Girl Effect

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING VISION

TABLE OF CONTENTS

FOREWORD

Girl Effect is an international non-profit that builds media that girls want, trust, and need. Over the past decade, Girl Effect has been using technology in innovative ways to reach girls and address their challenges. From websites and chatbots to IVR services, Girl Effect constantly adapts and refines tools in response to technological evolution and changes in girls' habits.

Girl Effect operates three chatbots on WhatsApp and Moya: AI-powered **Big Sis**, which provides young people in South Africa with trusted, non-judgmental advice about sex, relationships, and other sensitive topics; **Bol Behen**, a menu-based chatbot in India that answers questions about sexual health and well-being in 'Hinglish' (a mix of Hindi and English); and **WAZZII** in Kenya, which addresses young people's questions in Sheng (a mix of Swahili and English). By enhancing our chatbots with Artificial Intelligence (AI) capabilities, Girl Effect has seen significant increase in engagement in content consumption.

Since the creation of Big Sis in 2018 and indeed in the last year alone, the world of artificial intelligence and machine learning (ML) has transformed massively with the availability of large language models and tools to leverage these models rapidly evolving. In this new world of natural language processing (NLP), Girl Effect has embarked on a transformative journey to re-evaluate the ways in which the organization utilizes AI and ML to better serve users, harnessing the tremendous advancements in AI and ML to redefine the landscape of support and empowerment for girls and young people globally. The promise of AI to act as a proactive mentor, particularly through our envisioned chatbots, is not merely an ambition; it is the next critical step to unlock the power of girls and young people.

Recent breakthroughs in NLP have been nothing short of revolutionary. The emergence of generative AI, with pivotal releases such as OpenAI's ChatGPT and subsequent large language models and APIs, has ushered in a new era of conversational AI. These developments have profoundly expanded the capabilities of chatbots, allowing for more nuanced, empathetic, and robust interactions, especially in the critical domain of Sexual and Reproductive Health (SRH) and Mental Health (MH).

The landscape is further enriched by experimentation within the open source community, propelled by the release of resources such as Meta's LLaMa. The rapid problem-solving agility of this community has addressed "major open problems," accelerating the pace of innovation and opening a multitude of possibilities for applications in social good.

This document explores the synergy of proprietary and open-source AI advancements that offer a plethora of developmental opportunities for SRH and MH chatbots. These innovations enable the simplification of chatbot infrastructures into dynamic databases for real-time, custom dialogue, the incorporation of unique Girl Effect personas into the user experience design, and improved engagement in diverse vernaculars, keeping services culturally relevant. AI's evolution further enables the implementation of 'Safety by Design' — increasing the ability to detect sensitive disclosures while maintaining ethical standards. Simultaneously, AI enhances the ability to link girls to life-saving services, allowing Girl Effect to offer information when she's ready and proactively predict when she may need a service.

Girl Effect's vision is clear: **to leverage AI and ML in creating chatbots that not only converse but care, understand, and guide**. It is not simply about technology; it is about harnessing it to build a safer, more accessible world where every girl can realize her potential. This vision lays the foundation for that future, mapping out the strategies, innovations, and dreams that will drive Girl Effect forward in this new digital age. It can also serve as a starting point for others in the sector who hope to integrate new AI and ML capabilities into their chatbots.

This journey is about transformation—of technology, of lives, and of societies. We invite you to join us as we use AI and ML not just to reach girls but to touch their lives profoundly and positively. Together, we will create ripples of change, where empowered girls can initiate a tidal wave of progress that will shape our world.

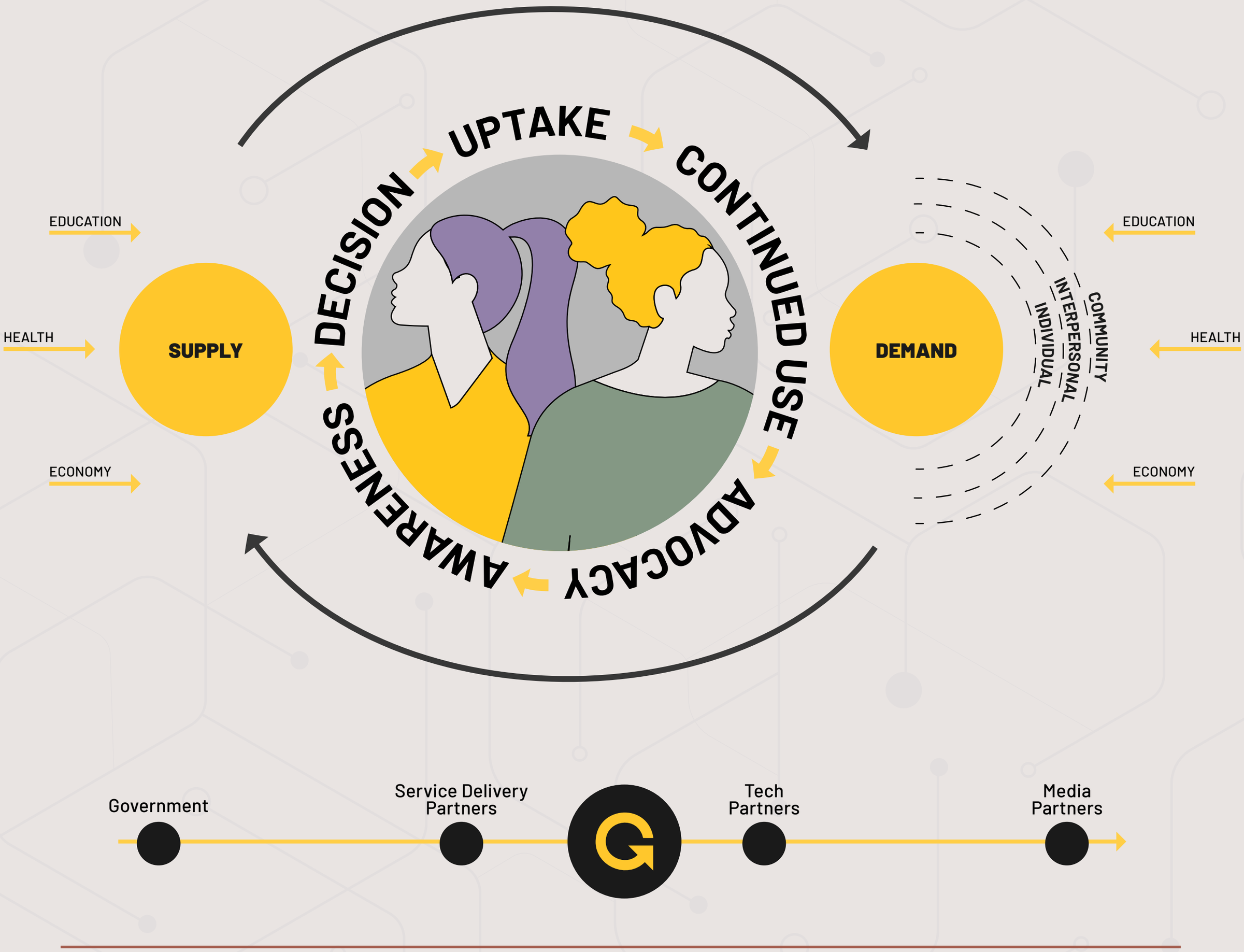
CHAPTER 1

The Role Of Chatbots In Health-Seeking Programming

THE ROLE OF CHATBOTS IN HEALTH-SEEKING PROGRAMMING

Girl Effect develops chatbots and conversational agents designed to promote health-seeking behaviors among young people. These tools aim to equip and inspire them with accurate information to make informed decisions about their sexual and mental health.

Figure 1. Girl Effect's High-Level Theory of Change



In the high-level Theory of Change displayed in Figure 1, Girl Effect primarily occupies the space of **demand generation**, using media and technology to engage girls at an individual level in the context of her world. Other organizations may occupy different roles especially in healthcare and service delivery that translate to different purposes and final design of chatbots used to serve users' FP and SRHR needs.

The following sections lay out the context for Girl Effect's decisions in purpose and design but may offer an example of how to think about the use of chatbots towards larger impact goals.

Ch 1 .1 Discussion in Girl Effect's Behavior Change Ecosystem

Young people and especially girls face significant obstacles worldwide, but particularly in the Global South where girls face unique challenges including early marriage and motherhood, limited access to education, and societal norms which restrict their potential for growth and development.

In particular, these girls struggle to find accurate and reliable information about SRHR, MH, relationships,well-being and puberty.

In many of the countries where Girl Effect operates, girls are denied opportunities to openly discuss SRHR and MH topics, face social stigma and shame for their curiosity, and often receive inaccurate and even harmful information as a result. The information gap, digital divide, and other gender-based barriers encountered by these girls prevent many of them from making informed decisions about their health and well-being, reaching their full potential, and making meaningful contributions to their communities – all of which, in turn, contributes to intergenerational poverty and inequality in which many girls and women find themselves trapped.

These barriers are connected to various behavioral drivers in a behavior change ecosystem that includes personal, interpersonal, and environmental components. Girl Effect's Theory of Change is designed to build demand for life-saving services.

Figure 2. Girl Effect's Theory of Change for Demand Generation: this behavior change ecosystem includes our eight core drivers of behavior change.

Girl Effect's theory of change for demand generation includes eight core drivers of behavior change as visualized in Figure 2:

Self identity: Who do I see myself as? Am I the type of girl who does this?

Social identity: Where do I want to fit in? What do I think my peers do, and what do they expect I should do?

Outcome expectations: What will I gain if I do this? What will I lose?

Perceived social support: Do I think others will support and help me?

Self efficacy: Do I have the confidence and ability to do this?

Attitude: What do I think about this behavior? Is it good or bad?

Perceived control: Is it up to me? What external obstacles might stop me?

Knowledge: What do I know about this (facts, or how-to's)? What can I do (skills)?

Girl Effect's approach to broader societal norms change is based on the concept of reciprocal determinism, which states that, as the social environment influences a girl, a girl also influences her social environment. By inspiring and equipping girls to adopt behaviors deemed important to them, it is assumed that girls will influence other changes in their social and enabling environments, whether consciously or inadvertently. This is supported by the diffusion of innovation theory, which states that an innovation (a new thought or action) diffuses when it is positively experienced by others. This is exemplified in the Theory of Change, which positions girls as change agents: when a girl thinks and acts differently, this starts to influence those around her, including other girls. The Theory of Change outlines that, for social norms to shift, there needs to be a critical mass – a tipping point – after which a new norm becomes socialized. For example, if only 20% of girls in a community access SRH services, this may not be enough to start shifting the norm. If, however, 40% of girls start accessing SRH services, this could be the tipping point for accessing SRH services to become socialized as normal in the community and a new social norm will begin to be established: that adolescent girls of a certain age access SRH services. Where exactly the tipping point lies has not been fully established, and would differ between behaviors and contexts [1].

While Girl Effect has limited influence on the structural and environmental barriers identified, like poverty and access to SRH services, Girl Effect's tried-and-tested approach has proven successful in addressing many of the barriers mentioned using a behavioral product ecosystem that includes several channels like social media, TV, radio, podcasts, and chatbots.

Girl Effect's chatbots are a particularly critical component of this ecosystem as they provide a safe discussion space for girls. Girl Effect has found through extensive research and analysis that discussion has a positive relation with drivers like knowledge, perceived social support, attitude, self-efficacy, and self identity in multiple geographies. In Tanzania, 48% of those who discussed SRH had accessed services, compared to 16% of those who had not discussed SRH and in Rwanda, 64% of those who discussed SRH widely (all topics) had accessed services, compared to 49% of those who had not discussed all SRH topics.

Girl Effect has proven that discussion is an important contributor to behavior change and hypothesizes that creating chatbots that function as discussion spaces for young people may be a key method by which Girl Effect can scale impact. The content on Girl Effect's chatbots is already designed to affect the psychological barriers proven in other markets, but new innovations in AI and ML have now opened up opportunities for Girl Effect to increase the impact of this content.

Ch 1 .2 Purposes of Chatbots

Chatbots are developed for a large variety of purposes in mental health and sexual and reproductive health and rights like healthcare information dissemination, patient triaging and engagement, symptom checking, and appointment scheduling. These are just a few examples that reduce human work load and enable scalability of operations.

As described in Section 1.1, Girl Effect uses chatbots as a space for discussion for young people. This purpose can be broken down into more specific components:

- Provide accurate MH and SRHR information that is vetted by experts in the field
- Build users' confidence in taking charge of their own SRHR & MH journey
- Correct assumptions about the user's external ecosystem and options available
- Direct users to services when they are ready

These purposes feed into the design of a chatbot and also drive the use and prioritization of different AI and ML techniques in Girl Effect's chatbot infrastructure.

Ch 1 .3 Designing a Chatbot

The chatbot design lifecycle Girl Effect uses can be found in Figure 3. It includes formative research, user experience design, monitoring, and evaluation, phases that are iteratively cycled through to create better and better versions of the chatbot. Various phases in this lifecycle can be optimized using AI and ML techniques but a large focus of this vision will be the user experience of the chatbot, detailed in Section 2.

