# Research Design: Medical AI Chatbot Intervention for Maternal and Child Health in Sierra Leone

## Table of Contents

# Introduction

A large language model (LLM) is a type of artificial intelligence (AI) system that uses machine learning algorithms trained on a very large corpora of text data to generate natural language. Recent advances in LLMs have opened the door to a wide range of applications that have the potential to benefit under-resourced communities, particularly in low-income countries. Currently, one of the most popular applications is ChatGPT, a chatbot introduced by the US-based AI research laboratory OpenAI in 2022 (OpenAI Blog, 2022). The capabilities of the chatbot have caught the public's attention, but the technology is still in its infancy. As the sophistication of the models scale, we should expect to see entire industries disrupted.

Until recently, concerns about AI disruption were focused on blue-collar work like food service and truck driving, but it's becoming clear that the industries that are the most at risk for disruption by LLMs are white-collar knowledge work like engineering, law, and medicine (Lowrey, 2023). It stands to reason then, that the communities that could gain the most from increased access to these ivory towers of knowledge are the ones that until now, have been cut off from them.

In fields like medicine, these under-resourced communities stand to benefit the most from increased access to knowledge. According to World Health Organization (WHO) statistics, the global country average physician ratio is 588 to 1. For the bottom 10% of countries, that gap is 25 times greater at 14,286 to 1 (WHO, n.d.). In order to bridge these divides, they need intermediary solutions. One such potential solution is a medical question and answer chatbot.

Maternal and child health stands out as an acute issue for under-resourced communities in low-income countries that could be helped by a pregnancy question and answers chatbot. According to WHO statistics, the global average maternal death rate is .2 in 100, but for countries in the top 10%, it's nearly 5 times greater, at 1 in 100 (WHO, n.d.). The global average infant death rate is 1.7 in 100, but for countries in the top 10%, it's over 3.5 times greater, at 6 in 100 (WHO, n.d.). Nowhere are these divides greater than in Sierra Leone, where 1.7 in 100 mothers die in childbirth and 8 in 100 children die before the age of 1. For this reason, I'm proposing a pregnancy question and answers chatbot

Given the potential for harm in interventions that disrupt established healthcare practices, it's important to first try to understand and assess its impact. For this reason, I'm proposing a pilot project to first validate the average truth value of the answers provided by the pregnancy question and answers chatbot, then deploy it in a two-armed intervention to a group of users to measure its impact on maternal and child health outcomes over the course of a pregnancy.

# Research Question

The broad research question underlying this proposal and future research in this space is, “How might applications of this new generation of LLMs impact the world’s poorest communities?” To contribute to this broader question, this evaluation will aim to answer a more narrow one: “How will a pregnancy questions and answers chatbot built using this new generation of LLMs impact maternal and child health outcomes in Sierra Leone?”

# Literature Review

To answer that, this project will leverage and build on three distinct bodies of research. Given that these bodies of research cover multiple academic fields–including economics, public health, and computer science–I’ve chosen to dedicate more space in this proposal to introducing and reviewing each below. The first is the most niche, focusing on the methodology of recruiting research subjects through Facebook ads and deploying interventions through Facebook messenger. The second is the largest, focusing on the use of chatbots in healthcare. The third is the newest, focusing on how to constrain LLMs functionality so as to produce chatbots that give factual and on-topic responses.

## Sampling and surveying using Facebook

The body of research on recruiting research subjects through Facebook ads and deploying interventions through Facebook messenger is largely from research by Leah Rosenzweig, who I am a research assistant under at the Development Innovation Lab (DIL) at the University of Chicago. Rosenzweig and coauthors highlight the advantages of this method, including its low cost and accessibility, given its ability to reach populations in low-income countries without the need for in-person survey teams. They also discuss the challenges and limitations, such as potential biases in the sample and the need for careful targeting of the advertisements (Rosenzweig et al., 2020). The methodology has been developed further through an ongoing chatbot intervention study to improve vaccine hesitancy in Kenya and Nigeria (Rosenzweig and Offer-Westort, 2022). I consulted with Rosenzweig directly regarding this proposed research design and she provided guidance and resources.

