

**R42**

## **A Human-Centric Approach to Evaluating the AI Mental Health App Landscape**

Cohort 7

### **Team AI Ethics**

Members: Eddie Carillo, Sophia Cummings,  
Uma Sarkar, Ashu Mishra, Wallie Butler  
Mentors : Ossian Broman, Laura Schwab-Reese

## Table of Contents

<b>Table of Contents</b> .....	<b>2</b>
<b>1. Abstract</b> .....	<b>3</b>
<b>2. Introduction</b> .....	<b>4</b>
2.1 Research and Discovery: Understanding Needs and Ethical Dimensions.....	4
2.2 Integrating Ethics into AI Solutions - Developing Our Framework.....	4
2.3 Operationalizing the Framework and Future Directions.....	5
<b>3. Current Landscape of AI Chatbots in Mental Health</b> .....	<b>6</b>
3.1 Historical Context.....	6
3.2 Overview of Existing Chatbots.....	7
3.3 Regulatory Landscape.....	8
<b>4. Ethical Concerns in AI Chatbots</b> .....	<b>10</b>
4.1 Data Privacy.....	10
4.2 Addictive Codependency.....	10
4.3 Quality of Advice.....	11
<b>5. Framework for Ethical Evaluation</b> .....	<b>12</b>
<b>5.1 Purpose and Goals</b> .....	<b>12</b>
<b>5.2 Sources of inspiration for our Framework</b> .....	<b>12</b>
5.3 Categories of Evaluation.....	14
5.4 Results of Evaluation.....	16
<b>6. Custom-GPT Ethical Advisor</b> .....	<b>18</b>
6.1 Introduction to Custom-GPT.....	18
6.2 Training and Methodology.....	18
6.3 Simulated Interactions.....	19
<b>7. Comparative Analysis of Other Platforms</b> .....	<b>22</b>
7.1 Overview of Considered Platforms.....	22
7.2 Benchmark testing.....	23
<b>8. Future Directions</b> .....	<b>25</b>
8.1 Goals for Expansion.....	25
8.2 Recommendations for Stakeholders.....	26
<b>9. Conclusion</b> .....	<b>27</b>
<b>10. References</b> .....	<b>28</b>

## 1. Abstract

The increasing prevalence of mental health issues globally has created a pressing need for accessible and effective therapeutic solutions. In recent years, artificial intelligence (AI) chatbots have emerged as a promising tool to bridge the gap in mental health care delivery, offering support to those who may not have access to traditional therapy. This whitepaper delves into the ethical landscape surrounding the use of AI chatbots in mental health therapy, providing a comprehensive evaluation of their effectiveness, potential risks, and ethical implications- all through a human-centric lens.

In our evaluation of existing frameworks for ethical assessment of AI mental health chatbots, we observed that most frameworks were designed primarily for creators of AI mental health companies. While these frameworks are invaluable for refining products, we identified an equally important need: empowering users to independently evaluate and understand their options. Recognizing that users have unique needs and values, we adapted existing frameworks to pose human-centric questions that users could seek answers to, assessing the safety, security, and appropriateness of various AI chatbots.

**To implement these findings, we introduce our custom GPT Ethical Advisor, a tool designed to evaluate AI chatbots against our robust ethical framework.** This advisor is trained to assess various platforms using specific evaluation criteria, ensuring that users can make informed decisions about their mental health care options. We present simulated user interactions to demonstrate the advisor's practical application and effectiveness.

In addition to our primary analysis, we offer a comparative review of other platforms considered during our research, providing a balanced view of the pros and cons of each. This comparative analysis highlights the strengths and weaknesses of different approaches, guiding stakeholders in making informed choices.

Finally, we outline future directions for research and development, emphasizing the need for ongoing innovation and ethical vigilance. Our recommendations aim to support developers, regulators, and users in navigating the evolving landscape of AI in mental health therapy.

## 2. Introduction

Our journey into AI Ethics began as part of the R42 Group, where we were assigned to the AI Ethics cohort with a clear mission: to identify a problem area ripe for AI innovation while ensuring that ethical considerations were deeply embedded in our approach. This mission was crucial, as the rapid advancement of AI technologies necessitated a focus on ethics to safeguard both users and societal values.

### 2.1 Research and Discovery: Understanding Needs and Ethical Dimensions

We were inspired by a previous R42 cohort's work in identifying the pillars of safe and ethical AI, and adopted their pillar of "Human Centricity" as our guiding ethical principle.

Human Centricity emphasizes designing technologies that prioritize human needs, values, and experiences—a critical consideration in the sensitive field of mental health as well as core to the Ethics discipline overall.

