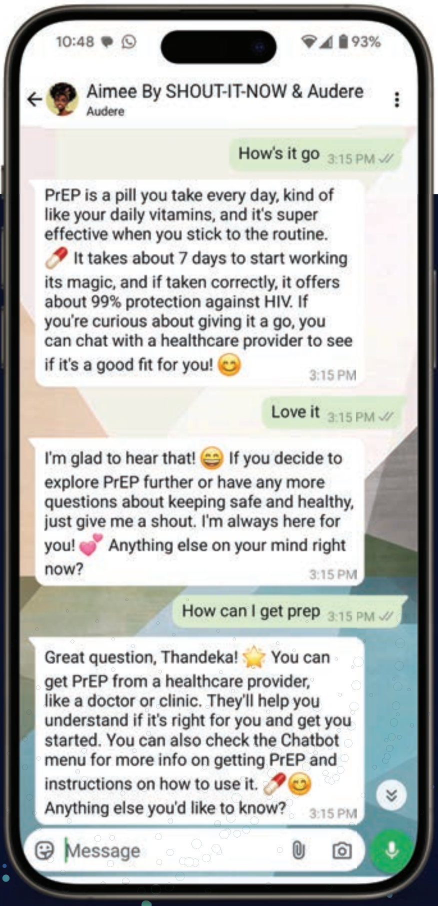
> *"Sometimes when people visit a clinic, it's really hard for clinicians to remain fully empathetic and not have some of the questions come off as abrasive, because they just don't have time to slowly gather those data. They have to ask these really personal questions ... And the individuals can feel stigmatized for their answers ... and just don't want to answer while they're looking somebody in the eye." - Shawna Cooper, Director of Product, Audere*

What is the GenAI use case?

Use case category: Direct-to-consumer (human-in-the-loop)
Health Areas: Communicable Diseases; Sexual and Reproductive Health

Audere is a nonprofit organization using GenAI to enhance HIV prevention and counseling services in South Africa. Their 'Self-Care from Anywhere' program was co-created with local community partners and SHOUT-IT-NOW, a South African nonprofit providing youth-focused HIV prevention and sexual health services. Powered by Audere's multimodal HealthPulse AI toolkit, 'Self-Care from Anywhere' addresses challenges faced by adolescent girls and young women (AGYW) in accessing sexual and reproductive health services due to stigma, logistical barriers, and overstretched healthcare systems.

Accessible via WhatsApp, an empathetic AI Companion provides health education, sexual health and HIV information, self-testing guidance and care linkage, while clinicians benefit from a summarized, AI-driven decision support view of AI Companion interactions through a Clinical Portal. The AI toolkit integrates computer vision, large language models, and predictive analytics to tackle sexual health, HIV prevention, and gender-based violence. In 2025, the program will expand to include HIV treatment counseling, mental health, and TB care for additional vulnerable populations, including men living with HIV, female sex workers, and people on the edge of sex work in South Africa and Zimbabwe.



10:48 93%

Aimee By SHOUT-IT-NOW & Audere
Audere

How's it go 3:15 PM

PrEP is a pill you take every day, kind of like your daily vitamins, and it's super effective when you stick to the routine. 💊 It takes about 7 days to start working its magic, and if taken correctly, it offers about 99% protection against HIV. If you're curious about giving it a go, you can chat with a healthcare provider to see if it's a good fit for you! 😊 3:15 PM

Love it 3:15 PM

I'm glad to hear that! 😄 If you decide to explore PrEP further or have any more questions about keeping safe and healthy, just give me a shout. I'm always here for you! 💕 Anything else on your mind right now? 3:15 PM

How can I get prep 3:15 PM

Great question, Thandeka! ⭐ You can get PrEP from a healthcare provider, like a doctor or clinic. They'll help you understand if it's right for you and get you started. You can also check the Chatbot menu for more info on getting PrEP and instructions on how to use it. 💊😊 Anything else you'd like to know? 3:15 PM

Photo Source: Audere

What tasks are LLMs being used for?