## Chatbot applications in healthcare in LMICs

The body of research on the use of chatbots in healthcare is foundational to my proposed intervention’s theory of change. A meta-analysis of eight developmental studies that assessed the practicality and usefulness of chatbots, and seven interventional studies found that while the first studies testing chatbots for public health seem very promising, there are aspects that should be improved, including the chatbots' designs, studies' methods, and analysis and reporting of results, concluding that more high-quality studies and improved reporting of chatbots' use are needed (Gabarron et al., 2020). Another meta-analysis of 12 studies

assessing the impact of mental-health chatbots found weak evidence that chatbots were effective in improving depression, distress, stress, and acrophobia, but no statistically significant effect of using chatbots on subjective psychological well-being (Ali Abd-Alrazaq et al., 2020). A study that evaluated 78 health-related applications with chatbots used in 33 countries found that they play a crucial role in addressing quality healthcare gaps, but they are still in an early stage of development. Further research on their development, automation, and adoption is needed to achieve a population-level health impact (Parmar et al., 2022).

In addition to meta analysis, there have been efforts to measure the effects of medical chatbots with randomized controlled trials (RCTs). A blind, non inferiority RCT studying whether an artificial conversational agent was able to provide answers to patients with breast cancer with a level of satisfaction similar to the answers given by a group of physicians found the chatbot's answers non-inferior (P<.001) (Bibault et al., 2019). A three-armed evaluation of a fertility awareness and preconception health chatbot found positive effects in the chatbot- and document-based treatments (+9.1pp and +14.9pp), as compared to the control (+1.1pp), but received feedback about the chatbot limitations in comprehension and coldness (Maeda et al., 2020).

Since the release of OpenAI's Generative Pre-trained Transformer (GPT) models in 2020 via their application programming interface (API), researchers have been building and evaluating more conversational chatbots. Results from a preliminary study on a conversational agent–built using GPT-3, fine-tuned on 0.26 million dialogues between patients and doctors–designed to promote well-being showed that the agent was perceived as helpful and engaging (Yan and Nakashole, 2021). Another evaluation of a chatbot–built using GPT-2, fine-tuned with 306 therapy session transcripts–designed to provide emotional support and psychoeducation to caregivers of people with dementia found that the chatbot was effective in reducing the caregivers' perceived burden and improving their quality of life (Wang et al., 2021). Lastly, a recent feasibility study on the utilization of GPT-3 in public health found that the AI tool has the potential to enhance various aspects of the public health field, including disease diagnosis, health education, and communication with patients. However, the study also highlighted the need for further research and consideration of ethical and privacy concerns before implementing GPT-3 in public health settings (Jungwirth and Haluza, 2023).

## Making GPT-3 responses more factual

If you were to ask GPT-3 for an answer to a rigorously understood and documented question like, "How can I tell how far along my pregnancy is?", you're likely to get a good answer. Major problems begin to arise when you ask it for answers to less understood and documented questions. Current models tend to begin "hallucinating", whereby they fill in gaps of knowledge with confabulations (Harford, 2023). For a medical questions and answers chatbot, this could lead to them causing more harm than benefit. In time, the sophistication of the models will likely increase, and confabulations should become less of a problem, but for now, it limits the usability of the base model for answering questions in important fields like medicine (Doshi and Bajaj, 2023).