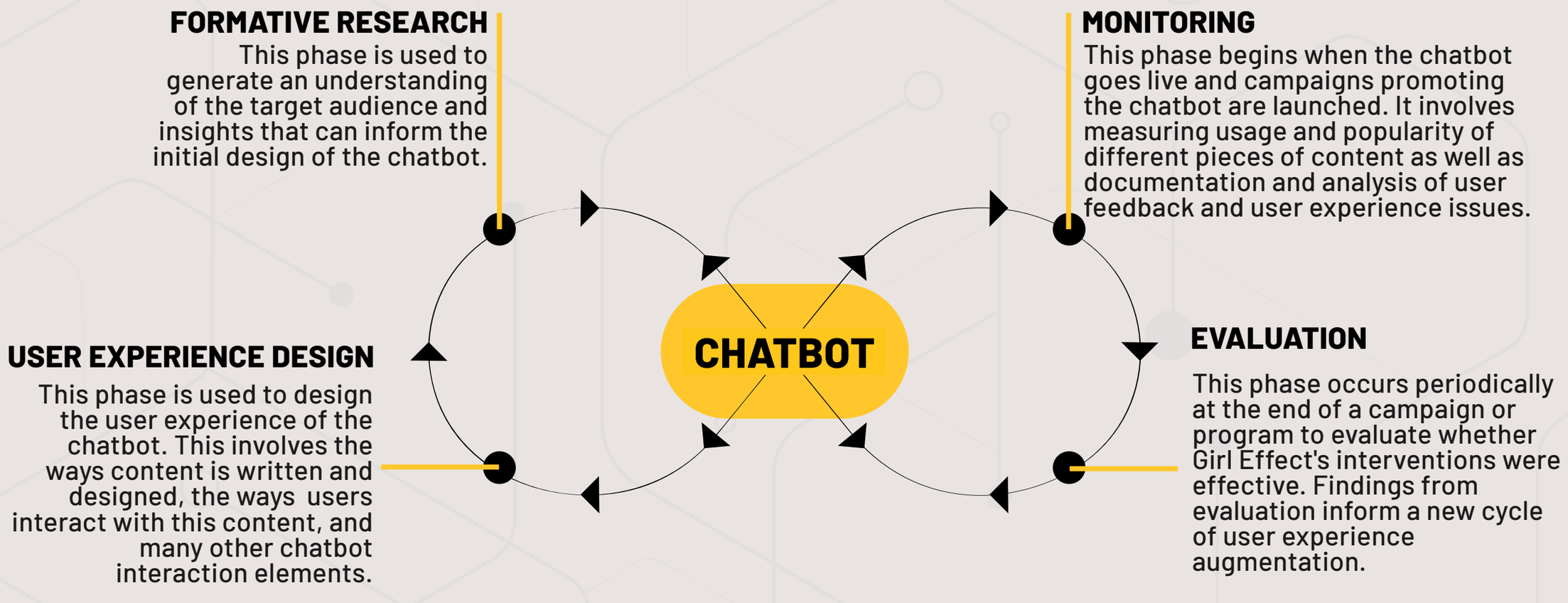


Figure 3. **Girl Effect's Chatbot Design Lifecycle**

Within Girl Effect's chatbot, message set design itself is grounded in Girl Effect's social ecological behavior change model. Different categories of messages (Personal Connection, Key Message, Engagement, Acknowledgement, Definition, Recognition of Feelings, Feedback, Refer, Info, Definition Check, Close) are stacked together to align with behavior change theory's best practices. These categories of messages that go beyond simple knowledge transfer demonstrate the balance that must be struck between a dialogue flow that a user engages with and a library of content that a user looks through. This complexity must be a consideration in how AI and ML techniques are integrated into the chatbot as well, making tradeoffs between the "natural" quality of generated text and the firmer guidance required to encourage girls on a positive behavior change journey. With these nuances acknowledged, the next section dives into the current state of AI and ML techniques.

CHAPTER 2

The Current State of Artificial Intelligence and Machine Learning

THE CURRENT STATE OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Broadly speaking, artificial intelligence (AI) is a wide-ranging concept that refers to machines or computers mimicking human intelligences, imitating tasks like learning, reasoning, problem-solving, and language understanding. Machine learning (ML) is a subset of AI that focuses on how systems can learn from data specifically, identify patterns, and make decisions with minimal human intervention. All the techniques detailed in the following subsections fall under ML because they depend on datasets, often massive in size, to perform tasks. AI and ML techniques are used to take data analysis beyond visual trend analysis and draw deeper insights – analytical and predictive – from broad sets of data that are complex and multi-dimensional.

This list of techniques is far from exhaustive. Girl Effect has focused on mentioning the techniques most relevant to chatbot development, but the field of AI and ML is far more vast. It should also be noted that the AI and ML landscape is rapidly evolving; techniques that are state-of-the-art now may grow obsolete in the timeframe of six months or less.

Ch 2.1 Supervised Learning

Supervised learning is responsible for most of the significant breakthroughs in AI, including machine translation, facial recognition, image classification, fraud detection, spam detection, and speech recognition [2].

Supervised learning systems learn to perform these tasks by analyzing tables of data including input and output data points and learning how to do one task from each table of data. If the same supervised learning system tries to learn to perform a task from a different table of data, it will forget everything it learned from the first table.

Supervised learning builds on long-used regression and classification models. In typical regression analysis, statistical methods are used to estimate the relationship between the input data points and output data points. This relationship is usually defined as a relatively simple function with a small number of fixed parameters and variables that is able to predict an output data point based on an input data point. Both regression and classification analysis involve fitting a pre-specified model to the data and using that model to predict the outcome for new data. For regression analysis, output data points must be numbers on a continuous spectrum. For classification analysis, the output data points are discrete categories. However, both regression and classification analysis perform poorly when modeling complex, non-linear patterns. Another limitation of supervised learning is that it depends on having a labeled dataset; labeling a large dataset can be a prohibitively cost- and time-expensive barrier to implementation. This limitation applies to both the older regression and classification algorithms as well as more modern supervised learning using deep neural networks which are discussed in Section 2.2.

Ch 2.2 Deep Learning Neural Networks

Deep learning neural network algorithms form the basis for modern artificial intelligence. These algorithms are the core of all the major types of artificial intelligence algorithms including supervised learning, unsupervised learning, and reinforcement learning [3].

An artificial neural network is a computing model whose layered structure resembles the neural structure of the human brain, providing myriad new ways for a computer to learn to solve problems. Similar to how neurons in the human brain receive inputs, process them, and pass the output to other neurons, the basic unit of a neural network is the neuron or node.

Given a dataset with input and output variables where multiple inputs determine the final output, the input layer of the neural network has one neuron for each input variable. The output layer has just one neuron that contains the final output. Between the input and output layers are hidden layer(s), termed "hidden" because they are not derived directly from the training dataset.

Each input neuron is connected to each neuron in the first hidden layer, each of the neurons in this hidden layer can be connected to the next hidden layer and so on until the neurons in the final hidden layer are connected to the final output neuron, a visual representation of which can be seen in Figure 4. Each neuron acts as a variable in a function and each connection between two neurons has a weight (also called parameter).

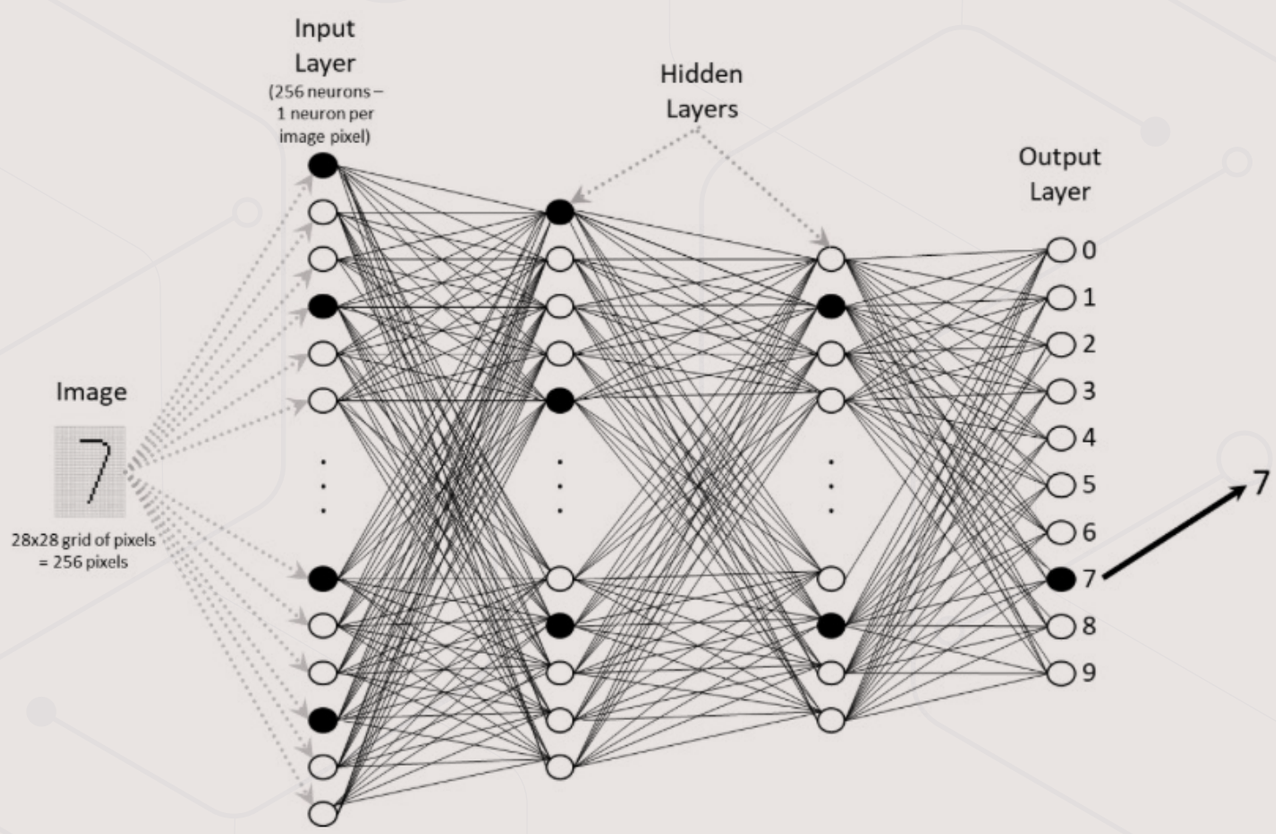


Figure 4. An example of a neural network used to classify handwritten numbers from the AI 101 textbook

At first, each of these connection weights is assigned a random value. Deep learning then involves using computational optimization methods to compute the optimal weights (or parameters). This process is essential to the development of language models which are discussed in Section 2.3.

Ch 2.3 Natural Language Processing and Large Language Models: Past and Present

Language models capture the likelihood of a sequence of words occurring for a particular language. These models are trained in an unsupervised fashion by analyzing massive amounts of text and seeing which words typically follow other words and word sequences. More specifically, language models are developed using an unsupervised learning technique known as self-supervised learning [4].

In supervised learning, discussed in Section 2.1, each row in the training dataset contains an output data point, also known as a label. Labels are typically added manually to a dataset and are often expensive and time-consuming to create. In contrast, in self-supervised learning, creative ways are found to use the data itself to create labels. In many language models, models are trained to predict each next word in a text. In self-supervised fashion, instead of using manually-generated labels, for each example text, the label is the next word in the text, providing its own supervision.

Most language models today are trained using a deep learning architecture known as a transformer which was invented in 2017 [5].

Ch 2.3.2 NLP and Language Models Prior to 2022

Prior to 2022, the most widely known language models included Generative Pre-Training (GPT) developed by OpenAI and Bidirectional Encoder Representations from Transformers (BERT) developed by Google AI Language, both released in 2018. BERT was primarily designed to provide high quality, context-aware word representations, making it excellent for tasks that require understanding the context in which words are used, like text classification (which Girl Effect has implemented in the chatbot Big Sis), named entity recognition, and question answering tasks. However, BERT alone is not designed to generate text, although its capabilities could be used in conjunction with other model components to generate text. While GPT could generate text, it had limitations, often producing nonsensical or inconsistent text particularly for longer form text. Many improvements were made to BERT in 2019 and 2020 including Carnegie Mellon University's XLNet, Meta's RoBERTa and BART, ByteDance's AMBERT, and Baidu's ERNIE but these were still not optimized for generating text.

GPT, BERT, and their language model successors still had to be fine-tuned to perform specific tasks. Task-specific fine-tuning involves further training the model using datasets for a specific NLP task like text classification. Through this training, parameters within the language model are adjusted, fine-tuned, to correctly classify text or perform any other task. This means that labeled datasets are still required to perform NLP tasks at high standards although these datasets can be significantly smaller than the datasets used to train systems specifically architected for the NLP task. For several NLP benchmarks, fine-tuned GPT outperformed systems designed for the task.

In 2019, OpenAI released GPT-2, a large language model that was its follow-on model to GPT, and then GPT-3 in 2020. Large language models build on language models by employing deep learning and particularly large numbers of parameters. Compared to GPT's 117 million parameters and BERT's 340 million parameters, GPT-2 has 1.5 billion parameters and GPT-3 has 175 billion parameters. Training these models requires access to massive computational power and so only companies with access to such finance could develop ever-larger and higher performance language models: Salesforce's CTRL in 2019 (168 billion parameters), AI21 Labs' Jurassic in 2021 (178 billion parameters), PanGu in 2021 (200 billion parameters), Inspur's Yuan in 2021 (245 billion parameters), Google Deepmind's Gopher in 2021 (280 billion parameters), Google's PaLM in 2022 (540 billion parameters), Google's Switch Transformer in 2022 (1.6 trillion parameters), and Beijing Academy of AI's Wu Dao (1.75 trillion parameters).

Training on massive numbers of parameters produced significant increases in performance for many NLP tasks. In addition to performance increases, GPT-2 and GPT-3 produced new emergent properties like the ability to generate text and incorporate few-shot learning.

For generating text, GPT-2 and GPT-3 (and more recent models) can be presented with the start of a text that was not in the training table. The model will use this text to predict the next word. Continuing, if the model is presented with the input text plus the word it generated, the model will predict the next word again until it generates several sentences.