- **Summarization:** Condensing client/AI Companion conversations into actionable insights for clinicians which can be viewed through the Clinical Portal.
- **Classification:** Flagging high-risk cases for human intervention, such as disclosures of abuse or suicidal ideation.
- **Extraction:** Retrieving key HIV vulnerability information from user conversations to inform risk assessments and tailored guidance.
- **Translation:** Multilingual capabilities; adapting responses to local vernacular and slang to ensure cultural relevance.
- **Conversation:** Conducting empathetic and stigma-free multi-turn interactions on sensitive topics; simulating different personas (e.g. health professional, sister, friend) according to user preference.

> *"Involving local partners from the beginning, co-designing with them, and ensuring local context, tone and slang and style are brought in ... We've created a slang dictionary for South Africa that has more than a thousand words in it ... For instance, if a girl says, 'I'm afraid I have the drop,' which in South Africa means an STI, the AI Companion knows that." - Shawna Cooper, Director of Product, Audere*

Designing for inclusivity:

- **Community co-creation:** developed in collaboration with community partners, and utilizing local guidelines to ensure context-appropriate information is provided.
- **Slang dictionary integration:** enabling nuanced communication, trust and rapport building.
- **Combating stigma:** focusing on reaching underserved populations with limited access to healthcare, and addressing stigma surrounding HIV.

Mitigating risks:

- Conversations are monitored in real time by an automated framework, with a portion reviewed by local clinicians, HIV testing service counselors and community representatives for local relevance, accuracy, and adherence to guidelines.
- High-risk cases such as disclosures of harm are flagged in real time for human intervention (clinicians).
- Users are able to request to communicate with a human at any time during AI Companion conversations.

What is the current deployment status?

Early pilot, research phase:

- 'Your Choice' Study: 7-month study with 130 clients and 20 healthcare professionals, tested an early alpha version of the AI Companion for pre-HIV-self-test counseling and prevention awareness; and clinical summaries of AI/client conversations for clinicians.
- 'Your Path' Study: 12-month study with 100 clients and 25 healthcare professionals, tested a beta version for supported HIV self-testing and the AI Companion for post-test counseling on confirmatory testing or prevention options like PrEP.
- Self-care from Anywhere field study: summative usability testing completed in early 2025 with 100 AGYW and 50 clinicians, with a 6-month field study commencing in Q2 2025 with 2000 AGYW in South Africa.

Across the above studies, co-design sessions, and summative usability testing, over 500 clients and nearly 100 clinicians have contributed to and used various alpha, beta, and release candidate versions of the solution.

How widely deployed could it be over time?

Target audiences: underserved populations in LMICs with limited access to healthcare and high risk and stigma surrounding HIV and SRH.

Scaling plans:

- Expansion to all SHOUT-IT-NOW clients—approximately 1.5 million people in South Africa.
- Expansion to all youth across South Africa through other community-based organizations and Ministry of Health support.

- Additional pilot studies for broader key populations in South Africa and Zimbabwe to demonstrate demand, linkage to care impact, and cost effectiveness.
- The platform is designed for adaptability to other health areas beyond HIV, including mental health, TB, and non-communicable diseases, as well as to other use cases such as clinical decision support.

How are they measuring success?

LLM evaluation:

- Evaluation of LLM outputs: accuracy and contextual relevance of candidate language models are assessed via an automated evaluation framework, which combines use of language models, quantitative metrics, and HITL evaluation. Once deployed, an automated monitoring framework ensures safety.
- Feasibility, acceptability, and usability for both clients and clinicians: 'Your Choice' trial demonstrated greater than 90% usability, acceptability and appropriateness.
- Client engagement metrics (awareness, interest, intent, engagement, and retention).
- User experience: qualitative data from 'Your Choice' trial demonstrated success in building client trust and comfort in addressing sensitive topics; healthcare providers saw value in client conversation summaries in addressing time limitations and provider mistrust.

Health outcomes:

Key health outcomes of interest—formal evaluation data forthcoming:

- Improving access to education about sexual health and HIV, self-awareness of HIV status, stigma-free access to prevention or confirmatory testing options.
- Improving efficiency and efficacy of clinical follow-up.