Research with LLMs on how to constrain their functionality so as to produce chatbots that give factual and on-topic responses is still in its early stages. Nevertheless, two approaches have been recently developed by OpenAI, each having its own benefits and limitations. OpenAI's fine-tuning command line interface (CLI) tools allow researchers to train their models on many more examples than can fit in the prompt, letting them achieve better results on many tasks (OpenAI Research, 2022). OpenAI's Text Embedding API endpoint converts strings of text into number sequences, which makes it easier for computers to understand the relationships between different texts (OpenAI Blog, 2022). This approach allows researchers to fit a group of texts from a knowledge base into the prompt to the model, along with instructions to, "answer the question as truthfully as possible using the provided context, and if the answer is not contained within the text below, say 'I'm sorry, that isn't within my expertise'". Nevertheless, this approach can only produce responses that are as factual and on-topic as the knowledge base is.

# Application Design

Given that this research proposal is meant to help answer a broader question about LLMs' impact on the world's poorest communities, it was important that the chatbot have a general design, showcasing the LLM first and foremost. For this reason, I stuck with a simple implementation with an emphasis on producing factual and on-topic answers. This implementation involves embedding each question generated by a user in a larger prompt that instructs the model to only respond if the question is related to pregnancy and to generate an answer based on a group of related facts drawn from a larger knowledge base.

## Coding the knowledge base

The chatbot uses OpenAI's Text Embeddings API endpoint (OpenAI Documentation, 2023). For the endpoint's API parameters, the chatbot uses the "text-embedding-ada-002" text embedding model, which is the current highest-performing model offered. Calling this endpoint on a string of text produces a vector (list) of 1536 floating point numbers. The distance between two vectors measures their relatedness, with small distances suggesting high relatedness.

The data source for the knowledge base comes from Kaggle, a platform for data science and machine learning enthusiasts to explore, analyze, and share data sets (Kaggle documentation, 2023). "Diagnose me" is an long-form question answering (LFQA) dataset of dialogues on Kaggle between patients and doctors based on factual conversations from icliniq.com and healthcaremagic.com that aims to collect more than 257k of different questions and prescriptions for patients.

To produce the knowledge base, the 257k patient questions from the medical dialogues dataset are passed through the endpoint and the resulting vectors are stored in a matrix, forming the

knowledge base. To generate the related facts, the user-imputed text is transformed and passed through the endpoint and then the resulting vector is compared against the knowledge base by taking the dot product of the two matrices. The doctor answers that were associated with the top five resulting values from the knowledge base are then extracted to be included in the prompt.

## Programming the chatbot responses

To generate the text completions, the chatbot uses OpenAI's Text Completions API endpoint (OpenAI Documentation, 2023). For the endpoint's API parameters, the chatbot sets the text completion model to "text-davinci-003", which is the current highest-performing model offered. The chatbot sets the endpoint's temperature value to 0. Temperature values can range from 0 to 1, with values like 0.8 making the output more random, and lower values like 0.2 making it more focused and deterministic. Lastly, along with model and temperature parameters, the chatbot passes a prompt, which it generates by wrapping the user-imputed text and the related facts in a larger prompt, which is structured as follows:

> Q: <user-inputted text>
>
> Answer the question above as truthfully as possible using the context below. If the answer is not contained within the context below, say "I'm sorry, that isn't within my expertise."
>
> <related facts>
>
> A: <response text>

## Hosting the knowledge base, source code, and application

The source code for the chatbot is hosted in a publicly available GitHub repository, which is a storage space on the GitHub platform where developers can store and share files, directories, documentation, and other resources related to a project (Github Documentation, 2023). The dataset of 257k doctor-patient dialogues is too large to be stored on GitHub, so it has been coded to be downloaded from the web to a local repository in order to produce the embeddings matrix. The embeddings matrix is also too large to be stored on GitHub, so it needs to be hosted on the web. This application stores the embeddings matrix using the free tier of a Pinecone vector database, which is a vector database service that makes it easy to build high-performance vector search applications (Pinecone Documentation, 2023).