GPT-3 is also able to exhibit strong performance on multiple NLP tasks like predicting the last word of a sentence, choosing the correct ending of a short paragraph, and translating sentences with especially high accuracy if few-shot learning is used. In few-shot learning, the language model (e.g. GPT-3) is given a few task examples that serve to "explain" the task that is to be performed to the language model. This means that GPT-3 (and later models) do not necessarily have to be fine-tuned to perform a specific task and can perform many NLP tasks with only a few examples provided, removing the large burden of providing labeled data to fine-tune a language model for a specific task like text classification. It should be noted that models that have been fine-tuned still typically perform better than a large language model using the few-shot approach.

Many improvements built on top of these large language models were made possible by the adoption of Parameter-Efficient Fine-Tuning (PEFT) techniques such as Low-Rank Adoption of Large Language Models (LoRA)[6]. A major cost- and time-expensive component of developing large language models is pre-training models with billions of parameters. As these models grow in size, full fine-tuning of all billions of parameters in the model becomes prohibitively expensive. Simply put, PEFT techniques like LoRA drop these costs significantly by freezing pre-trained model weights and instead adding a significantly smaller number of trainable parameters in layers of the language model that are particularly influential in producing better performance. Reducing the number of trainable parameters greatly reduces the computational cost and time of fine-tuning.

Beyond reduced computational cost and time, PEFT techniques are incredibly effective because they are stackable: improvements like instruction tuning can be applied and leveraged as other capabilities like reasoning or dialogue are added. Due to this, as new and better datasets and tasks become available, the model can be kept up to date cheaply rather than having to re-fine-tune the entire language model from scratch. Because PEFT updates are so cheap to implement, these models can be iterated on at a much faster rate.

Ch 2.3.2 NLP and Large Language Models 2022 Onwards

Recent times have seen the proliferation of both proprietary and open-source large language models. While open-source large language models have not yet caught up to the performance of proprietary models, they offer customization and flexibility options not available in proprietary models. Both will be discussed in the following subsections. It should be noted that the best option between proprietary and open source models will be which option the organization can continuously maintain and improve to the users' benefit. The resources needed to responsibly develop these models should also be considered in such a decision.

2.3.2.1 Proprietary Models:

Prior to 2022, many experts predicted that foundation models like GPT-3, BERT, LaMDA, and PaLM would still take five to ten years before they reached anywhere near the final stage of mainstream adoption[7]. This prediction was shattered in 2022 by OpenAI. Soon after the release of its successor foundational model GPT-3.5 in March 2022, OpenAI launched ChatGPT in November 2022, which built on the GPT-3.5 model[8]. ChatGPT was designed to work within policies based on human values. OpenAI's goal for ChatGPT was to make AI systems more natural and safe to interact with than GPT-3.5. To accomplish this goal, OpenAI fine-tuned the GPT-3.5 model using a technique called Reinforcement Learning from Human Preferences (described in more detail in Section 3.2.3.1). As a result, ChatGPT generates far more human-like responses than GPT-3 or GPT-3.5.

While GPT-3 had been used by NLP developers since 2021 when its API was made generally available, the public release of ChatGPT triggered a whole new class of large language model users on a larger scale than ever before who did not need niche coding knowledge to make use of large language models[9]. In March 2023, OpenAI also released GPT-3.5-turbo on their API [10]. GPT-3.5-turbo performs better than GPT-3 and is also priced at 10 times cheaper than previous GPT-3 models[11].

Around the same time that GPT-3.5-turbo was released on OpenAI's API, OpenAI also released its newest large language model, GPT-4, a model developed by focusing on aligning or training the model to much better follow user intentions[12, 13]. GPT-4 is more reliable, creative, and able to handle much more nuanced instructions than GPT-3.5 and the original ChatGPT, surpassing GPT-3.5 and the original ChatGPT in advanced reasoning capabilities and safety[14]. On top of this, GPT-4 is much better at generating nuanced multi-language text.

OpenAI has released few details on the training methods used to create GPT-4. Previously, GPT-4 was only available on OpenAI's API to select organizations off a waiting list but on 6 July, 2023, OpenAI released access to all paying API customers and announced plans to open access to new developers by the end of July. While GPT-4 is remarkable, especially in its multi-language capabilities, it is 20 times more expensive than GPT-3.5 for many use cases. This significant price differential requires analysis of which model should be used for which applications so that the use of GPT-4 can be limited to only absolutely necessary functions, optimizing costs while balancing use factor to end users.

While an API for GPT-3 did exist prior to 2022, its performance, cost, and technical expertise required meant only well-developed ML teams used it in apps (over 300 as of March 2021) [15]. What has revolutionized the practical use of large language models is OpenAI's release of several new APIs that allow OpenAI's models GPT-3, GPT-3.5, and GPT-4 to be simply and easily integrated into other applications with few restrictions and at low cost. These APIs include the chat completions API, the text embeddings API, the images API (using OpenAI's DALL·E models), the speech-to-text API, and the fine-tuning API. Newly released in November 2023, OpenAI's new offering, GPTs, now also allows "extra knowledge" to be directly uploaded as files for reference, removing the need to set up a separate system to use embeddings [16]. GPTs can also be created using OpenAI's Assistants API.

The APIs most relevant to chatbots are the following:

- **Chat completions API:** takes a list of messages as input and returns a specified model's generated message as output [17]
- **Text embeddings API:** used to measure the relatedness of text strings [18, 19]
- **Fine-tuning API:** used to customize OpenAI's large language models to specific tasks; we suspect this API utilizes PEFT techniques for task-specific fine-tuning [20]
- **Speech to text API:** used to 1) transcribe audio into whatever language the audio is in or 2) translate and transcribe the audio into English [21]
- **Assistants API:** used to build AI assistants within applications with instructions and extra knowledge [22]

While this section focused on OpenAI's offerings as the most accessible at the moment, there are several other proprietary options released or in development including Anthropic's Claude (released in March 2023 and considered comparable to ChatGPT) and Claude2 (released in July 2023 and likely to be as high performance as GPT-4), Cohere's Command (released), and Google's BARD (released in January 2023) that compete with OpenAI [23, 24].

2.3.2.2 Open Source Models:

While OpenAI's models currently remain state-of-the-art and relatively affordable for most applications, the space of open-source models is also transforming at astonishing rates. In 2022, multiple open-source large language models were released: EleutherAI's 20 billion parameter GPT-NeoX-20B in February 2022, Meta's 175 billion parameter OPT-175B in May 2022, and BigScience's Bloom available in a variety of sizes from 350 million to 176 billion parameters in July 2022 [25, 26, 27]. Designed to be as powerful as GPT-3, Bloom was soon after deployed on a free web app comparable to ChatGPT by AI startup Hugging Face, allowing anyone to try Bloom without having to download it [28].

While the open-source models released in 2022 are powerful, still none compare to the performance of ChatGPT and GPT-4, especially in generating sensical human-like text. While this remains true for now, recent developments in the open-source world signal the potential for rapid evolution of these models. In February 2023, Meta launched LLaMA, a 65-billion-parameter large language model, open sourcing the code but not the weights within the model [29].

Very quickly in March, LLaMa's weights were leaked, launching a flurry of experimentation by open-source developers [30]. In just a few months, many issues considered "major open problems" were solved: large language models can now be run on phones and personalized AI models can now be fine-tuned on one laptop in one evening, as just a few examples [31].

Open-source developments using PEFT have already produced remarkable results like Vicuna-13B, an open-source chatbot based on a large language model of only 13 billion parameters with a training cost of only $300 that beats Meta's LLaMA and Stanford's Alpaca in performance on several tasks [32]. Many open-source projects are saving time by training on small, highly curated open source datasets that produce models that have comparable performance to GPT-3.5 on many tasks. One example is Berkeley Artificial Intelligence Research's Koala [33]. While these models do perform well on many tasks, they generally do not have the breadth of knowledge and reasoning capabilities of larger models.

With the rapid pace of open-source development, it is important to keep open-source models in mind especially as they also allow cost reduction and customizability beyond the API access that is provided by proprietary companies like OpenAI and Cohere.

Ch 2.4 Other Machine Learning Techniques

While large language models are one of the most discussed applications of unsupervised learning techniques, there are other unsupervised learning techniques that can be used to enhance FP and SRHR chatbots [34]. Specifically in this section, reinforcement learning, cluster analysis and recommender systems are introduced.

Ch 2.4.1 Reinforcement Learning

Reinforcement learning (RL) has previously been primarily used to significant success in game-playing, robotics, and self-driving cars.

A reinforcement learning system can be considered an agent reacting to an environment in the following way:

1. **The agent takes actions according to the state of an environment defined by a starting behavior policy (like making a move in a game)**
2. **This environment periodically produces rewards (awarding points in a game)**
3. **The agent then modifies its behavior policy according to this reward**

The goal of reinforcement learning algorithms is to learn a behavior policy that maximizes the cumulative rewards over a period of time.

RL works well with problems that have "parse rewards". In problems like this, rewards are given infrequently (sparsely). More specifically, they are only generated after several actions have taken place. Because of this characteristic, RL has been found to be effective for NLP because the language models generate text word-by-word but the evaluation of a sentence does not occur until multiple words have been generated.

Ch 2.4.1 Cluster Analysis

In cluster analysis, the goal is to analyze large amounts of data and find groups of observations that are "similar." For chatbots, this might mean analyzing chatbot user interaction data to identify groups of users that are similar to one another.

These groups of users can then be targeted with tailor-made content or service offerings.

Clustering techniques start with a set of data that has input variables but no output variable. The task is to then find similarities in the data that define groups called clusters and place each user (or observation) in a cluster. To begin with, all data associated with a user or observation is mapped to a vector in a high-dimensional space. Cluster analysis algorithms can then group users or observations in a dataset into separate and distinct clusters in a high-dimensional space. The algorithm is designed to find the set of clusters that both maximizes the distance between the clusters in the high-dimensional space and also minimizes the distance between users or observations within one cluster.

The first cluster analysis algorithms were developed in the 1930s and had to be implemented by hand, limiting the number of observations and associated data that could be clustered. Now, statisticians and computer scientists have developed cluster analysis algorithms that can handle incredibly high-dimensional problems and massive training datasets.

Ch 2.4.3 Recommender Systems

Recommender systems are used to generate recommendations used by Amazon for book and shopping recommendations, by Pandora and Spotify for music recommendations, and by YouTube and TikTok for video recommendations, to name a few examples of their wide-ranging applications. Recommender systems begin with a table of rows of users and columns of items and cells that contain some type of interaction (in a chatbot, this could be engagement with a specific piece of content). The items can be anything like interactions with message sets or any other type of content-based interaction including the sentiment of a user-input message. In general, for each user, the table will have more empty cells than filled cells for the item columns like in the example shown in Figure 5. The goal of a recommender system is to be able to predict the user's values for the empty cells in the table.

		Jurassic	The		Pulp		Taxi
	Jaws	Park	Godfather	Casablanca	Fiction	...	Driver
User 1	5				3	...	
User 2	4					...	5
User 3				5	1	...	
...	...	...	...	...	...	...	...
User 17,632,592		4				...	2
User 17,632,593	4		5			...	2

Figure 5. An example user-item table for a movie recommender system [35]

There are two general types of techniques used for recommender systems: collaborative filtering and content-based techniques. Because most chatbots will not have enough content pieces to justify content-based techniques, only collaborative filtering is further explored here.

Collaborative filtering can be further broken down into user-user collaborative filtering and item-item collaborative filtering. User-user collaborative filtering is used to find users with similar tastes. To recommend a piece of content or service to a specific user, other users that have rated pieces of content similarly to the specific user can be found and the content or services that those users rated highly can be recommended to the specified user. Item-item collaborative filtering uses a similar reasoning but is used to determine content or service similarity which has less use factor in chatbots.

CHAPTER 3

Applications of AI & ML Techniques to Chatbots

APPLICATIONS OF AI & ML TECHNIQUES TO CHATBOTS

Previous sections have laid out the role chatbots can generally play in FP and SRHR and specified the context and purpose of Girl Effect's chatbots. Building on the context of the current state of available AI and ML techniques provided in Section 2, this section focuses on how these AI and ML techniques can specifically be applied to enhance FP and SRHR chatbots. These applications have been split up into subsections based on what chatbot purpose the AI or ML technique contributes to.

These techniques have been shortlisted by first surveying available AI and ML techniques and mapping them to use cases in different phases of Girl Effect's chatbot lifecycle. The AI and ML techniques that resulted from this mapping were then roughly evaluated across a variety of criteria like utility, how useful this technique will be for Girl Effect's desired impact; technological feasibility, whether tools at scale exist to deploy this technique; and the required financial and staff resources needed for development. This section lays out the most promising techniques that came out of this process and their use cases. It should be noted that the selected techniques are those that Girl Effect found most useful. Other organizations may find other techniques better suited to their purposes.

Ch 3.1 Knowledge Transfer: Chat Completions API + Text Embeddings API

One of the primary purposes Girl Effect's chatbots serve is providing accurate and up-to-date FP and SRHR information in a relatable way. It is paramount that Girl Effect's chatbots provide only vetted FP and SRHR information as there are myriad sources online that misguide our users. Girl Effect's current chatbot content is written and vetted by gender and SRHR experts, as well as content writers and creators who are aware of the cultural context that users live in. As such, relying on only the large corpus of text data fed to a large language model to generate culturally specific FP and SRHR answers to users' questions is not a sufficiently robust method. Moreover, large language models are known to "hallucinate," i.e. produce responses with incorrect facts when used in an uncontrolled manner.