Cost-effectiveness:

Costs include usage fees for LLM services (e.g. token or API usage costs), data hosting costs, system maintenance and support, and implementation operational costs (e.g. clinician salaries). In 2023, early use of GenAI systems demonstrated between $2.65–$3.50 per 15-minute conversation, across ChatGPT and ClaudeAI versions.

System optimizations including development of an omnichannel LLM orchestration service, dynamic prompt system, and evaluation of lower cost alternatives have cut the cost to a fraction of the original token fees. Cost of use will be analyzed during the field study in 2025, and will include monitoring costs to ensure system safety, accuracy, and ethical use.

> [Describing 'Your Choice' research study results]:
> *"Women were very engaged and prepared to connect with the bot, and, importantly, disclose more information than they would to human counselors. Because the disclosure was more comprehensive, the risk assessments were more accurate … In fact, many women who self-disclosed ended up realizing that their risk was significantly higher than they themselves anticipated, and opted for HIV PrEP or self-tests, or to receive HIV care. We didn't expect that kind of chain to lead to behavior change so quickly." - Zameer Brey, Deputy Director, Technology Diffusion, Gates Foundation*

Future measurement plans

- 'Your Path' study concluded in December 2024, with analysis of results ongoing.
- Three scaling studies are planned for 2025, targeting different high-risk demographics for HIV counseling, prevention and treatment support. These also include expanded language support for isiZulu and Shona, and will include evaluations of commercial vs. fine-tuned LMs for each task where LLMs are utilized.
- A/B testing between intervention arms with and without GenAI integration.
- Evaluation of GenAI computer vision capabilities vs. purpose-built models for rapid test identification and interpretation via the automated evaluation framework.
- Plans to evaluate savings of preventive intervention against normal treatment costs.

CASE STUDY

5 NOORA HEALTH: REMOTE ENGAGEMENT SERVICE (RES) QUERY CLASSIFIER

> *"Where Gen AI is coming into picture for us, and why we are so excited about it ... Every day on the messaging platform, we are processing about 10,000 messages, out of which about 2,000 end up being health-related questions … We are servicing eight regional languages … spread out across India, Bangladesh and Indonesia. So, these 10,000 messages are coming across the languages and also across different formats … How do you identify the non-medical queries from the medical queries? How do you categorize the medical queries into different relevant buckets? How do you then attribute a risk level to a certain medical query so that you are able to attend to the more urgent queries first? To then generate a relevant response which is not just medically accurate, but also empathetic, which also takes the user context into picture." - Anubhav Arora, Co-Executive Director - Programs & Platforms, Noora Health*

What is the GenAI use case?

Use case category: Direct-to-provider
Health Area: Health Systems Strengthening (HSS)

Noora Health's Care Companion Program (CCP) seeks to equip family caregivers with the knowledge and skills to enable them to deliver home-based care. The program combines in-hospital training sessions with a mobile-based Remote Engagement Service (RES) providing real-time, personalized responses to caregiver queries, addressing knowledge gaps and reducing caregiver burden in managing domestic care across a broad range of health areas, with a majority of caregivers trained in maternal and newborn care.

Previously, the RES has required in-house clinical workers to respond to user queries. As of early 2024, GenAI is being gradually integrated to streamline this process through automated message classification, with plans for response generation. Further GenAI deployment is also envisioned to create user profiles based on user goals, engagement data, government data and other sources, enabling more personalized and persuasive recommendations.

What tasks are LLMs being used for?

- **Summarization:** Consolidating caregiver and patient information from multiple data sources to generate caregiver profiles and family summaries for clinical workers.
- **Classification:** Classifying user queries by intent, subject matter, and risk level.
- **Extraction:** Extracting health details from caregiver messages to inform response recommendations.
- **Translation:** Handling queries in eight regional languages; adapting to local dialects and colloquialisms.
- **Conversation:** Empathetic, personalized responses to user messages (in development phase).

Designing for inclusivity:

- **Multilingual capability:** classifying incoming queries in eight regional languages.
- **Community co-creation:** developed in collaboration with community partners.
- **Government partnership:** strong partnerships with state and national governments across India, Bangladesh, Indonesia and Nepal.