The framework for the chatbot is built using Flask, which is a Python web framework that allows developers to quickly build web applications with minimal dependencies and a simple syntax (Flask Documentation, 2023). The web application for the chatbot is hosted on Heroku, which is an easy-to-use cloud platform that allows developers to deploy, manage, and scale web applications (Heroku Documentation, 2023). Apart from requiring a Heroku account to host the web application and a Pinecone account to host the embeddings matrix, the code for the current

iteration of the chatbot is completely open-source and available on GitHub (dustinmarshall, 2023). Contained in the repository, is over 400 lines of code, written in python, javascript, html, and css. The base folder contains a README.md file with complete replication instructions, which reads as follows:

1. Clone this repository to your local machine
2. Open a terminal and navigate to where the repository is saved
3. Run the following code from the command line to securely save private variables associated with your Kaggle, OpenAI, and Pinecone accounts to your local environment:
   “export KAGGLE_KEY=YOUR-KEY-HERE”
   “export OPENAI_API_KEY=YOUR-KEY-HERE”
   “export PINECONE_API_KEY=YOUR-KEY-HERE”
   “export PINECONE_ENVIRONMENT=YOUR-ENVIRONMENT-HERE”
4. To download the dataset from Kaggle, run the following code from the command line:
   “kaggle datasets download -d dsxavier/diagnoise-me”
5. To clean the doctor-patient dialog data, run the following code from the command line:
   “python3 /embeddings/clean_data.py”
6. To compute the embeddings and store them in your Pinecone Index, run the following code from the command line:
   “python3 /embeddings/compute_embeddings.py”
7. To create the app on Heroku and link it to your existing GitHub repo, run the following code:
   “python3 /application/create_app.py”

The current iteration is fully functional and available to use at docgpt.herokuapp.com. I encourage you to take a moment to visit the website so that you can experience the chatbot application directly. After doing that, I would also suggest that you visit the GitHub repository so that you read the source code and review the instructions for replication.

## Validating the truth value of the model’s responses

In order to trust the truth value of the model’s responses, we have to first first trust the truth value of the underlying medical dialogues dataset. The doctor-patient dialogues originate from two leading online medical advice services. HealthcareMagic is an online Q&A based medical advisory service with a global network of more than 18000 certified doctors from over 80 specialties (HealthcareMagic About Us, 2023). iCliniq is a Medical Second Opinion platform where users can get medical advice from medical practitioners, physicians and therapists from US, UK, UAE, India, Singapore, and Germany (iCliniq About Us, 2023).

In order to validate that the model is correctly generating responses using the knowledge base, the model needs to to be manually tested. To do this, I plan to take the following steps:

1. Take a random sample of 1000 questions from the medical dialogues dataset and pass the questions through my model to produce a response.
2. Take the 1000 responses and run them through the OpenAI embedding endpoint to produce a text embedding vector for each.
3. Take the 1000 medical dialog responses and run them through the OpenAI embedding endpoint to produce a text embedding vector for each.
4. Take the dot product of each response pair to produce a similarity score.
5. Take the average of the 1000 similarity scores.

The resulting average similarity score will give me a better understanding of how good the model is at responding to questions using the knowledge base. Having done this–and taking onboard the assumption that the truth value of the knowledge base itself can be trusted–we can feel more confident moving forward with the intervention as planned.

## Transitioning from proof-of-concept to real world application

The current iteration of the chatbot is only meant to serve as a proof-of-concept for this proposal. The chatbot application that is being actually proposed in this proposal will deviate from the current iteration in a few important ways. The first difference will be that the actual chatbot will restrict responses to being only about pregnancy. The second difference will be that, instead of having a custom interface and domain, its interface will be a Facebook Business page and research participants will interact with it through Facebook Messenger. On the backend, the source code will still be hosted on GitHub, but the web application will be hosted on ChatFuel, which is a cloud-based chatbot builder platform that allows users to create and deploy chatbots on messaging platforms like Facebook Messenger (Chatfuel Documentation, 2023). Hosting the application on ChatFuel will allow us to run pre- and post-intervention surveys through Facebook Messenger. The details of which I'll outline below.