Through engagements with therapists and patients, we uncovered a pressing need for AI-driven solutions to address the shortage of counselors and enhance the effectiveness of the available therapeutic support ecosystem. We observed a significant desire for AI-driven solutions to address the shortage of mental health counselors and enhance the effectiveness of existing therapeutic support.

In alignment with our team, we zeroed in on the mental health space as a promising domain for AI innovation. This choice was driven by two pivotal factors: the surge in mental health challenges, exacerbated by the COVID-19 pandemic, and the urgent need for user-oriented ethical frameworks in understanding emerging technologies.

The mental health sector is undergoing significant transformation, with an increasing demand for innovative solutions to support both patients and professionals. However, this space is also overcrowded with numerous apps positioned as solutions, creating a challenge for users to evaluate the ones that truly meet their needs. Users often lack the tools to assess whether these apps are aligned with their fundamental needs. This crowded landscape (41+ AI therapists) presented an ideal opportunity for us to integrate human-centricity into helping users navigate this crowded landscape while addressing ethical considerations comprehensively [1].

### 2.2 Integrating Ethics into AI Solutions - Developing our Framework

Armed with research insights, we targeted a high-impact area: empowering users seeking mental health care through AI to make evidence-based decisions tailored to their unique needs. To address this, we set out to develop a framework upon which to evaluate available mental health apps.

The crowded landscape of AI-powered apps in this space highlighted a critical need for a structured approach to evaluating the available products effectively.

We crafted a framework grounded in Human Centricity, designed to evaluate key aspects such as data privacy and safety across 30+ dimensions.

This framework is aimed to help users navigate the crowded market by providing a reliable method that assesses if these apps meet their fundamental needs and adhere to ethical principles. It narrows the larger ethical question into the parameters of Privacy, Effectiveness, Equity, Safety, Transparency, and Engagement.

Our framework was initially tested on a select few mental health apps. After iterative refinements, we successfully expanded it to assess over 10 different applications in the market (US and International).

This validation process demonstrated the framework's effectiveness in equipping users with the necessary tools to make informed decisions and trust AI-powered mental health solutions.

### 2.3 Operationalizing the Framework and Future Directions

We leveraged a Custom-GPT, AI platform to operationalize our framework, ensuring it was easily accessible and usable for a large audience. The successful deployment of this framework demonstrated its capacity to support diverse AI applications, affirming its potential to drive ethical innovation in mental health and beyond.

We have tested the framework across various user scenarios to ensure it effectively meets their needs, and can scale seamlessly to become the go-to resource for users seeking support in the Mental Health space.

Our journey from initial research to the final whitepaper demonstrates our commitment to ethical innovation. The mental health sector, with its distinct challenges and opportunities, provides an ideal context for applying our principles and creating an impactful solution.

We are dedicated to launching a human-centric applied ethical framework that will demonstrate that ethics can be seamlessly incorporated into AI-powered mental health solutions. This framework will serve as a foundational tool for users to evaluate and select applications that prioritize their needs and well-being. By leveraging learnings from the launch, we aim to refine our approach and scale our solution to address a wider range of unmet user needs in the mental health space.

## 3. Current Landscape of AI Chatbots in Mental Health

### 3.1 Historical Context

The history of mental health chatbots dates back to the 1960s when Professor Joseph Weizenbaum of MIT created ELIZA. ELIZA functioned using substitution methodology and pattern matching which allows it to simulate a human conversation. A famous variation of ELIZA, called DOCTOR, was programmed to behave like a psychotherapist. Essentially, DOCTOR would respond to questions by flipping the prompted questions back on the user, thus simulating a Rogerian psychotherapist [2]. The therapist use of DOCTOR reveals a crucial aspect of mental health care– the human desire for a non-judgemental listener. Chat therapists could theoretically fulfill this need, after all, they are consistently available, non-judgmental, and confidential, offering a space for individuals to explore their thoughts and feelings without fear of social stigma or judgment. This introspective use of ELIZA was not intended by Weizenbaum, but it raised the question: Can artificial intelligence be as effective as human therapists? While the interactions with ELIZA specifically lacked depth as opposed to genuine therapy, this instance opened a window toward the potential of AI to provide non-critical support.

Following ELIZA, chatbots continued to evolve, reflecting improvements in the technology, as well as different expectations from users. In the current landscape, there are a plethora of AI chatbot applications that offer mental health support. Below is a timeline of the milestones in therapeutic chatbot development.