Mitigating risks:

- Queries classified as non-health are reviewed by lightly trained human agents to minimize false negatives.
- All generated outputs will be reviewed by a team of in-house nurses for accuracy, appropriate tone, usefulness and understandability.

> *“The next level of categorization that we are now working towards is within the medical category ... which condition area does it belong to, whether it is related to community care, or diabetes, or cardiac and then within each of those is a question related to diet, or antenatal visits, or a warning sign ... And the third use case related to that is the risk: assigning the risk level to each query—whether it's an emergency that requires immediate escalation; or whether it's urgent, where, you know, it seems important, but it requires more probing with the user; or if it is non-urgent or routine—can the answer be picked by our knowledge base?” - Anubhav Arora, Co-Executive Director - Programs & Platforms, Noora Health*

What is the current deployment status?

- The LLM query classification system is in daily use by the 20-person clinical team across India, Bangladesh and Indonesia, processing approximately 10,000 messages per day.

How widely deployed could it be over time?

- The existing CCP (without GenAI integration) has reached over 14 million caregivers across more than 11,500 health facilities since 2014, including over 9.6 million from Q3 2023 to Q2 2024.
- The RES had over 1 million subscribers at the end of Q3 2024, with approximately 200,000 monthly active users (users receiving messaging campaigns), averaging over 70,000 monthly interactive users (users who interacted at least once), and 18,000 monthly queried users (users who asked a question).
- A 20-person RES team currently handles roughly 10,000 user queries per day, in eight regional languages.
- CCP is projected to support 25 million patients annually by 2027.

How are they measuring success?

LLM evaluation:

- Precision and recall for message classifications.
- False negative rate (health queries misclassified as non-health): low false negative rates across languages (English, Hindi, Punjabi, Marathi, Telugu and Kannada).

Health outcomes/system efficiency:

- Clinical efficiency and productivity gains: pilot testing of GenAI query classifier saw an 80% reduction in the number of queries reviewed by nurses.
- Positive impact on internal clinical team decision-making and job satisfaction.

Health outcomes with wider CCP (including in-person training - not GenAI-enabled):

- 18% reduction in risk of neonatal mortality within the first month of delivery.
- Decreases in newborn readmissions (56%) infant complications (14%) and maternal complications (12%) two weeks post-delivery.
- 78% increase in adoption of skin-to-skin thermal care.

Cost-effectiveness:

- Platform/software costs; RES team staffing costs; user acquisition via IVRS; message delivery costs.
- For the non-GenAI-enabled platform: $1.07 per caregiver trained, $1.71 per patient reached (Q1-Q2 2024), including platform and software costs, RES team staffing, user acquisition, and message delivery costs. An initial, short-term increase in cost per user is anticipated with GenAI integration.

Future measurement plans

- **Enhanced Data Infrastructure & Labeling:** In partnership with local organizations to improve vernacular language coverage, reduce misclassifications, ensure consistent measurement of user engagement, and refine model outputs.
- **A/B Testing:** Comparison of different GenAI features (e.g. personalization, multimodality) against existing intervention to identify the most effective strategies for improving engagement, knowledge retention, and self-reported behavior change.
- **User-Centric Feedback Loops:** Periodic caregiver interviews and surveys to capture qualitative insights on AI-generated messages (e.g., empathy, clarity, trust); review of new content and feature rollouts by community committees or focus groups to ensure technology and content remains culturally appropriate, understandable, and helpful.
- **Longitudinal Tracking of High-Risk Users:** Identify and follow specific cohorts (e.g., high-risk mothers, chronic disease patients) for 6–12 months post-discharge, to validate whether AI-driven follow-up correlates with measurable reductions in complications or improvements in treatment adherence.
- **Cost-Effectiveness Analyses:** Incorporating detailed costing models (e.g., cost per subscriber, staff time saved) side-by-side with clinical outcomes, to determine if and when AI investments yield net savings, and where to further optimize the platform at scale.