To ensure that only vetted content is referenced to generate text or answers, OpenAI's chat completions API can be combined with its text embeddings API. Embeddings can be used to match user input to the most relevant existing and vetted Girl Effect content. The chat completions API will then be prompted to respond to the user input by referencing only the vetted Girl Effect content and no other external content, instructed to refuse to answer if the responses do not exist in the vetted Girl Effect content provided in the "user" message. This technique is typically called Retrieval Augmented Generation (RAG). In testing, this method in English and Hinglish has already proven useful in providing nuanced answers to freeform user input sourced only from Girl Effect content.

Using OpenAI's chat completions API combined with the text embeddings API has multiple advantages over current chatbot infrastructures. Girl Effect currently uses RapidPro to manage content and dialog flows in chatbots as connected message sets or flows. When information must be updated, the process of updating this content within our dialog flows is complex and labor-intensive because interconnected flows must be updated manually through RapidPro's user interface. An embeddings-based chatbot on the other hand references content stored as strings in CSV files or vector databases in more complex situations. OpenAI's Assistants API can also be used for this purpose without the need to set up extra infrastructure. This content is much simpler to edit and requires little technical knowledge to update.

Furthermore, the chat completions API can parse any type of user input. Currently, Girl Effect uses menu button options augmented with a BERT-based classifier called QnA fine-tuned on a question answering dataset to parse user input.

Because QnA is specific to a question answering dataset, when a user replies with any text that falls outside of this scope, the chatbot is unable to parse the input text and must simply reply with a canned response stating it cannot understand the user. With the chat completions API, GPT can be instructed to tell the user that it does not have the information to reply to the user if the user's text falls out of the scope of Girl Effect content, but it can still provide a response that is contextualized to the user's free-form input.

Because of the simplicity of OpenAI's APIs, the infrastructure for an embeddings-based chatbot is very simple and cheap to set up. Production level chatbots using this method can be set up in the span of a few months if the curated content is available. The GPT-3.5-turbo chat completions API is priced at $0.0015 / 1K tokens for input and $0.002 / 1K tokens for output and the Ada v2 embeddings API is priced at $0.0001 / 1K tokens. (1000 tokens is about 750 words). Further testing is necessary to determine the optimal price/performance trade-off.

Ch 3.2 User Experience Design

While knowledge transfer is a very important component of Girl Effect's chatbots, the user experience of the chatbot deeply influences whether a user stays on the chatbot long enough to embark on a positive behavior change journey. The user experience of a chatbot is determined by many different factors; the chatbot's ability to carry on natural and flowing conversations, the layout of the user interface, and the chatbot's ability to provide personalized responses are all examples of these factors.

Girl Effect's chatbot user experience is designed to reflect the local context and culture of the user. Content is written in a youthful tone with emojis and local slang integrated. While GPT-3.5 and GPT-4 generate very human-like text, Girl Effect has imbued its current chatbots with a specific personality and tone as well, one which is designed to be relatable and trustworthy to our users to keep them engaged. On the other hand, Girl Effect's current chatbots cannot carry out flowing conversations; GPT-3.5 and GPT-4 can.

The following sections describe AI and ML techniques that can be implemented to combine GPT's forte of generating natural and human-like conversation with Girl Effect's forte of tone and personality. They are grouped by short term, intermediate and long term according to the simplicity of implementation and timelines required to implement. AI and ML techniques in the short term subsection can technically be implemented at production level in the next three months to one year. AI and ML techniques in the intermediate subsection can be technically implemented at production level in one to three years. AI and ML techniques in the long term subsection can be technically implemented at production level in three to five years.

Ch 3.2.1 Short Term: GPT-4 Steerability

While OpenAI's GPT-3.5 has a personality with a fixed verbosity, tone, and style, with the release of GPT-4, OpenAI has also added functionality to prescribe GPT-4's style and task by describing those directions in the "system" message, a new input parameter to the API's chat completion function. System messages allow API users to significantly customize their users' experience within bounds [14]. In testing, Dimagi has shown that prescribing a task list, a detailed personality, guidelines on message length and tone, and incorporating several reminders of the chatbot's boundaries in the "system" message produces a remarkably high-performing chatbot that is able to stay committed to its assigned task and adjust its language usage to its prescribed personality in most circumstances. Dimagi experiments with this type of prompt engineering have produced chatbots that are very difficult to misdirect with multiple guardrails set up to anticipate and handle anomalous user input.

This level of advanced steerability is only possible using GPT-4. As previously mentioned, GPT-4 is 20 times as expensive as GPT-3.5-turbo, costing $0.03 / 1K tokens on input and $0.06 / 1K tokens on output. It should be noted that adding significant numbers of words in the system message to steer the generated text increases the cost of every input (1000 tokens is about 750 words).

Ch 3.2.2 Mid Term: Fine-Tuning API, Speech to Text API, and User Preference-Driven UX Design

3.2.2.1 Fine-Tuning:

While GPT-4's steerability produces astonishingly human interactions, the cost of using its API builds up quickly as chatbot users and input and output messages grow in number. In scaling up a chatbot, these costs may become unsustainable unless the price of GPT-4 drops considerably in the near future. It is important to explore other routes of generating text with specified tone and personality for a range of cost options.

OpenAI's fine-tuning API can be used to further train GPT's models to generate text for specific tasks or with certain qualities. Girl Effect can use this API to train GPT's models on Girl Effect's existing youth-friendly content sets. Fine-tuning a GPT-3 model may prove more cost-effective and have similar tone results to GPT-4 in the long run.

To implement fine-tuning, a tone dataset must be generated. In Girl Effect's experience, this process can be labor-, time-, and cost-intensive, although it can now be supported by GPT-4. This step is often the largest barrier to effectively and successfully fine-tuning a model to generate text with a certain tone and personality. Once this dataset is curated, it is fed into OpenAI's fine-tuning API. OpenAI's models available for fine-tuning have different prices for training and usage: Ada is $0.0004 / 1K tokens for training and $0.0016 / 1K tokens for usage; Babbage is $0.0006 / 1K tokens for training and $0.0024 / 1K tokens for usage; Curie is $0.0030 / 1K tokens for training and $0.0120 / 1K tokens for usage; and Davinci is $0.0300 / 1K tokens for training and $0.1200 / 1K tokens for usage (1K tokens is about 750 words).

Other vendor large language models like Claude, Cohere, and Scale could also be considered.

For this option, Girl Effect or other interested organizations would first have to set up the infrastructure to host open source large language models like Meta's LLaMa or others offered on Hugging Face [36]. Hosting and managing open-source models adds complexity compared to simply using proprietary APIs, indicating extra costs in infrastructure and technical expertise which may outweigh the aforementioned cost and time benefits of PEFT techniques. It is also not guaranteed that using PEFT techniques with smaller open source language models will perform at the level of other large language models to produce Girl Effect's desired tone and personality. These options are still being explored by the open source community.

It should be noted that the OpenAI API only supports fine-tuning via supervised learning. The developers of ChatGPT found that fine-tuning via supervised learning produced less than optimal results and this motivated the creation of a more complicated technique known as RLHF, which is discussed below in Section 3.2.3.1. Organizations will need to perform testing to determine if supervised fine-tuning produces adequate tone. If not, organizations will need to invest in more complex and expensive techniques such as RLHF or DPO (Section 3.2.3.2) to produce the desired tone.

3.2.2.2 Speech to Text API:

Through formative research and focus group testing, Girl Effect has found that many chatbot users prefer communicating on messaging platforms using voice notes.

While Girl Effect can currently craft voice notes to send to users, a feature deployed in our Hinglish chatbot Bol Behen, it has not been possible to understand and respond to user voice notes so our current chatbot platforms do not accept voice notes as input.

In September 2022, OpenAI released and open-sourced its neural net called Whisper which approaches human level robustness and accuracy in English speech recognition [37]. Whisper is not just limited to English, but also enables transcription in multiple languages as well as translation from those languages into English. Previously difficult to run, since March 2023, Whisper has been made available on OpenAI's API as the speech to text API [10]. Whisper API availability opens up a large opportunity space to implement speech recognition combined with the chat completions API to generate responses to speech inputs simply and easily. Other widely available models like Google's Text-to-Speech system can even be implemented atop the chat completions API to send users voice notes in response.

OpenAI's Whisper model is currently priced at $0.006 / minute of recorded speech. As an example of one in many text-to-speech models, Google's Text-To-Speech offers free speech synthesis of up to 4 million characters per month. After the free usage limit is reached, it is priced at $4 per 1 million characters.

3.2.2.2 User Preference-Driven UX Design:

Currently, due to the technical constraints, Girl Effect's chatbots have one user experience. This user experience has been crafted to appeal to the average chatbot user, grounded in focus group testing with feedback from girls but there is a large variety of users who have different preferences in how they speak to a chatbot and how they expect chatbots to respond. These preferences can be of many forms: one user might prefer higher frequency and shorter length texts while another user prefers communication primarily through voice notes; one user might prefer a friendly, casual tone when discussing sensitive SRHR topics whereas another might be put off by a carefree tone in relation to those topics. Hosting different versions of a user experience would be prohibitively complex within our current chatbot infrastructure.

The combination of OpenAI's chatbot completion API and embeddings API described in Section 3.1 offers new opportunities to explore different versions of user experiences because it does away with the complex structuring of message sets and dialog flow that must be manually updated if changes in tone, personality, text length, and other user experience components are required. Using the chatbot completion API instead allows Girl Effect to define many user experience parameters like length of text per message or tone in the "system" message. A library of different "system" messages can be created to represent different types of user experiences including a "system" message that is designed to intake and output text from voice notes.

A user experience design that responds to the preferences of the user may be set up in the following way. The chatbot and data platform must first be set up to record metrics related to engagement, like average chatbot session duration, the number of times a user initiates a conversation with the chatbot, the number of messages exchanged, sentiment of the users' inputs, or whether the user took an action like accessing a service as a few examples. Once a library of user experiences is created, one user experience is randomly selected for each new user. As users interact with the chatbot, Girl Effect's data infrastructure collects the aforementioned engagement metrics per user and labels these interactions as associated with the randomly selected user experience. As the user continues engaging with the chatbot, the chatbot randomly switches to another user experience in the library and our infrastructure records engagement metrics associated with this different user experience. As user engagement data labeled by the user experience grows, Girl Effect will begin to be able to ascertain which user experience the user engages with most positively and prioritize that user experience for that user.

Ch 3.2.3 Long Term: Reinforcement Learning from Human Feedback, Direct Preference Optimization and Cluster Analysis

Sections 3.2.3.1 and 3.2.3.2 focus on techniques that could create models that generate nuanced, culturally relevant, youthful text and personality with potentially lower cost. However, these techniques are not currently available on OpenAI's or other competitors' APIs and would require the use of open-source models. In the long term, many open source models may rival or even outperform proprietary models like GPT, Claude, and Cohere so it is still important to consider these techniques even if they require open source infrastructure. Sections 3.2.3.3 and 3.2.3.4 focus on ML techniques aside from large language models that could enhance Girl Effect's chatbots' abilities to create well-crafted and personalized experiences for our users.

3.2.3.1 Reinforcement Learning from Human Feedback (RLHF):

Reinforcement Learning from Human Feedback (RLHF) is a technique most notably used by OpenAI to train both its original ChatGPT and GPT-4 models and is credited for both models' ability to generate nuanced, human-like text and conversation.

RLHF involves combining aspects of supervised learning (Section 2.1) and another machine learning technique known as Reinforcement Learning with Human Feedback (RLHF). In RLHF, human feedback is used to guide the learning process.

To effectively implement this technique, an open-source model that performs well and supports the multiple languages Girl Effect works in must first be identified. Then a training dataset would need to be created that contains hundreds or thousands of training examples. Each training example would consist of a typical user question or input and two or more ways of responding. The responses should have the same content but be different in tone. This dataset creation process could be supported by using the open-source language model or GPT to generate responses, supplemented by human-generated responses. Human raters would then have to rank the responses to each user input. This data would then be used to train a reward function and this reward function would be used to fine-tune the model to achieve the desired tone and personality.

This technique requires large amounts of technical and financial capacity and is difficult to implement which is why it should only be considered in the longer term if other options are not sufficient.

3.2.3.2 Direct Preference Optimization (DPO):

Direct Preference Optimization (DPO) is a cutting edge technique that is a promising alternative to conventional preference learning methods like RLHF [38]. It uses human preference data to optimize AI policies using supervised learning, eliminating the need for fitting a reward model and for using reinforcement learning to train the policy, both of which are required steps in RLHF. Experiments show that DPO can fine-tune large language models to align with human preferences as well as or better than existing methods and even exceeds RLHF's abilities on a variety of tasks while being simpler to implement and train.

DPO is a very new technique and it is yet to be determined how difficult it may be to implement in practice even if it is posited that it is simpler than RLHF. If it can be implemented, it may prove to be a more powerful and effective technique than RLHF for fine-tuning tone.

3.2.3.3 Cluster Analysis:

Another method Girl Effect may use to create personalized experiences for users is cluster analysis. Cluster analysis techniques can be used to group similar users together. As a first step, these clusters can then be used to make personalized chatbot content recommendations based on what content other users in the group engaged with.

Building on this as Girl Effect begins to understand the user characteristics of different clusters of users, personalized content journeys based on social and behavior change theory can be crafted for larger clusters that contain the most typical users. As a user engages more with the chatbot and Girl Effect is able to place the user in a specific known cluster, the chatbot can more quickly and proactively guide the user on a positive social behavior change journey towards better sexual and reproductive health outcomes.

Cluster analysis requires structured and coherent datasets to draw insights. The process of setting up a data infrastructure that can support these techniques is time-, labor-, and cost-intensive. Once this infrastructure is set up, cluster analysis techniques have to be trained and tuned which can take a wide range of time spans depending on the complexity of the datasets.