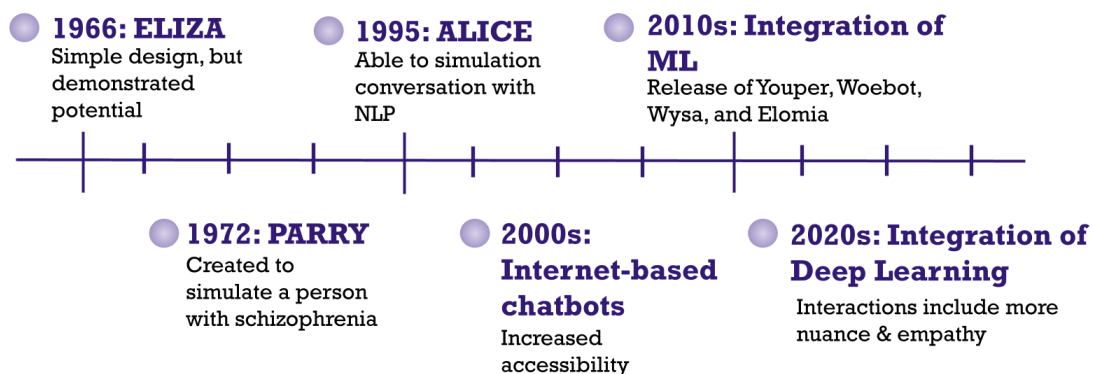


Figure 1: Timeline of Milestones in Therapeutic Chatbot Development [2][3][4]

## 3.2 Overview of Existing Chatbots

As the landscape of AI chatbots in therapeutics continues to evolve and expand, so do the capabilities of the chatbots. The leading applications for AI therapy leverage significant innovation and unique advancements from natural language processing techniques to deep learning algorithms that enable nuanced, empathetic conversations. In this section, we look into the leading AI therapy chatbots in the market today, examining the key features as well as each app's unique approach to accessible mental health support.



Figure 2: Existing Chatbots on the Market [4][5]-[14]



### 3.3 Regulatory Landscape

In the United States, the regulation and legal aspects governing AI chatbots are still in a nascent stage with no existing national data privacy law. Currently, twelve U.S. states have enacted laws to protect data privacy, specifically in the context of AI. California leads the way with the California Consumer Privacy Act (CCPA) and its successor, the California Privacy Rights Act (CPRA), which grants residents significant control over their personal data and impose stringent requirements on businesses regarding data collection and usage. Other states, including Virginia (Consumer Data Protection Act), Colorado (Colorado Privacy Act), and Connecticut (Data Privacy Act), have also implemented comprehensive data privacy regulations. These laws often mandate transparency, consent, data access, correction, and deletion, as well as specific provisions for automated decision-making and profiling. While these state-level laws provide essential protections, the absence of a comprehensive federal data privacy law leaves the U.S. without a uniform standard for AI and data privacy. Nonetheless, various existing federal and state laws regulate specific sectors, types of data, or concerns related to these technologies.

#### **Healthcare Regulations and Oversight:**

The Food and Drug Administration (FDA) plays a critical role in overseeing digital health technologies, including AI-powered tools. The FDA's regulatory framework, however, primarily focuses on medical devices and software that meet the definition of a medical device [3]. AI chatbots designed for mental health may fall under this category if they are intended for diagnosing, curing, mitigating, or treating mental health conditions. The FDA's approach to AI in healthcare is evolving, with recent initiatives like the Digital Health Innovation Action Plan and the proposed framework for AI/ML-based software as a medical device (SaMD)[4]. However, the dynamic nature of AI technologies, including continuous learning systems, poses challenges for regulatory oversight.

#### **Data Privacy and Security:**

The Health Insurance Portability and Accountability Act (HIPAA) is a key piece of legislation that governs the privacy and security of patient health information in the U.S. AI mental health chatbots that handle protected health information must comply with HIPAA's requirements, including the privacy rule, security rule, and breach notification rule[5]. However, compliance can be complex, particularly when AI chatbots are developed by tech companies that may not be traditional healthcare providers or when the chatbot's services are offered directly to consumers outside of a healthcare setting.

#### **Consumer Protection and Ethical Considerations:**

The Federal Trade Commission (FTC) is another relevant regulatory body, with a mandate to prevent deceptive or unfair business practices. The FTC has expressed interest in the ethical and truthful marketing of AI technologies, including mental health chatbots. Issues such as



the accuracy of AI chatbots' responses, the transparency of their capabilities and limitations, and the disclosure of data use practices are critical from a consumer protection perspective. Misrepresentation of the chatbot's abilities or failure to adequately inform users about data handling can lead to legal consequences [6].