Next steps for Noora Health:

As a winner of the Chat for Impact India Accelerator in partnership with Meta, Noora Health will be running a pilot using Llama3 to provide dynamic, voice-enabled, personalized pregnancy care and guidance. They have developed a prototype, with further development and testing planned for 2025 to evaluate frequency and depth of caregiver engagement, impact on appropriate knowledge and behavior uptake indicators, and health outcomes indicators including complications and readmissions.

Photo Source: Noora Health

CONCLUDING CALLS TO ACTION: ADVANCING GENERATIVE AI FOR HEALTHCARE

GenAI has enormous potential to improve healthcare access, engagement, and outcomes in LMICs, but realizing this potential will require sustained collaboration, investment, and strategic action from stakeholders across the ecosystem, including funders, implementers, researchers and policymakers.

First, there is an urgent need to strengthen the case for investment in GenAI for health. Future work will need to generate cost-effectiveness data and return-on-investment analyses to inform funding decisions. Building a compelling case for private sector engagement will be crucial to successfully scaling interventions beyond the pilot phase.

Second, the standardization of measurement frameworks is imperative. The current lack of consistent outcome metrics hampers comparative analyses and impedes assessment of long-term impact. We recommend that funders take an active role in driving the use of rigorous measurement frameworks by making it a requirement for grantees. This will both enhance accountability and create a valuable evidence base to guide future investments and policy decisions.

We propose the following calls to action to maximize the health impact of GenAI interventions:

1. **Promote Continuous Learning and Collaboration:** Rapidly evolving technologies necessitate practical guidance and regular dissemination of example case studies. Establish cross-sector platforms for knowledge exchange, ensuring that lessons learned are disseminated widely. Transparency about both successes and failures will accelerate progress.
2. **Establish actionable measurement strategies** for evaluating benefits, costs, and risks to inform implementation. Funders should mandate the use of standardized, rigorous measurement frameworks that capture both health outcomes and cost-effectiveness data. Implementers should collaborate to share methodologies and insights. A number of RCTs evaluating AI-driven interventions, tailored for low-resource settings, are planned for 2025, and will likely set the foundations for establishing evidence-based metrics.
3. **Enhance language and localization** to ensure equitable GenAI adoption, particularly in LMICs. Addressing gaps in model quality for underserved languages and health-specific contexts requires investment in building datasets and development of performance benchmarks tailored to local needs, including voice-based solutions for low-literacy populations.
4. **Strengthen technical capacity and shared infrastructure**, for example by centralizing technical resources and expertise to support funders, health system leaders, and implementers. This can reduce fragmentation and duplication of efforts, and facilitate scalable, cost-effective solutions that are adaptable across diverse health systems.
5. **Address the digital divide** by investing in foundational healthcare and digital infrastructure ensures the equitable deployment of GenAI tools where they are needed most.

This is a pivotal moment in the evolution of GenAI as a tool for healthcare worldwide. By tackling these barriers, we can maximize GenAI's potential to facilitate innovative, effective, and equitable solutions to critical healthcare challenges across the globe.

RESEARCH TEAM

Stanford University, Center for Digital Health
& Center for Advanced Study in the Behavioral Sciences

CDH
STANFORD
CENTER FOR
DIGITAL HEALTH

CASBS
STANFORD UNIVERSITY
CENTER FO
ADVANCED
STUDY IN T
BEHAVIORA
SCIENCES

Isabella de Vere Hunt, MD

James Parkhouse, MA, PhD

Lara Rich, MEd

Kang-Xing Jin, BA

Mubarik Imam, MPA/ID, MBA

Jiyeong Kim, MPH, PhD

Mariana Ramirez Posada, MD

Michael Avanti Lopez, JD

Zachary Ugolnik, MTS, PhD

Sarah Soule, MA, PhD

Eleni Linos, MD, DrPH

ACKNOWLEDGEMENTS

We are grateful to Stanford University's RAISE Health Initiative for their feedback and contributions to this White Paper.

Funding:

This paper was drafted with support from Advancing Health Online (AHO). AHO, a fiscally sponsored project of Global Impact, is an initiative launched and financially supported by Merck & Co. Inc. Rahway, NJ USA (known as MSD outside the United States and Canada) and Meta in June 2021 to advance public understanding of how social media can be utilized to better understand and increase the health and resiliency of communities around the world by bringing actors together from technology, health, global development, and the academic sector. AHO's goal is to support effective integration of social media as a core component of social behavior change (SBC) to improve health outcomes. To achieve this, AHO collaborates with organizations working at the intersection of technology, global health, and SBC.