Ch 3.3 Low-Resource and Mix-Coded Languages

Most current large language models are optimized for the English language although they have been fed on a corpus of text that includes several other languages. Girl Effect and many other international development organizations work in contexts where it is essential to deliver content in multiple languages across the African continent and India. Most languages spoken in these regions are low-resource languages (languages that lack large monolingual or manually crafted linguistic resources sufficient for building statistical NLP applications). Not only are users in these regions speaking in low-resource languages, they also often mix languages and local slang, speaking in what is termed mix-coded languages. Some examples of mix-coded languages are Spanglish (Spanish and English), Hinglish (Hindi and English), and Sheng (Swahili and English). Ideally, Girl Effect's chatbots would not only understand user text in these languages but also generate mix-coded language text that feels natural and relatable to the user.

The following subsections explore AI and ML techniques that can be implemented to enhance chatbots' multi-language capabilities. They are grouped by short-term, intermediate, and long-term techniques according to the simplicity and timeline for implementation. AI and ML techniques in the short-term subsection can technically be implemented at production level in the next three months to one year. AI and ML techniques in the mid-term subsection can be technically implemented at production level in one to three years. AI and ML techniques in the long-term subsection can be technically implemented at production level in three to five years. To remain up to date with continuously evolving mix-coded languages, mid-term and long-term techniques would also have to eventually employ continuous learning, updating their training as new user text data comes in with the most up-to-date usage of the language.

Ch 3.3.1 Short Term: GPT-4 Prompt Engineering and Embeddings API

For the task of understanding multi-language and mix-coded Languages, in initial testing, both GPT-3.5 and GPT-4 were successful in understanding Swahili, Hindi and even Sheng and Hinglish. This may be due to the fact that Hindi is not a low resource language, as one of the most widely spoken languages in the world, and despite Swahili being a low resource language, because of its wide use across multiple African countries, it is one of the most well-resourced African languages for the purposes of NLP.

For the task of generating multi-language and mix-coded languages, in testing, there were marked differences between GPT-3.5 and GPT-4. As mentioned in previous sections, one of the largest advantages GPT-4 has over GPT-3.5 is its multi-language capabilities. When prompted to generate just Swahili or just Hindi, GPT-3.5 performed reasonably well, but when asked to generate Sheng or Hinglish, GPT-3.5 was inconsistent, often defaulting to pure Swahili or Hindi.

On the other hand, GPT-4 has shown much more accuracy and nuance in both Swahili, Hindi, and even Sheng and Hinglish simply by including the instruction to speak in the selected language with some further instruction on personality and tone in the "system" message.

Another method that has proven useful in generating mix-coded language is using OpenAI's embeddings API. This method, as described in Section 3.1, forces the chat completions API to reference only Girl Effect vetted content when responding to a user's input. In testing, when using content sets written in a selected mix-coded language, the chat completions API using GPT-3.5 generated convincing Hinglish and even used emojis in similar ways to the provided content. Girl Effect has yet to test this method in Sheng, but if successful, using the embeddings API and Girl Effect vetted content may be a cost-effective method of getting potentially similar performance to GPT-4 in generating mix-coded language text.

Ch 3.3.2 Mid Term: Fine-Tuning API

If the methods described in Section 3.3.1 are insufficient, then GPT's fine-tuning API can be used in a similar way to the way it was used in Section 3.2.2.1 except applied to mix-coded languages. Girl Effect can use this API to train GPT's models on Girl Effect's existing mix-coded language content sets.

To implement fine-tuning, a mix-coded language dataset must be generated. In Girl Effect's experience, this process can be labor-, time-, and cost-intensive, although it can now be supported by GPT-4. This step is often the largest barrier to effectively and successfully fine-tuning a model to generate nuanced mix-coded language text. Once this dataset is curated, it is fed into OpenAI's fine-tuning API. OpenAI's models available for fine-tuning have different prices for training and usage: Ada is $0.0004 / 1K tokens for training and $0.0016 / 1K tokens for usage; Babbage is $0.0006 / 1K tokens for training and $0.0024 / 1K tokens for usage; Curie is $0.0030 / 1K tokens for training and $0.0120 / 1K tokens for usage; and Davinci is $0.0300 / 1K tokens for training and $0.1200 / 1K tokens for usage (1K tokens is about 750 words).

Other vendor large language models like Cohere and Scale could also be considered.

Ch 3.3.3 Long Term: Fine-Tuning or Building Open Source Models

Possible longer term methods include continuing to train an existing open source model for our desired mix-coded languages or building our own mix-coded large language model using open source tools.

The advantage of both of these possibilities is that open-source models and tools are already readily available to implement free of cost. The disadvantage lies in the uncertainty around how large of a model would be needed in both cases to generate text with the nuance and complexity we desire for our chatbots. Computation costs could be high without an understanding of how large these models need to be. Furthermore, organizations would likely have to hire or contract a highly specialized AI and ML team to train and develop either of these options.

Despite these barriers, it may be a worthwhile pursuit to generate open-source models that are capable of generating mix-coded languages that millions of people use but are not well-resourced anywhere except in vendor large language models. Releasing these open-source models could indirectly enable exponentiating benefits for our target audiences by providing novel open-source large language models to smaller organizations in the region who do not have the financial resources to utilize vendor large language models but have the technical expertise to leverage open source models and tools.

Ch 3.4 Safety by Design: Fine-Tuning

While Girl Effect's chatbots are designed to take users on a positive behavior change journey towards better FP and SRHR outcomes, they must also be equipped to handle instances of sensitive disclosures and follow safeguarding protocols in such situations. Previously, Girl Effect has invested in developing a BERT-based safeguarding classifier called SaferChatbots to more accurately detect sensitive disclosures. Girl Effect's method of detecting these involves checking every user's freeform input text for a large variety of trigger words which includes all types of iterations of trigger words to account for misspellings and minute language idiosyncrasies. SaferChatbots was initially proposed to reduce human burden because the trigger word method flagged many inputs that were not sensitive disclosures but still required human review. It was hoped that SaferChatbots would result in fewer false positive sensitive disclosures reducing the amount of human review time. Unfortunately, due to a limited dataset and BERT's low familiarity with sensitive disclosure text data, SaferChatbots only passes a high enough accuracy threshold on a limited number of labeled categories out of the 20 categories it has been trained on, rendering it too inaccurate to be safe to implement. It remains in testing.

With the availability of GPT-3.5 it's possible that with OpenAI's fine-tuning API, the much more powerful GPT-3.5 model can instead be fine-tuned to detect sensitive disclosures. If successful and accurate, this fine-tuned model could be used as both the first classifier that user input and chatbot response will be run through to detect sensitive disclosures and also used to identify safeguarding risks and ensure all of GPT's responses meet Girl Effect's safety standards. In testing, Girl Effect would still employ its current back-up system of flagging any high risk trigger words to ensure no disclosures slip past the fine-tuned model.

To implement fine-tuning, a labeled safeguarding dataset must be generated. In Girl Effect's experience, this process can be labor-, time-, and cost-intensive although this step can now be supported by GPT-3.5 and GPT-4. This step is often the largest barrier to effectively and successfully fine-tuning a model to perform a specific task. Once this dataset is curated, it is fed into OpenAI's fine-tuning API. OpenAI's models available for fine-tuning have different prices for training and usage: Ada is $0.0004 / 1K tokens for training and $0.0016 / 1K tokens for usage; Babbage is $0.0006 / 1K tokens for training and $0.0024 / 1K tokens for usage; Curie is $0.0030 / 1K tokens for training and $0.0120 / 1K tokens for usage; and Davinci is $0.0300 / 1K tokens for training and $0.1200 / 1K tokens for usage (1K tokens is about 750 words).

Other vendor large language models like Cohere and Scale could also be considered.

There are three approaches to safeguarding, all of which use PEFT techniques. The first is to use two different chatbots, one fine-tuned for tone, verbosity, and style as discussed in Section 3.2, and another chatbot fine-tuned to detect unsafe inputs and responses.

The second approach would be to use a single chatbot that is fine-tuned to perform tone, verbosity, style, AND safety. To create a single chatbot, the safeguarding dataset would need to be integrated with the dataset for tone, verbosity, and style. The integrated dataset would then be used to fine-tune the system on tone, verbosity, style, and safety at the same time. This is preferable to fine-tuning first on tone, verbosity, and style and then separately on safety because the safety fine-tuning may cause the system to "forget" the tone, verbosity, and style training.

The third approach is to begin by fine-tuning a language model for tone, verbosity, and style, creating one layer of fine-tuned parameters, and then fine-tuning the same language model for safety, adding a new layer of fine-tuned parameters. These two processes could be switched as well. This approach works because PEFT techniques are stackable improvements in large language models.

As discussed in Section 3.2, the OpenAI API only supports fine-tuning via supervised learning. This may or may not produce adequate tone, verbosity, style, and safety. If not, we will need to invest in more complex and expensive techniques such as RLHF or DPO (Section 3.2.3.2) to produce the desired tone, verbosity, style, and safety.

Ch 3.5 Service Linkage

After encouraging our users on a positive social behavior change journey towards taking charge of their SRHR health, Girl Effect uses its chatbots to help users connect to the next step in their journey of accessing life-saving SRHR services when they are ready. At the moment, Girl Effect's chatbots offer this information when it is directly requested by the user. Information about different Girl Effect-vetted services is available on our country branded websites. Within Girl Effect's current chatbots, APIs are used to collect this information from Girl Effect's websites (which are all pre-vetted and posted by Girl Effect's team) and provide it to chatbot users based on the location they have selected. Currently, users must go through a tree of multiple choice questions to select their location for the API in our chatbot to then filter through available services. Due to the complexity of parsing a user's free input location, Girl Effect employs a rule-based system but updating this system with new locations and associated vetted services is labor-intensive.

The following subsections explore AI and ML techniques that can be implemented to enhance chatbots' abilities to direct users to services. They are grouped by short-term, and mid- to long-term techniques according to the simplicity and timelines for implementation. AI and ML techniques in the short-term subsection can technically be implemented at production level in the next three months to one year. AI and ML techniques in the mid- to long-term subsection can be technically implemented at production level in one-and-a-half to five years.

Ch 3.5.1 Short Term: Chat Completions API + Text Embeddings API

In a similar fashion to knowledge transfer, OpenAI's chat completions API can be combined with its text embeddings API to match the user's input location with the most relevant set of vetted Girl Effect services stored as strings and embeddings in a CSV file or vector database. This input location could come from direct user input or determined using location technology to automatically detect the girl's location. The chat completions API will then be prompted to respond to the user's location by referencing only the vetted Girl Effect services and no other external content, instructed to answer with no services and general advice if the location does not exist in the Girl Effect's vetted library of services. This method would remove the labor intensive step of continually updating our tree of possible locations within our current chatbot infrastructure.

Ch 3.5.2 Mid to Long Term: Supervised Learning, Cluster Analysis and Recommender Systems

While currently Girl Effect's chatbots only direct users to a service when they access message sets related to services themselves, Girl Effect hopes to eventually gain a deep enough understanding of its users that the chatbots can then offer personalized service recommendations based on how users interact with the chatbot.

The simplest method by which Girl Effect hopes to eventually more actively direct users to services is using supervised learning techniques. Supervised learning techniques, like linear regression or logistic regression, can be used to find common characteristics amongst users who access services and then predict other users' likeliness to access services based on these common characteristics.

Another method Girl Effect may use is cluster analysis. Cluster analysis techniques can be used to group similar users together and make service recommendations based on what services other users in the group were interested in or accessed.

Both supervised learning and cluster analysis require structured and coherent datasets to draw insights. The process of setting up a data infrastructure that can support these techniques is time-, labor-, and cost-intensive. Once this infrastructure is set up, both supervised learning techniques and cluster analysis techniques have to be trained and tuned which can take a wide range of time spans depending on the complexity of the datasets.

Recommender systems are typically good choices for recommending content or services but Girl Effect does not currently offer information about a large enough range of services to justify the development of recommender systems that are often quite complex and expensive to set up. Other organizations whose purposes are more focused on recommending a large range of products and services might find these systems useful.

Ch 3.6 Analysis

As briefly mentioned in Figure 3, significant phases of Girl Effect's design lifecycle include formative research, monitoring, and evaluation. Many activities in these phases generate freeform text data from open-ended responses, interviews, focus groups, and also text users send to chatbots. These datasets are often large, unstructured, and take many labor hours to analyze. In these cases, text analysis methods using NLP can be employed to draw quicker and deeper insights from this data. Techniques like sentiment analysis, topic modeling, and named entity recognition can help identify key themes, sentiments, and entities within our data. These insights can then be used to better inform new iterations of the chatbots.

Using GPT models through OpenAI's chat completions API and prompting the model to determine sentiment or identify key themes can rapidly provide a good first-pass, rough analysis but this may not be as accurate as a fine-tuned model designed for sentiment analysis or topic modeling. BERT and BERT derivatives are better suited for sentiment analysis and models like latent Dirichlet allocation (LDA) or BERTopic are better suited to topic modeling or thematic analysis but most of these models will not provide perfect results "out of the box" – they often must be carefully fine-tuned on task-specific data, increasing the time and complexity of using these methods for more nuanced text analysis [39].

The choice of technique and model should be determined by the purpose of the analysis. In Girl Effect's monitoring phase, content, UX and technology teams are often looking for a near real-time understanding of what users are engaging with or finding issues with. For this purpose, it is likely more appropriate to use unmodified GPT models to provide a rough analysis of how users are engaging with the chatbots. In the formative research and evaluation phases where Girl Effect is trying to draw nuanced insights out of interviews and focus groups to inform the design and evaluation of the chatbot, fine-tuned BERT or LDA may be more appropriate.