Overall, the regulatory environment for AI mental health chatbots in the U.S. is characterized by a patchwork of existing laws and a developing landscape of AI-specific guidance and regulation. As the technology continues to evolve, there is a growing need for clear and comprehensive regulatory frameworks that address the unique challenges posed by AI in mental health, balancing innovation with safety, privacy, and ethical considerations.

### **Mandated Reporting**

AI mental health chatbots present a unique challenge in responding to suicidal and harmful thoughts from users. Chatbots are not currently mandated reporters. This means they are not legally required to report information such as suicidal thoughts, reports of abuse, or confessions of harm to authorities. This is partly because AI chatbots are not yet recognized as capable of making the nuanced judgments required for mandated reporting. Mandating AI chatbots to report could raise significant privacy issues, especially if the data is handled by third-party companies [7]. Ensuring user confidentiality while meeting reporting obligations is a complex challenge that has not been solved yet.

In addition, their responses are based on pre-programmed algorithms, which can sometimes result in generic or inadequate responses to serious situations. The concern is that an AI may not fully grasp the complexity or urgency of a user's condition, potentially leading to delayed or inappropriate intervention. Misreporting or failing to report could have serious consequences, making the role of AI in mandated reporting contentious[8]. As of now, most chatbots are designed to recommend resources such as hotlines or possible options for human intervention for the user.

## 4. Ethical Concerns in AI Chatbots

### 4.1 Data Privacy

The collection, storage, and use of personal data in AI mental health chatbots raise several critical issues, primarily revolving around privacy, security, and consent. These chatbots often handle sensitive and confidential information, including users' mental health conditions, treatment history, and emotional states. The potential for data breaches or unauthorized access poses a significant risk to users' privacy and can lead to severe consequences, such as identity theft or personal distress [9].

Furthermore, there is the question of how transparently these technologies communicate their data practices to users. Ensuring informed consent is crucial, yet users may not fully understand how their data is collected, used, or shared, especially when algorithms are involved. The use of personal data for training and improving AI systems also raises ethical concerns about de-identification and the potential for re-identification of anonymized data. As AI mental health chatbots become more prevalent, these issues underscore the need for stringent data protection measures, clear communication with users, and adherence to ethical standards in data handling [10].

A notable case involving privacy issues with mental health chatbots occurred with the platform BetterHelp. In this instance, BetterHelp faced scrutiny and potential class action lawsuits due to allegations of sharing users' sensitive mental health information with advertisers without proper consent. The company ultimately settled with the Federal Trade Commission (FTC) for \$7.8 million in 2023 [11]. This case highlighted significant concerns about the potential for misuse of personal data collected by mental health apps, especially when users trust these platforms with deeply personal information during vulnerable times.

### 4.2 Addictive Codependency

The potential for users to become overly reliant on AI mental health chatbots is a notable concern. These tools offer convenient and immediate support, making them appealing, especially for those with limited access to traditional mental health care. However, this convenience can lead to excessive dependence, potentially discouraging users from seeking in-person therapy or engaging with social support networks. Unlike human therapists, chatbots lack nuanced emotional understanding, which is crucial in complex mental health situations. The perceived anonymity and ease of use of chatbots can sometimes encourage excessive engagement, where users repeatedly turn to the AI for reassurance, reinforcing unhealthy coping mechanisms. This reliance raises ethical questions about the appropriate scope and limits of AI in mental health care, emphasizing the need for balanced use and clear guidelines to ensure users maintain access to holistic and human-centered therapeutic resources [12].

### 4.3 Quality of Advice

Instances of harmful or misleading advice given by AI mental health chatbots have emerged as a significant concern in the field. In various cases, chatbots designed to offer mental health support have provided responses that were either overly simplistic or misaligned with established therapeutic practices. For example, some chatbots have suggested potentially dangerous self-help strategies without proper context or caution, such as minimizing the severity of a user's symptoms or endorsing unverified coping techniques. These instances can lead to increased anxiety, confusion, or a false sense of security among users, particularly when the chatbot's advice lacks the nuance and personalization that a human therapist would provide. Such issues highlight the critical need for robust oversight and continuous refinement of AI mental health tools to ensure they deliver safe, accurate, and supportive guidance [13].

In May 2023, the National Eating Disorders Association (NEDA) faced significant backlash when its chatbot, Tessa, was found to be providing harmful and inappropriate advice to individuals seeking support for eating disorders. Tessa, designed to offer guidance and resources for those struggling with eating disorders, instead offered advice that was counterproductive and potentially dangerous. Users reported that the chatbot suggested unhealthy eating habits and weight loss tips, which contradicted the principles of recovery and could exacerbate disordered behaviors. The controversy led to NEDA suspending the chatbot [14].