Core collaborators:

Laiah Idelson, Strategic Partner Manager, Health, Meta
Mark Allen, Executive Director, Health Equity, Strategic Partnerships and Digital & Data Solutions, MSD
Jessica Katz, Partner, Rabin Martin
Thea Lacerte, Senior Consultant, Rabin Martin
Allyson Arocha, Consultant, Rabin Martin

AHO
ADVANCING HEALTH ONLINE

With special thanks to our dissemination partner, Bay Area Global Health Alliance:

Sara Anderson, Executive Director
Lisa Foster, Senior Advisor

BAY AREA
Global
Health
ALLIANCE

We are incredibly grateful to the teams at Jacaranda Health, Viamo, Girl Effect, Audere and Noora Health for all the time and effort spent liaising with our team to provide the data detailed in the case studies. We would also like to thank the Turn.io team for sharing their accelerator data with us to include in our scoping analysis, and Chris Struhar for his role in conceptualizing the LLM task strengths framework.

Stanford Workshop, October 2024

MAYA ADAM, Director of Health Media Innovation and Clinical Associate Professor, Department of Pediatrics, Stanford School of Medicine

TILL BÄRNIGHAUSEN, Alexander von Humboldt Professor & Director of the Heidelberg Institute of Global Health, Heidelberg University; Senior Faculty, Africa Health Research Institute; Adjunct Professor of Global Health, Department of Global Health and Population, Harvard T.H. Chan School of Public Health

SANJAY KINRA, Professor of Clinical Epidemiology, London School of Hygiene & Tropical Medicine

LISA BOURGET, Senior Director, Strategy, Management, and Partnerships, Duke Global Health Innovation Center

NICOLE MARTINEZ-MARTIN, Assistant Professor, Stanford Center for Biomedical Ethics

NEESH PANNU, Vice Dean of Research, Faculty of Medicine and Dentistry, University of Alberta; Alberta Health Services Chair in Health Informatics

DREW BERNARD, Digital Communications Leader, Team Upswell

SHAWNA COOPER, Director of Product, Audere

STEPHEN MEYER, Director of Partnerships, Viamo

CK CHERUVETTOLIL, Former Senior Strategy Officer at the Bill & Melinda Gates Foundation

NATALIA LINOU, Policy specialist, UNDP

SARA ANDERSON, Executive Director of the Bay Area Global Health

Global Digital Health Forum Roundtable, December 2024

ANNE MAKENA, Co-Director, Africa Oxford Initiative

DINO RECH, CEO, Audere

BILAL MATEEN, Chief AI Officer, PATH

RACHEL SIBANDE, Artificial Intelligence-Africa Senior Program Officer, Gates Foundation

KARINA RIOS MICHEL, Chief Creative and Technology Officer, Girl Effect

SOMA MITRA-BEHURA, Lead AI Researcher, Girl Effect

YASMIN CHANDANI, CEO, InSupply Health

JAY PATEL, Chief Technical Officer, Jacaranda Health

PAUL MACHARIA, Research Scientist, Kenyatta National Hospital & University of Nairobi

ISABELLE AMAZON-BROWN, Independent Designer and Researcher, Ethical Chatbots & AI

DANIEL FUTERMAN, Head of Engineering, Reach Digital Health

WINNIE KARANU, AI National Skills Director, Microsoft

ALLYSON AROCHA, Consultant, Rabin Martin

Semi-structured interviews, November-December 2024

SHAWNA COOPER, Director of Product, Audere

SATHY RAJASEKHARAN, Co-executive director of Jacaranda Health

NNEKA MOBISSON, Co-Founder and CEO of mDoc

STEPHEN MEYER, Director of Partnerships, Viamo

BRIAN DERENZI, VP, Research and AI, Dimagi

KARINA RIOS MICHEL, Chief Creative and Technology Officer, Girl Effect

ALEX FULCHER, Senior Director of Technology, Girl Effect

SOMA MITRA-BEHURA, Lead AI Researcher, Girl Effect

ISABELLE AMAZON-BROWN, Independent Designer and Researcher, Ethical Chatbots & AI