CHAPTER 4

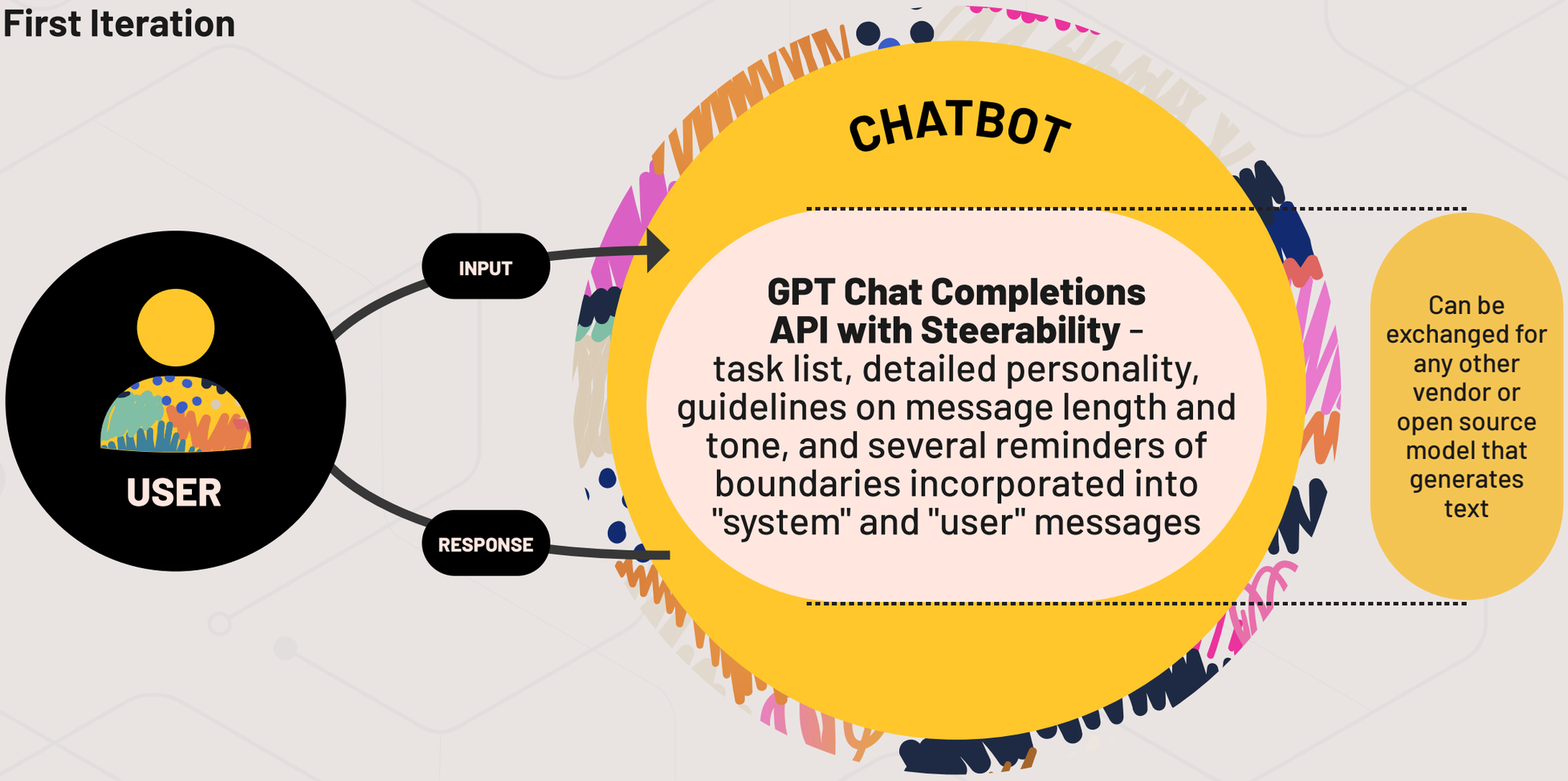
Possible Iterations of AI & ML Infrastructures for Chatbots

POSSIBLE ITERATIONS OF AI & ML INFRASTRUCTURES FOR CHATBOTS

This section suggests iterations of an AI and ML-augmented infrastructure for a chatbot. These iterations build on each other by adding different AI and ML techniques described in Section 3 to each previous iteration to incrementally develop a safer, robust, and personalized chatbot.

The FIRST ITERATION is the simplest infrastructure needed to stand up a chatbot using a large language model available on an API that generates text.

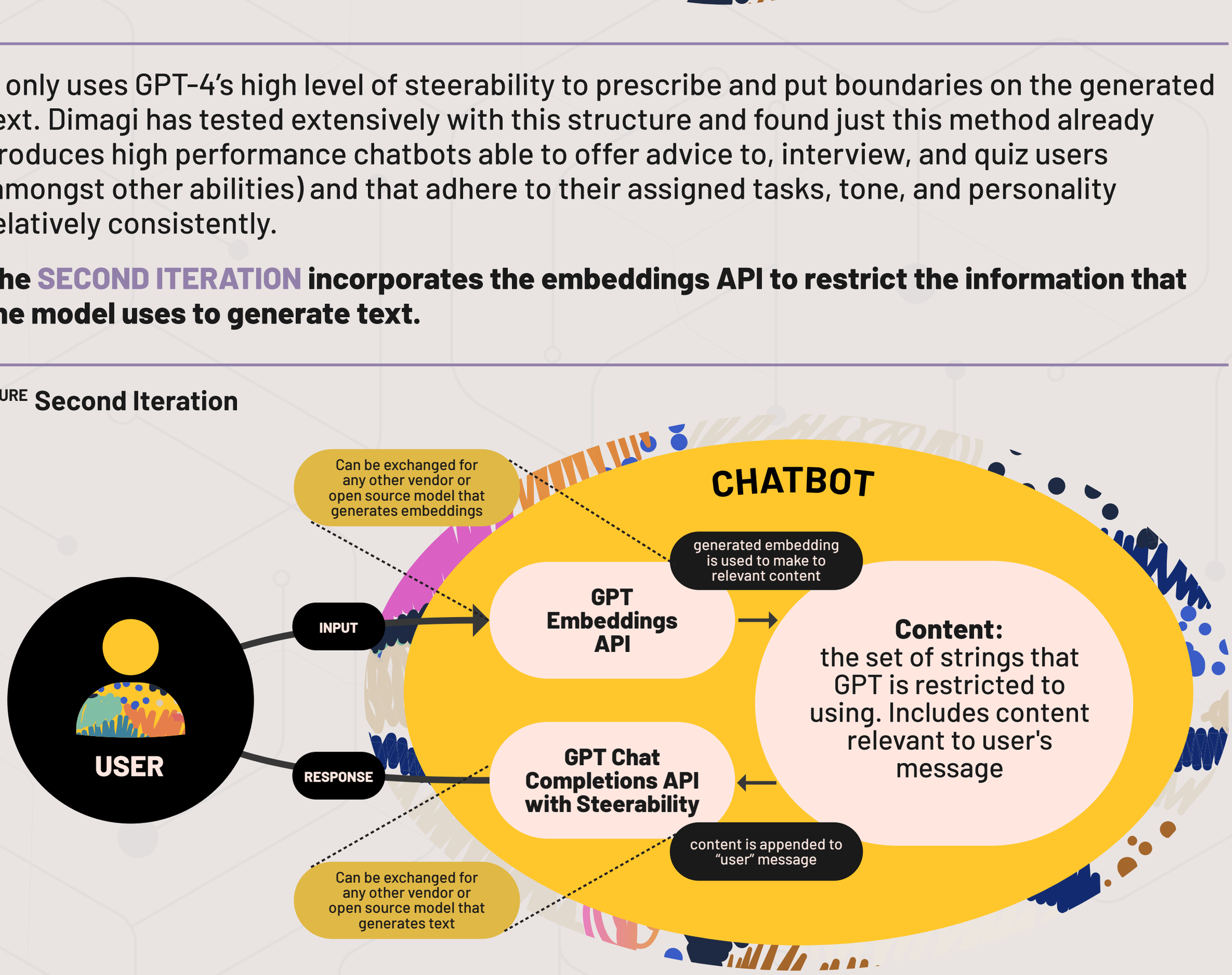
FIGURE **First Iteration**



It only uses GPT-4's high level of steerability to prescribe and put boundaries on the generated text. Dimagi has tested extensively with this structure and found just this method already produces high performance chatbots able to offer advice to, interview, and quiz users (amongst other abilities) and that adhere to their assigned tasks, tone, and personality relatively consistently.

The SECOND ITERATION incorporates the embeddings API to restrict the information that the model uses to generate text.

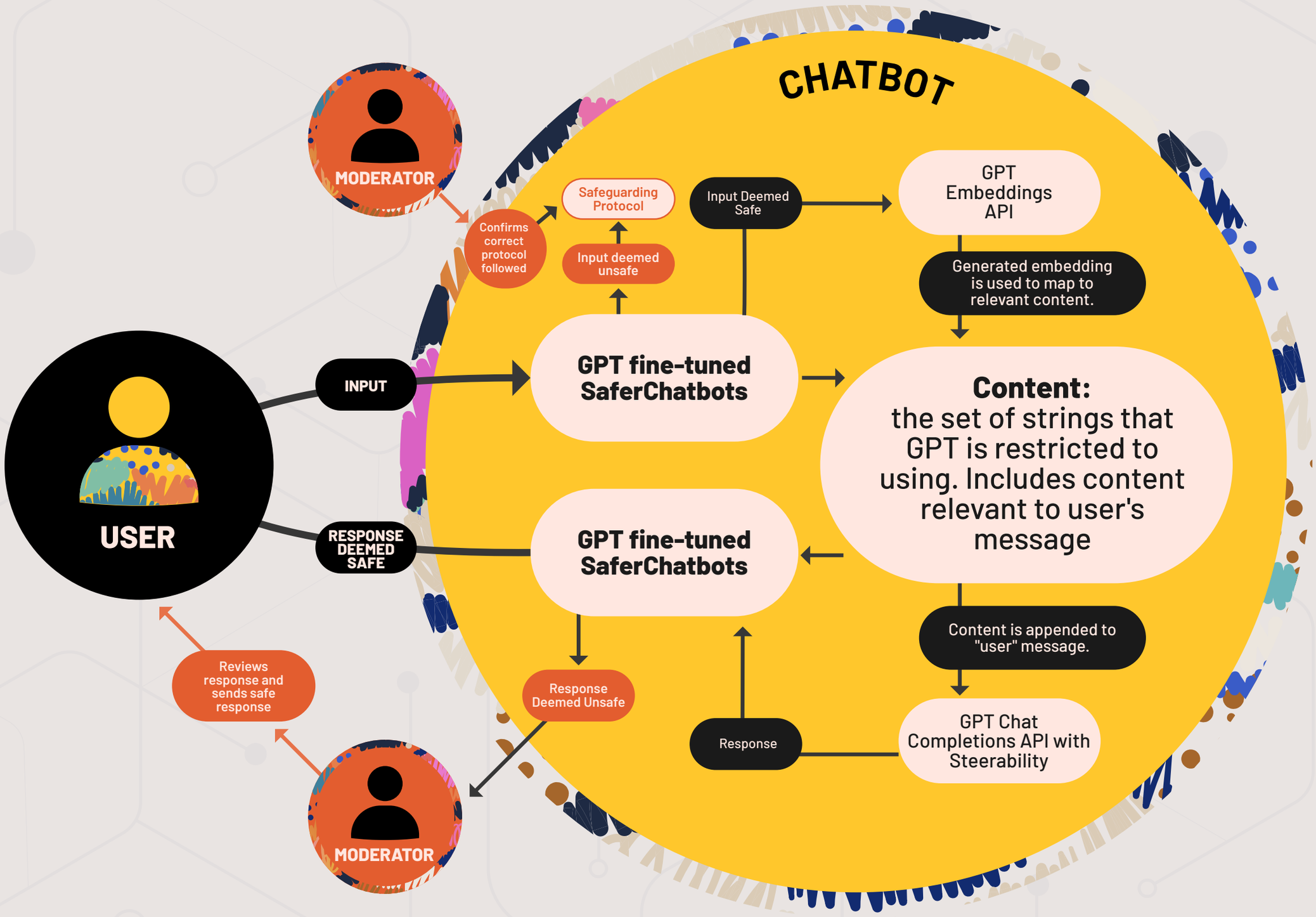
FIGURE **Second Iteration**



It uses GPT-3.5 or GPT-4's high level of steerability to prescribe tone and personality and the embeddings API to prescribe the content allowed in the generated text. Girl Effect has tested this structure and found it simple to use and remarkably powerful at providing answers and responses to even misspelled, grammatically incorrect, and low context user inputs in multiple languages.