This case was particularly alarming because it highlighted the limitations of AI in handling complex and critical mental health scenarios. The chatbot's inability to provide appropriate escalation or connect the user with emergency support resources demonstrated the potential risks of relying solely on AI for mental health care. This incident underscores the need for AI systems to be integrated with human oversight and intervention mechanisms to ensure that users in crisis receive appropriate and timely support.

## 5. Framework for Ethical Evaluation

### 5.1 Purpose and Goals

The ethical risks associated with the improper design and implementation of AI in mental health solutions can have significant impacts. Coupled with the low barrier to entry for new products and the absence of overarching regulatory institutions, these risks can lead to dangerous ramifications. This highlights the clear need for a robust ethical framework. However, most existing frameworks are designed primarily for developers of AI-based mental health applications. Our goal was to create a set of criteria that allows for the evaluation of these applications from the perspective of the average user, guided by our key ethical principles.

### 5.2 Sources of inspiration for our Framework

In developing our criteria and evaluation questions for mental health apps, we drew on a variety of existing frameworks and guidelines to ensure that our approach was both comprehensive and grounded in established ethical and practical standards. These sources ranged from the READI framework, which focuses on the safe and ethical deployment of AI in mental health, to the American Psychological Association's Ethical Principles and App Evaluation Model, which emphasize ethical conduct and practical considerations in mental health practices. Additionally, we considered broader ethical guidelines, such as those outlined in the Blueprint for an AI Bill of Rights and the Belmont Report, which underscore the importance of safety, equity, privacy, and informed consent in the development and use of AI systems in healthcare.

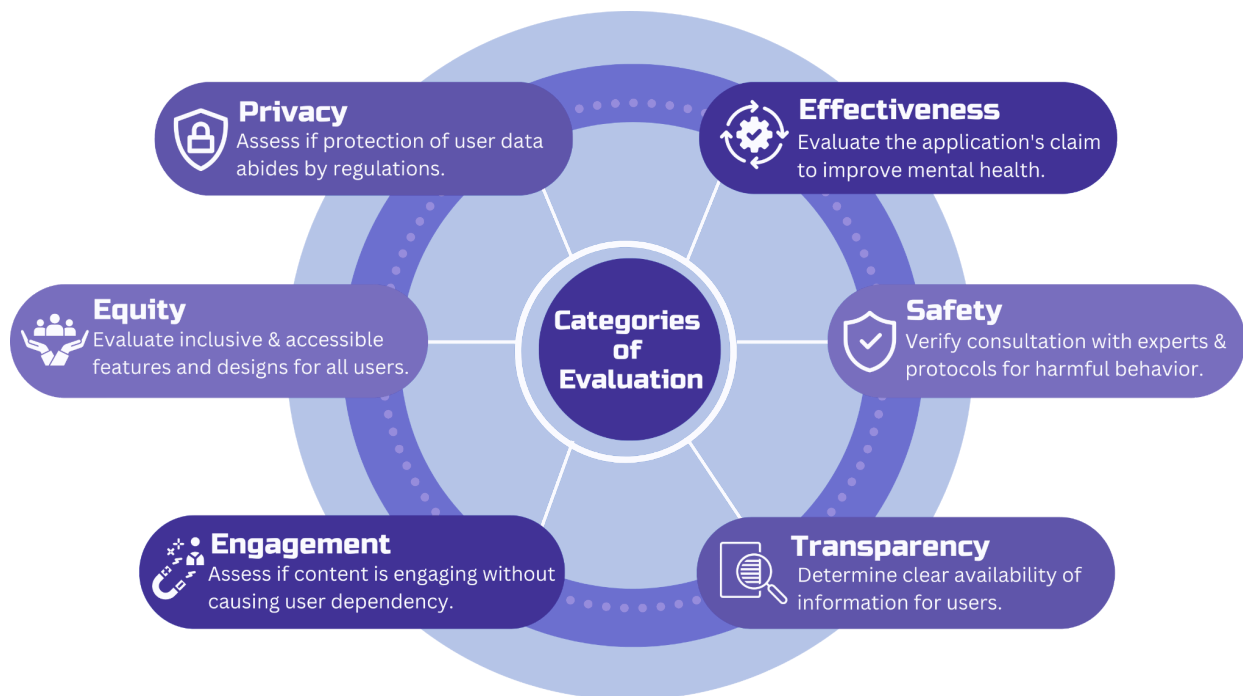
Although these frameworks varied in focus—some tailored for developers and policymakers, others for clinicians and researchers—a few key principles consistently emerged. These included the importance of ensuring safety, promoting effectiveness, safeguarding privacy, fostering equity, enhancing user engagement, and maintaining transparency. Recognizing the significance of these principles, and keeping with our theme of human-centricity, we adapted them to be more applicable to the actual users seeking a mental health app, instead of the developers of the products. Our hope was this adaptation would empower users to make informed decisions when choosing a mental health app that best suits their needs and allow them to navigate the vast landscape of mental health apps on the market.