ANUBHAV ARORA, Co-Executive Director - Programs & Platforms, Noora Health

GABRIELLE ARRUDA, Lead for digital health messaging, ImpulsoGov

DANIEL FUTERMAN, Head of Engineering, Reach Digital Health

EMILY MANGONE, Program Officer, Family Planning, Gates Foundation

ZAMEER BREY, Deputy Director, Technology Diffusion, Gates Foundation

TOPAZ MUKULU, Strategy Analyst, Patrick McGovern Foundation

GUSTAV PRAEKELT, Co-founder, turn.io

ALEX NAWAR, Social Impact Lead, OpenAI

MERRICK SCHAEFER, Director of the Center for Innovation and Impact, USAID's Global Health Bureau

SEAN BLASCHKE, Senior Health Specialist, Digital Health & Information Systems, Unicef

MANISH PANT, Policy Specialist, Digital Health, UNDP

ELIZABETH SHAUGHNESSY, Director of Digital Programming; Co-lead of NetHope's AI Working Group, NetHope

ROBERT KOROM, Chief Medical Officer, Penda Health

DAPHNE NGUNJIRI, Chief Executive Officer, Access Afya

BILAL MATEEN, Chief AI Officer, PATH

NICOLE SPIEKER, Chief Executive Officer, PharmAccess

LINDA RAFTREE, Founder, The MERL Tech Initiative

APPENDIX

In our scoping analysis, we reviewed the following accelerators or similar grant programs:

Major tech platforms

- Meta: Llama Impact Grants and Llama Impact Innovation Awards
- OpenAI/Turn.io/Agency Fund: Chat for GenAI Accelerator
- Google: AI for the Global Goals; Accelerator: Generative AI; and Startups AI for Health programs
- Microsoft: AI Accessibility Innovation Grants

Major non-governmental and non-IGO public health funders:

- The Gates Foundation: Global Grand Challenges
- The Patrick McGovern Foundation: AI in Digital Health Grants
- The Global Fund: AI for Tuberculosis Diagnosis initiative

International Development and Government-Funded AI Programs:

- Canada IDRC: Artificial Intelligence for Global Health and AI for Pandemic & Epidemic Preparedness programs

Social Impact and Startup-Focused Accelerators:

- Fast Forward: Startup Accelerator
- Data.org: Generative AI Skills Challenge

REFERENCES

1. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI [Internet]. Rochester, NY; 2023 [cited 2024 Oct 3]. Available from: https://papers.ssrn.com/abstract=4443189

2. Friedland A. What Are Generative AI, Large Language Models, and Foundation Models? [Internet]. Center for Security and Emerging Technology. 2023 [cited 2024 Dec 6]. Available from: https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

3. World Bank Country and Lending Groups – World Bank Data Help Desk [Internet]. [cited 2024 Dec 26]. Available from: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

4. Short SE, Mollborn S. Social Determinants and Health Behaviors: Conceptual Frames and Empirical Advances. Curr Opin Psychol. 2015 Oct;5:78–84.

5. Meng XL. Data Science and Engineering With Human in the Loop, Behind the Loop, and Above the Loop. Harvard Data Science Review [Internet]. 2023 Apr 27 [cited 2025 Jan 5];5(2). Available from: https://hdsr.mitpress.mit.edu/pub/812vijgg/release/3

6. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-Augmented Generation for Large Language Models: A Survey [Internet]. arXiv; 2024 [cited 2025 Jan 30]. Available from: http://arxiv.org/abs/2312.10997

7. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A Comprehensive Overview of Large Language Models [Internet]. arXiv; 2024 [cited 2025 Feb 7]. Available from: http://arxiv.org/abs/2307.06435

8. What Is an API (Application Programming Interface)? | IBM [Internet]. 2024 [cited 2025 Feb 6]. Available from: https://www.ibm.com/think/topics/api