Iterations from here on out have not been tested and are infrastructures Girl Effect hopes to explore in the future. **The THIRD ITERATION adds both a fine-tuned GPT model (or other vendor or open source model) for detecting sensitive disclosures and harmful content and a human-in-the-loop safeguarding process to ensure robust safety standards are followed.**

FIGURE **Third Iteration**



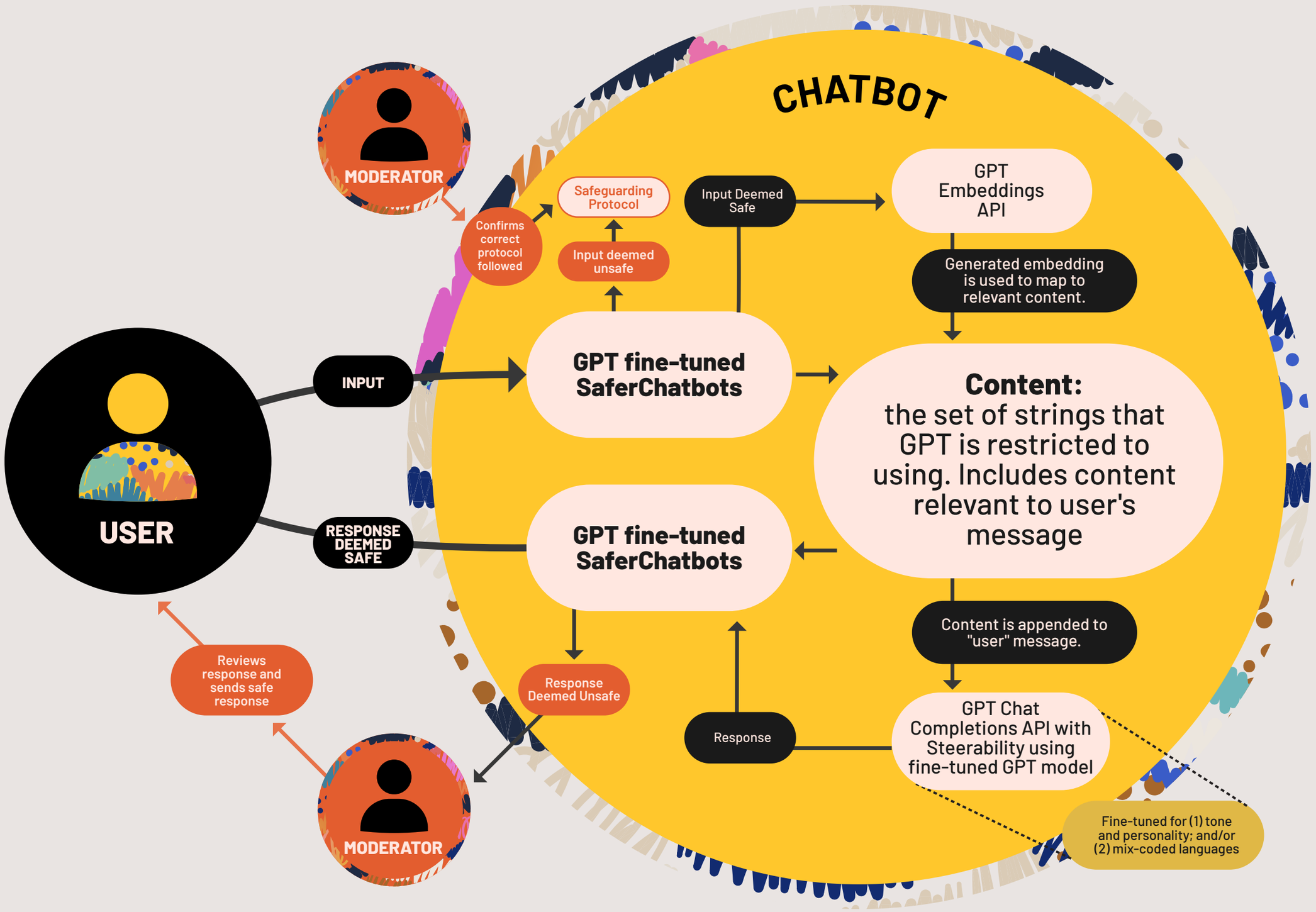
At this point and for future iterations, Girl Effect might consider using RapidPro or another chatbot content management system to manage the system of API calls to different embeddings CSV files/databases or fine-tuned models. This iteration of the chatbot currently only generates text from a vetted set of Girl Effect content and has human-in-the-loop safeguarding protocols integrated.

This iteration would bring Girl Effect close to the current state of its chatbots and would have a number of advantages over its current chatbots:

- **the chatbot will be able to understand any user input and provide a response;**
- **making content changes will be as simple as updating strings in a CSV or database;**
- **every response the chatbot sends will be newly generated, no response exactly repeated, resulting in a more natural and human-like user experience; and**
- **the load on the human moderator may be reduced if our fine-tuned GPT SaferChatbots is accurate enough.**

The FOURTH ITERATION adds model(s) fine-tuned for tone and mix-coded languages that generate text crafted in Girl Effect's preferred tone, personality, and language.

FIGURE **Fourth Iteration**

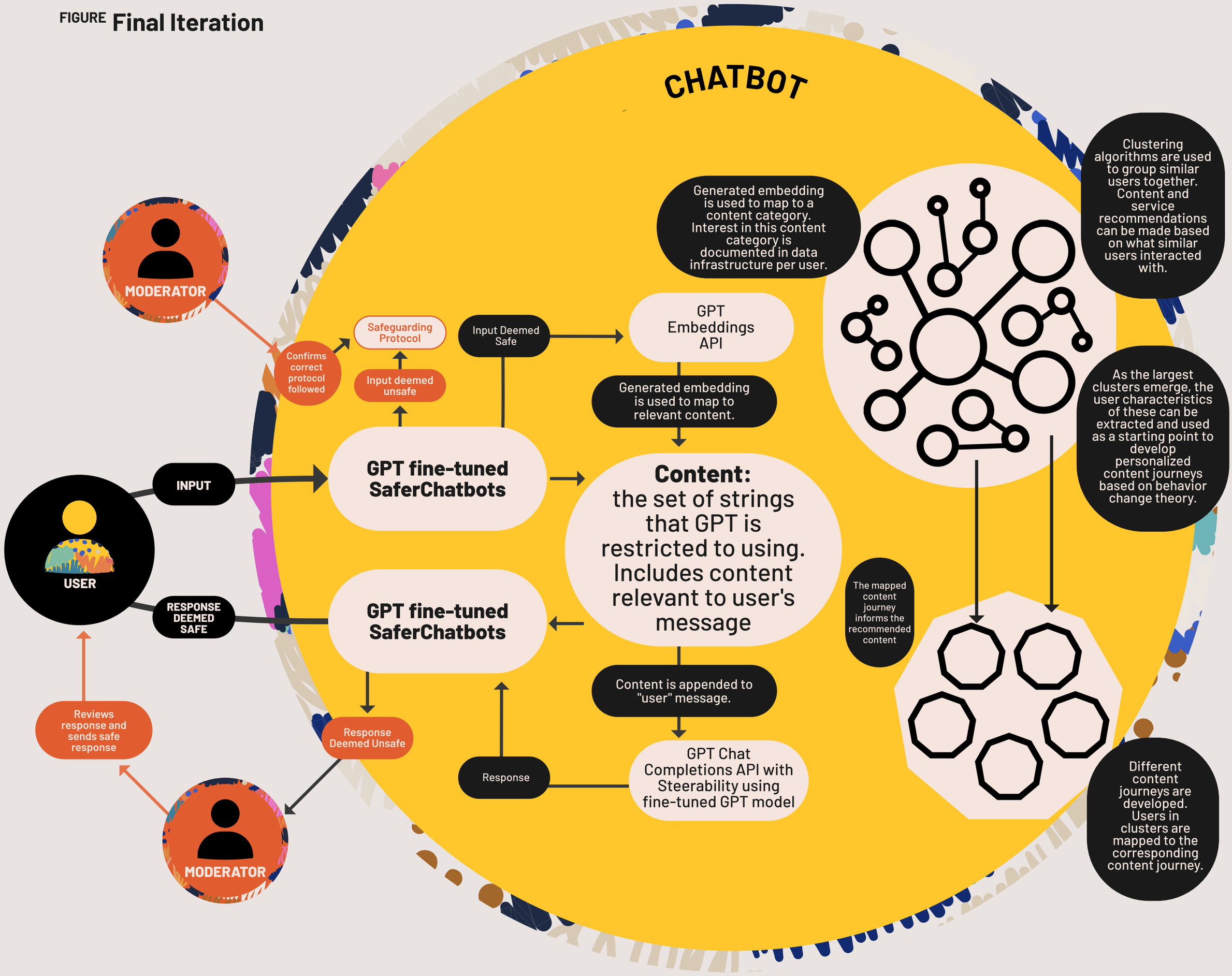


This iteration of the chatbot now only generates text from a vetted set of Girl Effect content, has human-in-the-loop safeguarding protocols integrated, and speaks in the tone and language that our users find relatable.

This iteration would match our current chatbots and would have all the advantages mentioned in the third iteration with the added advantage of all text generated in Girl Effect's preferred tone, personality, and language.

The FIFTH AND FINAL ITERATION (in this vision) adds clustering analysis techniques used to recommend content and services and further proactively guide users on a personalized content journey designed using behavior change theory. This addition relies on a robust data infrastructure behind the scenes that stores user data in a structured and coherent enough way to allow the use of many other types of machine learning techniques that may help us understand our users better.

FIGURE **Final Iteration**



This final iteration is Girl Effect's dream version of a chatbot that is not only able to give our users the information and services they need when they are ready but also track where our users are in their behavior change journey and use these insights to gently steer our users towards better SRHR outcomes.

Not included in this diagram is user preference-directed UX design but this application would also be made possible by the integration of a robust data infrastructure behind the scenes.

CHAPTER 5

Data & AI Ethics, Privacy and Protection: Girl Effect's Approach

DATA & AI ETHICS, PRIVACY AND PROTECTION: GIRL EFFECT'S APPROACH

Ch 5.1 Data Ethics

Ch 5.1.1 Girl Effect's Values and Principles

Along with discussion around the use of AI and ML in improving chatbots comes the obvious consideration for data ethics and privacy. Furthermore, when using generative AI services like GPT in particular, the data being fed into OpenAI and other companies' language models must be adequately protected. This may involve special data agreements with these companies. Every organization is different and requires nuanced approaches to responsibly using data. Here we have described Girl Effect's data values, principles, and standards.

We ensure the privacy and protection of all our data.	The data we hold is girls' assets, and we treat all data with the utmost respect and care. Privacy is our default setting and we ensure that our practices comply with both national and international data protection laws.
We collect a minimum amount of personal data.	We only collect the information necessary to meet legitimate business purposes and to deliver, provide, maintain or develop services. We do not keep personal information for longer than is necessary.
We are transparent and accountable about our privacy, security and safety practices.	When collecting personal data, Girl Effect explains why the information is being collected, how it will be used, by whom and for how long. This is set out in our Privacy Notice/s on our website. We document all processes related to privacy, security and safety and ensure that any breaches or violations are reported and addressed promptly and effectively.
We implement special safeguards to protect the welfare of children and young people.	We implement additional measures to identify and mitigate risks to children and young people who share their data as part of any of our initiatives. We consider the most vulnerable girl as our baseline for determining privacy, safety and security policies and practices.
We only use children and young people's personal information to meet legitimate business purposes.	We only access, collect, share, disclose and further use children and young people's personal data to meet legitimate business purposes or to meet our legal obligations. Partners are never allowed to access children and young people's personal information unless it is for a legitimate business purpose which supports Girl Effect in fulfilling its goals.
We put girls' privacy and protection above our own institutional benefit and any monetization of data will be returned to the project involved.	We ensure that children and young people, as well as other data subjects, are treated equally and are able to access their rights without discrimination. Every individual has the right to fair, lawful and secure processing of their data and Girl Effect takes all steps to ensure the safety and security of all data processing. These rights include the right of access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object and rights in relation to automated decision making and profiling. These protections are implemented regardless of the nationality or place of residence of the individual.
We provide children and young people with opportunities to exercise meaningful choice and control over their personal data.	We provide information to support children and young people's understanding of data privacy, security and safety and help them to exercise their rights in this regard. All information provided to children and young people is adapted to their levels of understanding and we will give easy, clear choices to whether to provide their personal data or subscribe to a particular service.

Ch 5.1.2 Standards

Standard 1. RESPONSIBLE DESIGN
Girl Effect integrates and documents measures to ensure the safe and ethical management of data into the design and development of all its existing and new research products, initiatives and research.

Standard 2. RESPONSIBLE DATA COLLECTION
Each Girl Effect product, initiative or research has a lawful basis for collecting data and active, informed consent is obtained for the collection and use of any personal or sensitive data.

Standard 3. RESPONSIBLE DATA STORAGE AND TRANSMISSION
Girl Effect staff and partners ensure that data, particularly sensitive or personally identifiable data, are protected and secure from unauthorized surveillance, capture, modification, downloading, copying, transfer, sharing, tampering, unlawful destruction, accidental loss, and/or improper disclosure during capture, storage and transmission.

Standard 4. RESPONSIBLE DATA USE, RE-USE AND DATA SHARING
Data are used and re-used, both internally and externally, to help Girl Effect affect change and enhance impact but any use, re-use or data sharing respects consent limitations and considers privacy and security risks.

Standard 5. RESPONSIBLE DATA RETENTION AND DESTRUCTION
Personal data are only retained for as long as it is necessary and are destroyed, aggregated or otherwise rendered anonymous once the specified purpose of the data has been fulfilled.

Ch 5.2 AI Ethics and Privacy

Like all efforts that use large language models (LLMs), Girl Effect's application of LLMs for chatbots to support adolescents to access sexual and reproductive health (SRH) information and learn about referral services and counseling presents potential ethical risks and harm.

Four categories of risk should be mitigated or managed, including risk to:

- **Adolescents who use a chatbot –** including possible physical, emotional, psychological, legal, economic, political, or reputational harms.
- **Groups or communities of people who use (or are excluded from use of) a chatbot –** including possible harms related to bias, exclusion, discrimination, privacy loss, and mis- or disinformation.
- **Organizations (including Girl Effect itself) who provide, promote, or receive referrals from a chatbot –** including loss of trust, financial and legal risk, and reputational risk if a chatbot causes harm or is perceived to be acting in an unacceptable or undesired way or providing information that is questioned/questionable.
- **Wider ecosystems –** including risks of mistrust in wider health or NGO systems, excessive use of natural resources (such as energy and water required for data processing), and harms along the labor supply chain (e.g., if those who train or label AI are suffering from unsafe/unfair labor practices).