We drew inspiration from the following:

- **READI (Readiness for AI Deployment and Implementation)**  
 Stanford University's READI [15] framework is designed to evaluate AI-mental health applications to ensure they are safe, effective, and ethically sound before deployment. It addresses the challenges of integrating AI into mental health care, helping developers, clinicians, and healthcare organizations make informed decisions about deploying AI tools.
- **Ethical Principles of Psychologists and Code of Conduct**  
 The American Psychological Association's Ethical Principles of Psychologists and Code of Conduct [16] provides guidelines to ensure psychologists maintain high ethical standards in their professional roles. It addresses key principles such as beneficence, responsibility, integrity, justice, and respect for people's rights. This framework helps psychologists make ethical decisions, safeguarding the welfare of clients, patients, and the community.
- **APA App Evaluation Model**  
 The APA's App Evaluation Model [20] is specifically designed for clinicians to assess mental health apps, focusing on critical factors like access, privacy, clinical foundation, usability, and data integration. Recognizing that selecting apps differs from traditional therapeutic decisions, this model provides a structured rating system to help clinicians make informed choices, ultimately supporting improved clinical decision-making and patient outcomes.
- **Blueprint for an AI Bill of Rights**  
 The Blueprint for an AI Bill of Rights [17] outlines principles to ensure that AI systems are safe, fair, and respect privacy. It provides guidelines for developers and policymakers to protect public rights, mitigate risks, and promote equity in AI usage. It aims to ensure that AI technologies benefit society while upholding civil liberties and democratic values.
- **Artificial Intelligence in Health, Healthcare, and Biomedical Science: An AI Code of Conduct Principles**  
 The "Artificial Intelligence in Health, Health Care, and Biomedical Science: An AI Code of Conduct Principles and Commitments" [18] framework provides ethical guidelines for AI use in healthcare. It emphasizes safety, effectiveness, equity, transparency, inclusive collaboration, continuous safety assessments, and environmental considerations to ensure AI technologies benefit patients and the health system.
- **The Belmont Report**  
 The Belmont Report [19] outlines ethical principles for research involving human subjects. It emphasizes respect for persons (informed consent), beneficence (maximizing benefits and minimizing harm), and justice (fair distribution of research benefits and burdens). Its goal is to promote ethical practices in biomedical and behavioral research.

### 5.3 Categories of Evaluation

When developing our own framework, we narrowed down our evaluation into six distinct categories.



*Figure 3: Pillars of our Framework*

For privacy, we look at whether the application clearly explains how it uses and shares your data. Additionally, we examine how the app safeguards your personal health information. When it comes to effectiveness, we verify whether the app provides evidence to support its claims of improving mental health. This could be in the form of clinical studies, peer-reviewed research, or even user feedback.

Safety is another important area. Here, we ask whether the app includes features to prevent harmful behavior, especially in crisis situations. We also check if the app offers easy ways for users to get help or report problems. Transparency is what it sounds like. Essentially, the more information the app provides—and the easier it is to access—that information, the better. For example, one question we ask is whether the app explains what it can and cannot do. Since we're talking about AI-based apps, it's also important that they explain how the AI component of the app works.

Engagement examines whether the app can maintain a user's attention long enough for them to see results or benefits, without becoming overly dependent on the app. We look for

features that help users stay on track but also highlight any that may grab a user's attention a little too much.

Finally, equity ensures that the app is accessible to everyone. Among other factors, we consider features like the languages available and whether the app is designed with consideration for people with disabilities. By using these categories as a framework, we aimed to create a tool that empowers users to make informed decisions when choosing a mental health app.