9. Global Burden of Disease 2021: Findings from the GBD 2021 Study | Institute for Health Metrics and Evaluation [Internet]. [cited 2025 Feb 7]. Available from: https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study

10. Health-related SDGs [Internet]. Institute for Health Metrics and Evaluation. [cited 2025 Feb 7]. Available from: https://www.thelancet.com/lancet/visualisations/gbd-SDGs

11. Our work: communicable and noncommunicable diseases, and mental health [Internet]. [cited 2025 Feb 7]. Available from: https://www.who.int/our-work/communicable-and-noncommunicable-diseases-and-mental-health

12. Non communicable diseases [Internet]. [cited 2025 Feb 7]. Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

13. Environmental health [Internet]. [cited 2025 Feb 7]. Available from: https://www.who.int/health-topics/environmental-health

14. U.S. Agency for International Development. Health Systems Strengthening | Global Health | Health Health Systems and Innovation [Internet]. [cited 2025 Feb 7]. Available from: https://www.usaid.gov/global-health/health-systems-innovation/health-systems-strengthening

15. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025 Feb 5;388:e081554.

16. EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act [Internet]. [cited 2025 Feb 7]. Available from: https://artificialintelligenceact.eu/

17. Partnerships will ensure inclusivity for Nigeria's AI strategy [Internet]. [cited 2025 Feb 7]. Available from: https://www.luminategroup.com/posts/news/partnerships-nigeria-ai-strategy

18. Kenya launches project to develop National AI Strategy in collaboration with German and EU partners | Digital Watch Observatory [Internet]. 2024 [cited 2025 Feb 7]. Available from: https://dig.watch/updates/kenya-launches-project-to-develop-national-ai-strategy-in-collaboration-with-german-and-eu-partners

19. INDIAai | Pillars [Internet]. IndiaAI. [cited 2025 Feb 7]. Available from: https://indiaai.gov.in/

20. ebia-summary_brazilian_4-979_2021.pdf [Internet]. [cited 2025 Feb 10]. Available from: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-summary_brazilian_4-979_2021.pdf

21. 2024 RAISE Health Summary Paper [Internet]. Google Docs. [cited 2025 Mar 6]. Available from: https://drive.google.com/file/u/0/d/1Gu8Q2FFktCA1f-F-JG04SunHOnPxvkY2/view?pli=1&usp=embed_facebook

22. February 2025 CC// 10. Exclusive: Donors commit $10M to include African languages in AI models [Internet]. Devex. 2025 [cited 2025 Feb 12]. Available from: https://www.devex.com/news/sponsored/exclusive-donors-commit-10m-to-include-african-languages-in-ai-models-109044

23. From Connectivity to Services: Digital Transformation in Africa [Internet]. World Bank. [cited 2025 Feb 7]. Available from: https://projects.worldbank.org/en/results/2023/06/27/from-connectivity-to-services-digital-transformation-in-africa

24. The-Mobile-Gender-Gap-Report-2024.pdf [Internet]. [cited 2025 Jan 6]. Available from: https://www.gsma.com/r/wp-content/uploads/2024/05/The-Mobile-Gender-Gap-Report-2024.pdf?utm_source=website&utm_medium=button&utm_campaign=gender-gap-2024

25. Large Language Models for Health Equity [Internet]. [cited 2025 Feb 7]. Available from: https://www.path.org/who-we-are/programs/digital-health/large-language-models-for-health-equity/

26. AfriMed-QA [Internet]. [cited 2025 Feb 7]. Available from: https://afrimedqa.com/

27. AI Index Report 2024 – Artificial Intelligence Index [Internet]. [cited 2025 Jan 6]. Available from: https://aiindex.stanford.edu/report/

28. Global Digital Health Monitor [Internet]. Global Digital Health Monitor. [cited 2024 Dec 8]. Available from: https://digitalhealthmonitor.org

29. AI White Paper [Internet]. Reach. [cited 2024 Dec 8]. Available from: https://www.reachdigitalhealth.org/ai-white-paper

30. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. The Lancet Digital Health. 2025 Jan 1;7(1):e64–88

Stanford MEDICINE

Department of Medicine