# Research Design

I am proposing that pregnant women in Sierra Leone in their first trimester of birth be recruited via Facebook Ads to participate in the intervention. Individuals that opt-in will complete a pre-intervention survey to collect baseline demographic and maternal health information and select for individuals that live in Sierra Leone and are in their first trimester of pregnancy. Selected individuals will then be sorted into either the control group and receive nothing, or the treatment group. The treatment group will be introduced to the pregnancy questions and answers chatbot and invited to use throughout the remainder of their pregnancy.
Post-pregnancy, the individual will be invited to participate in a post-intervention survey to assess the maternal and child health outcomes of the intervention.

## Targeting with Facebook Ads

Facebook's detailed targeting is a targeting option available in the "Audience" section of ad set creation that allows you to refine the group of people we show your ads to. While you can't target pregnant women in Sierra Leone directly with detailed targeting, there are ways to target them indirectly by choosing to target people living in Sierra Leone,  then choosing "Parenting" and "Motherhood" as targeted interests, then choosing "New parents (0-12 months)" as the targeted demographic. If you wanted to increase the proportion of pregnant women that respond to your ad, you could also search for Facebook groups or pages related to pregnancy that pregnant women might probably follow, but this would likely introduce too much selection bias into the sample. For this reason, we'll limit the detailed targeting to women interested in "Parenting" and "Motherhood" and in the "New parents (0-12 months)" demographic.

## Soliciting Research Participants

The proposed chatbot application will be deployed via Facebook Messenger, from a private Facebook Business Page. Research participants will be solicited via a Facebook ad inviting pregnant women in their first trimester to participate in an informational intervention study for the duration of their pregnancy. The ad will explain that those who opt in will receive a small monetary compensation for completing a short pre-intervention survey and then receive another larger monetary compensation for completing a post-intervention survey after their pregnancy. Those who opt-in will be granted access to the private page and will receive a direct message from the page via Facebook Messenger to begin the survey.

## Completing the Pre-Intervention Survey

Participants will then complete a short pre-intervention survey to collect their demographic information, including their sex, age, ethnicity/race, socioeconomic status, education level, marital status, place of residence, and languages spoken. Any respondents that are male or don't live in Sierra Leone will be compensated and exit the survey. The remaining participants will complete an additional survey to collect baseline maternal health information like date of their last missed period, whether they plan to receive any antenatal care (ANC) attendance in their pregnancy, whether they plan to receive any skilled birth attendance (SBA) during their birth, whether they plan to receive any postnatal care (PNC) attendance after giving birth, whether or not they meet a threshold of frequency of consumption of various food groups and the use of micronutrient supplements, and whether or not their medical history puts their pregnancy at high-risk of complications. Upon completion of the survey, they will receive compensation and be sorted into treatments and control groups.

## Sorting into Treatment and Control Groups

The control group will be notified that they will be followed up with via Facebook nine months from the date of their last missed period for a post-intervention survey. The treatment group will receive the same notification, in addition to being introduced and granted access to the

pregnancy questions and answers chatbot via the same Facebook Messenger window. Treated individuals will be notified that the chatbot is not intended to be a replacement for professional medical advice and that all of their interactions with the chatbot will be recorded for research purposes.

## Completing the Post-Intervention Survey

Then, nine months from the date of each individual's stated last missed period, they will be contacted again via Facebook Messenger to complete a post-intervention survey. At this point, treated individuals will no longer have access to the chatbot. The post-intervention survey will include questions like whether they received any antenatal care (ANC) attendance in their pregnancy and if so, how many visits, whether they received any skilled birth attendance (SBA) during their birth, whether they plan to receive any postnatal care (PNC) attendance after giving birth, whether or not they experienced any complications during their pregnancy, whether or not they experienced any complications during their birth, whether or not they gave birth to a full-term baby, and the date of the birth.