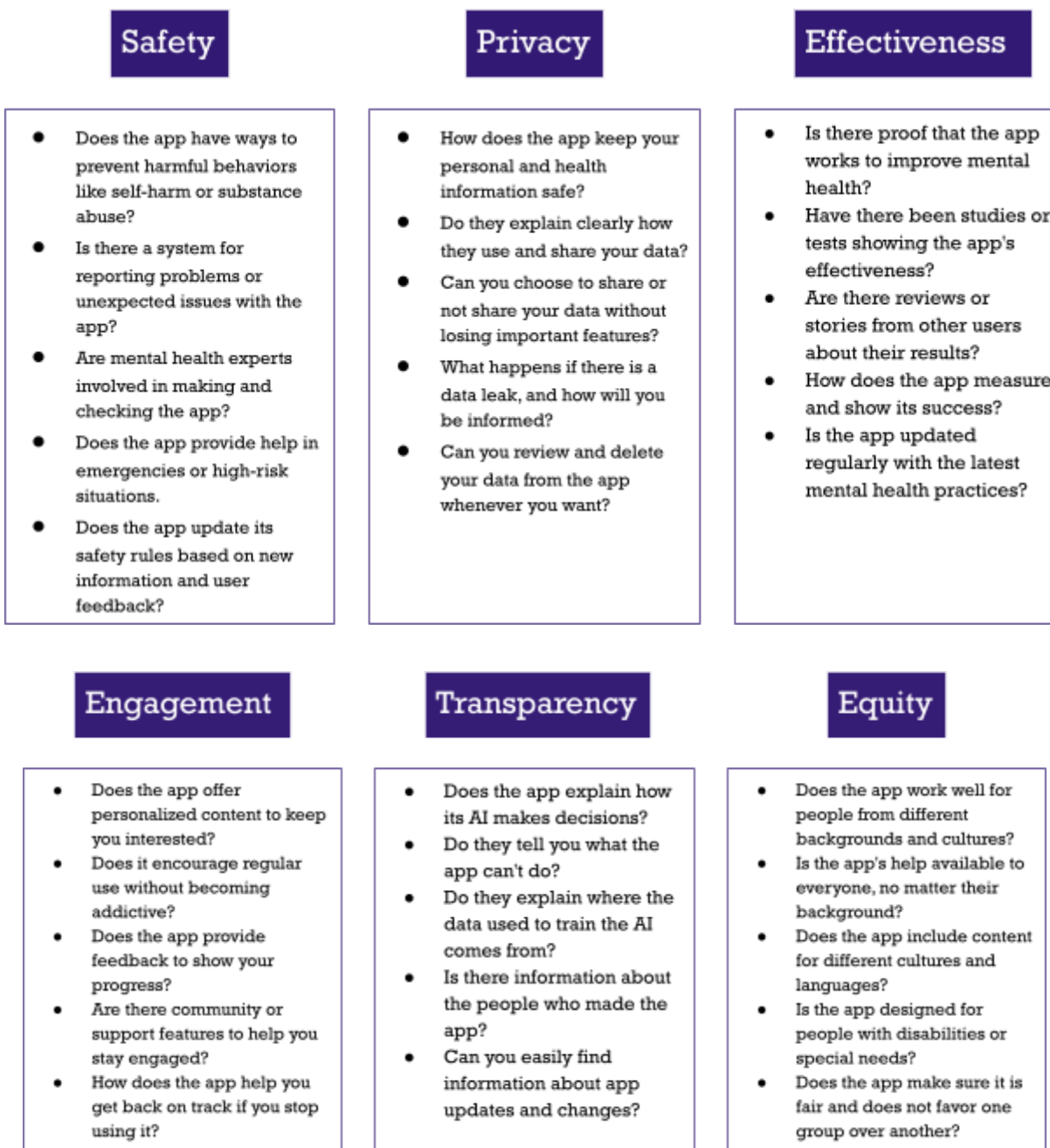


Figure 4: Evaluation Questions



## 5.4 Results of Evaluation

We evaluated ten AI-based mental health applications: Bloom, Elomia, Sonia, Youper, Kintsugi, Woebot, Mindsum, Sintelly, Mindspa, and Wysa. Our evaluation focused on information accessible to the average user, relying primarily on publicly available sources such as the applications' websites and app store pages. Each app was assessed based on the criteria mentioned above: Privacy, Effectiveness, Safety, Engagement, Transparency, and Equity.