While all programming, and any programming that involves technology and data involves risk, specific risks related to LLMs have been widely cited, including:

- Data used to build LLM applications may not be a true or appropriate representation of the context or the population.
- Data used to train LLMs is drawn from large data sets that are rife with harmful bias, including racial and gender bias.
- Data quality and quantity issues lead to lower accuracy, validity, and relevance, especially for non-dominant languages, including Sheng and "Hinglesh" in which Girl Effect is operating chatbots.
- Errors in outputs (so-called "hallucinations") due to insufficient language data and contextual awareness to train an LLM properly or a chatbot being programmed to provide answers even when it does not have sufficient information to correctly answer.Data sets used to train AI systems can become detached from their original and intended context over time, reducing applicability, producing errors and/or doubling down on biases.
- Privacy risks are intensified due to the large volumes of data used and enhanced data aggregation, which can allow for identification of individuals within a data set because more personal identifiers are linked.
- Local privacy risks, especially in contexts where phones are shared, adolescents (especially girls) have less power and control over devices, and users are less aware of how to manage their own privacy.
- Lack of transparency because machines are learning on their own and it is difficult to fully understand how decisions are made within opaque algorithms.
- People may perceive AI systems as being more objective than humans and as having greater intelligence or capacity than humans, even though all AI systems are built by humans and carry human biases.
- Assumptions that an LLM-powered application built in one context can be easily transferred and applied to a new or different context without retraining and sufficient testing.
- Blatantly insensitive answers to safeguarding issues, mental health questions, gender-based violence content, and other sensitive topics.
- Trialing of unproven approaches with vulnerable people and groups.
- Extractive processes and low levels of participation in design of LLMs and chatbots by those whose data is used to create them.
- Lack of accountability to the people and communities whose data enables and/or who are using LLMs and/or chatbots.
- Challenges with ensuring meaningful, active, and informed consent for collecting, using, and storing data from vulnerable people who have lower digital literacy and data privacy awareness, and related challenges with ensuring transparency with users that they are chatting with a "bot" and not a human.

One response to addressing emerging risks with AI, even before the mainstream explosion of LLMs on the scene in early 2023, has been the development of ethical principles and guidance. A 2019 report identifies no less than eighty-four sets of ethical principles. Across them, researchers found overlap in five key areas: transparency, justice and fairness, non-maleficence, responsibility, and privacy. Girl Effect's Responsible Data Principles and Standards, developed in 2018, converge with these five common principles (See Table 1).

Girl Effect is working now to understand exactly how to apply these principles and standards to emerging efforts to incorporate LLMs and Generative AI into chatbots, and to understand what additional guardrails might be needed to ensure safe design and use of chatbots.

In its work to design and roll out chatbots, Girl Effect will:

1. **Create and implement ethics, principles and guidelines:** Develop clear ethics guidelines and train staff, partners, contractors, and vendors to ensure that they understand and are held accountable for their roles and responsibilities in developing ethical chatbots.
2. **Involve chatbot users and content experts in design:** Work with representatives of the community of users (especially those who are habitually excluded) and service providers to develop, test and validate the chatbot, its conversational flows, content, tone, personality, accuracy, and user interface.
3. **Consistently improve**: Stay abreast of the latest developments in LLM techniques, especially in terms of continuous learning and to ensure constant advances in the quality and to reduce bias, drift, and other common challenges with LLM applications.
4. **Minimize the collection of data:** Only collect data that is required for a specific, legitimate purpose and clearly communicate to people and communities about the data's intended use. Anonymize data as soon as possible and destroy once the data has served its legitimate purpose.
5. **Ensure data privacy and security:** Secure data in accordance with relevant legal regulations and best practices (strong passwords, access controls, encrypted transmission and storage, etc) to reduce the likelihood of data breaches and privacy violations and regularly help users learn about contextually relevant practices to protect themselves.
6. **Ensure safeguarding good practices:** Develop, test, and regularly validate specific responses to sensitive topics and questions, disclosures, and urgent requests for help or support from chatbot users. Make sure that these are properly addressed by the chatbot, and that response, reporting, and referral pathways are clear and functional. If sensitive data is shared in safeguarding or protection cases, ensure that data is protected, securely stored, and kept confidential in case it is needed for a legal response.
7. **Include complaints mechanisms:** All chatbots will include a mechanism whereby users can report problems or challenges with the technology. Complaints will be regularly reviewed and responded to and used to improve the product.
8. **Assess and evaluate regularly:** Monitor regularly and periodically evaluate LLMs and chatbots for fairness, robustness, transparency, accountability, bias, drift, and resilience to hacking. Make this an iterative process and involve users in identifying and resolving biases where possible.
9. **Promote transparency:** The development process and decision-making criteria for LLMs and chatbots will be open, transparent, and well-documented. Explainability reports can help with transparency and will be produced regularly.
10. **Ensure accountability:** Establish accountability mechanisms, including legal contracts and written agreements, that hold all stakeholders and partners responsible for following Girl Effect's ethics guidelines.

Comparison of Girl Effect's Principles and Standards and the 5 Common AI Ethical Principles

Girl Effect's Responsible Data Principles and Standards	Common AI Ethical Principles [40]
Privacy and protection	Non-Maleficence, Privacy
Data minimization	Privacy, Responsibility
Transparency and accountability	Transparency, Responsibility, Justice and Fairness
Safeguarding	Non-maleficence
Lawful bases	Justice and Fairness
Data subject rights	Justice and Fairness
Meaningful choice and control	Justice and Fairness
Standard 1. Responsible design	Responsibility, Transparency, Accountability
Standard 2. Responsible data collection	Justice and Fairness, Transparency
Standard 3. Responsible data storage and transmission	Privacy
Standard 4. Responsible data use, re-use and data sharing	Privacy, Accountability, Responsibility
Standard 5. Responsible data retention and destruction	Privacy

REFERENCES

1. Wallin, J., Michel, K., Hemmings, J. and Evans, L. (2020). Girl Effect: Understanding Girls, Expanding Choices. Social and Behaviour Change White Paper prepared for Girl Effect (Global).
2. Shwartz, S. (n.d.). Supervised Learning. AI Perspectives. Retrieved June 2023, from https://www.aiperspectives.com/supervised-learning/
3. Shwartz, S. (n.d.). Deep Learning. AI Perspectives. Retrieved June 2023, from https://www.aiperspectives.com/deep-learning/
4. Shwartz, S. (n.d.). Deep Learning. AI Perspectives. Retrieved June 2023, from https://www.aiperspectives.com/natural-language-processing/#57_Language_models
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. https://arxiv.org/abs/1706.03762
6. Hu, E., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/abs/2106.09685
7. Hype Cycle for Emerging Tech, 2022 [Diagram]. (2022). Gartner. https://emtemp.gcom.cloud/ngw/globalassets/en/articles/images/hype-cycle-for-emerging-tech-2022.png
8. OpenAI. (2022). Introducing ChatGPT. Retrieved June 2023, from https://openai.com/blog/chatgpt
9. VentureBeat. (2021). OpenAI makes GPT-3 generally available through its API. Retrieved June 2023, from https://venturebeat.com/ai/openai-makes-gpt-3-generally-available-through-its-api/
10. OpenAI. (2023). Introducing ChatGPT and Whisper APIs. Retrieved June 2023, from https://openai.com/blog/introducing-chatgpt-and-whisper-apis
11. Lebedeva, I., Moklyak, O. (2023). ChatGPT vs GPT-4 vs GPT-3: key differences and applications for business. Retrieved June 2023, from https://greenice.net/chatgpt-vs-gpt-4-vs-gpt-3/#:~:text=The%20API%20uses%20the%20new,%240.004%20per%201k%20output%20tokens
12. OpenAI. (n.d.). GPT-4 is OpenAI's most advances system, producing safer and more useful responses. Retrieved June 2023, from https://openai.com/gpt-4
13. OpenAI. (2022). Aligning language models to follow instructions. Retrieved June 2023, from https://openai.com/research/instruction-following
14. OpenAI. (2022). GPT-4 Technical Report. Retrieved June 2023, from https://openai.com/research/gpt-4 or https://arxiv.org/pdf/2303.08774.pdf
15. OpenAI. (2021). GPT-3 powers the next generation of apps. Retrieved June 2023, from https://openai.com/blog/gpt-3-apps
16. OpenAI. (2023). Introducing GPTs. Retrieved November 2023, from https://openai.com/blog/introducing-gpts
17. OpenAI. (n.d.). Text generation models. Retrieved June 2023, from https://platform.openai.com/docs/guides/text-generation
18. Shwartz, S. (n.d.). Word embeddings. AI Perspectives. Retrieved June 2023, from https://www.aiperspectives.com/natural-language-processing/#56_Word_embeddings
19. OpenAI. (n.d.). Embeddings. Retrieved June 2023, from https://platform.openai.com/docs/guides/embeddings/what-are-embeddings
20. OpenAI. (n.d.). Fine-tuning. Retrieved June 2023, from https://platform.openai.com/docs/guides/fine-tuning
21. OpenAI. (n.d.). Speech to text. Retrieved June 2023, from https://platform.openai.com/docs/guides/speech-to-text
22. OpenAI. (n.d.). Assistants API. Retrieved November 2023, from https://platform.openai.com/docs/assistants/overview

REFERENCES

23. Anthropic. (2023). Introducing Claude. Retrieved June 2023, from https://www.anthropic.com/index/introducing-claude
24. Cohere. (n.d.). The Cohere Platform. Retrieved June 2023, from https://docs.cohere.com/docs/the-cohere-platform
25. Leahy, C. (2022). Announcing GPT-NeoX-20B. EleutherAI. Retrieved June 2023, from https://blog.eleuther.ai/announcing-20b/
26. Meta. (2022). Democratizing access to large-scale language models with OPT-175B. Retrieved June 2023, from https://ai.meta.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/
27. BigScience. (2022). BigScience Large Open-science Open-access Multilingual (BLOOM) Language Model. Retrieved June 2023, from https://huggingface.co/bigscience/bloom or https://arxiv.org/abs/2211.05100
28. Wiggers, K. (2022). A year in the making, BigScience's AI language model is finally available. TechCrunch. Retrieved June 2023, from https://techcrunch.com/2022/07/12/a-year-in-the-making-bigsciences-ai-language-model-is-finally-available/?guce_referrer=aHR0cHM6Ly93d3cuYWIwZXJzcGVjdGl2ZXMuY29tLw&guce_referrer_sig=AQAAAIwq3wHKzNFS6zS5su2dkrGCJoQSTNL9tZiu8m_78Nn8RIvZIDtqdWU_GOX8tuwnuDrSpdp_pxGeNM-vv5cD9OI7dFtoLjJM6bxZLeeQS3h1LxdFQtdAv6VfiO2W2F1SO3bF-Q7zbTfk8U4qMZMeI1XG4IvzHhdMjiWpCJIX6dyE&guccounter=2
29. Meta. (2023). Introducing LLaMA: A foundational, 65-billion-parameter large language model. Retrieved June 2023, from https://ai.meta.com/blog/large-language-model-llama-meta-ai/
30. Cox, J. (2023). Facebook's powerful large language model leaks online. Vice. Retrieved June 2023, from https://www.vice.com/en/article/xgwqgw/facebooks-powerful-large-language-model-leak-online-4chan-llama
31. Patel, D. and Ahmad, A. (2023). Google "We have no moat, and neither does OpenAI". SemiAnalysis. Retrieved June 2023, from https://www.semianalysis.com/p/google-we-have-no-moat-and-neither
32. The Vicuna Team. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. Lmsys Org. Retrieved June 2023, from https://lmsys.org/blog/2023-03-30-vicuna/
33. Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S. and Song, D. (2023). Koala: a dialogue model for academic research. Berkeley Artificial Intelligence Research. Retrieved June 2023, from https://bair.berkeley.edu/blog/2023/04/03/koala/
34. Shwartz, S. (n.d.). Unsupervised Learning. AI Perspectives. Retrieved June 2023, from https://www.aiperspectives.com/unsupervised-learning/
35. Shwartz, S. (n.d.). Recommender Systems. AI Perspectives. Retrieved June 2023, from https://www.aiperspectives.com/unsupervised-learning/#44_Recommender_systems
36. Hugging Face. (n.d.). The AI community building the future. Retrieved June 2023, from https://huggingface.co/
37. OpenAI. (2022). Introducing Whisper. Retrieved June 2023, from https://openai.com/research/whisper
38. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D. and Finn, C. (2023). Direct preference optimization: your language model is secretly a reward model. arXiv. Retrieved June 2023, from https://arxiv.org/abs/2305.18290
39. Shwartz, S. (n.d.). Latent Dirichlet allocation. AI Perspectives. Retrieved June 2023, from https://www.aiperspectives.com/natural-language-processing/#7328_Latent_dirichlet_allocation
40. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389-399. Retrieved July 2023, from http://ecocritique.free.fr/jobin2019.pdf

AUTHOR:

Soma Mitra-Behura – Data Scientist, Girl Effect

Drafted November 2023

CONTRIBUTORS:

Karina Rios Michel – Chief Creative and Technology Officer, Girl Effect
Janet Kasdan – Fractional CTO, Girl Effect
Steven Shwartz – Artificial Intelligence and Machine Learning Expert
Linda Raftree – Artificial Intelligence and Machine Learning Ethics Expert
Johanna Wallin – Social Behavior Change Expert

ACKNOWLEDGEMENTS:

This work would not have been possible without the support of **Dimagi**, a global social enterprise enabling impactful frontline work through scalable digital solutions and expert services.

CONTACT:

Soma Mitra-Behura soma.mitra-behura@girleffect.org
Karina Rios Michel karina.michel@girleffect.org

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING VISION