### Key Findings from Applying the Framework:

1. **Lack of Mechanisms for High-Risk Situations:** Many mental health apps lack mechanisms to prevent harmful behaviors like self-harm and fail to provide support in high-risk situations. This is concerning, given the vulnerability of the population using these apps and the expectation that any tool or person involved in mental health care should protect users from harm. At most, these apps may direct a user to a crisis hotline or suggest they call professionals, which might not be enough.
2. **Evidence-Based Claims:** When it comes to the evidence supporting the claims made by mental health apps, only a few are backed by peer-reviewed studies and clinical trials. Those with such evidence tend to be the more prominent apps, like Wysa or Woebot, which are no longer publicly available. Smaller apps available to the average user often provide little to no evidence or merely claim to be based on CBT. This is problematic, as users seeking to improve their mental health through an app need assurance that it will genuinely be effective.
3. **Data Review and Deletion:** On a positive note, a significant number of apps, if not most, provide users with a good level of control over their personal information and data. For example, many apps allow users to review their data and even request the deletion of their personal information.
4. **Involvement of Mental Health Experts:** Many of these apps claim to involve mental health experts in their development and review; however, the extent of this involvement is not always clear. This lack of transparency is concerning, as these experts' contributions are a major factor in whether the app is both safe and effective for users.
5. **Data Safety and Security:** The safety and security of personal health information and data is one area where most apps seem to perform well. However, it is not always clear or explained how this is achieved. For example, one app might have a comprehensive privacy policy detailing its data security methods, while another may only state compliance with relevant privacy laws. It's important for users to understand how their data is being stored and to feel assured that it is safe.
6. **Personalized Content and Re-engagement:** Ensuring that someone benefits from using a mental health app requires them to actually use it consistently. It's encouraging that many apps offer personalized content and methods to re-engage users, such as push notifications. However, it's important to consider that promoting

too much engagement can have negative effects, like fostering over-dependency on the app [16]. This concern is especially relevant when the app involves a chatbot.

7. **Functionality Information:** Applications generally provide information on their functions, capabilities, and limitations, which is important as it helps users understand what to expect, how to utilize the app effectively, and whether it's right for them.
8. **Explanations of AI Functions:** When it comes to explaining how an app's AI-based functions work, only a few of the more prominent apps offer any information, and even then, the details can be limited. Often, they provide a set of AI principles rather than specifics on training or model details. It's important for users to be aware of these details to fully understand how the AI operates and make informed decisions about using the app.
9. **Data Usage and Sharing:** Applications generally provide some level of explanation on how they use and share user data; however, the level of detail can vary greatly from app to app. Data usage and sharing policies should be comprehensive, easily accessible, and understandable to ensure users are well-informed about how their data is being used.

### Challenges in Evaluation:

Applying the evaluation questions to the list of applications proved to be time-consuming and tedious. Locating the necessary answers required sifting through lengthy policy pages, navigating obscure links, and deciphering jargon. Additionally, some crucial information was either unavailable or difficult to access.

Given the nature of this process, it was concluded that simply providing a set of questions or a guide would not be sufficient to help users make informed decisions. Most individuals do not have the time or resources to conduct exhaustive research. Therefore, developing a tool to apply our work and simplifying the process seemed like a better alternative, ensuring that users can access clear and relevant information efficiently, and also increasing the likelihood that vulnerable users carry out this evaluation process at all.

## 6. Custom-GPT Ethical Advisor

### 6.1 Introduction to Custom-GPT

In the evaluation of existing frameworks for ethical assessment of AI mental health chatbots, we observed that most frameworks were designed primarily for creators of AI mental health companies. While these frameworks are invaluable for refining products, we identified an equally important need: empowering users to independently evaluate and understand their options. To address this challenge, we leveraged OpenAI's Custom-GPT functionality to create the **Custom-GPT Ethical Advisor**.

The Custom-GPT Ethical Advisor simplifies the evaluation process for users. Trained on our ethical framework, the advisor assesses AI chatbots based on the six parameters we identified as crucial for user safety and well-being from our comprehensive framework. Rather than burdening users with a lengthy questionnaire, the advisor provides concise, relevant information tailored to the user's queries. This ensures that vulnerable users, who may lack the time or resources to conduct thorough research, can stay informed about potential risks and make well-informed decisions. Our goal is to make the process of evaluating AI chatbots as easy and actionable as possible, guiding users toward the safest products that best fit their needs.

### 6.2 Training and Methodology

The training of the Custom-GPT Ethical Advisor involved multiple stages of testing and refinement. We provided the advisor with an extensive set of instructions, including our user-facing framework and a detailed list of evaluation questions. Ensuring the advisor incorporated this framework into its ethical analysis was a key priority.

The Custom-GPT Ethical Advisor underwent a rigorous training process to ensure its ability to provide accurate and ethical guidance to users. This training involved multiple stages of testing and refinement, focusing on equipping the advisor with the necessary knowledge and skills to evaluate AI-based mental health applications effectively.

Key components of our training process:

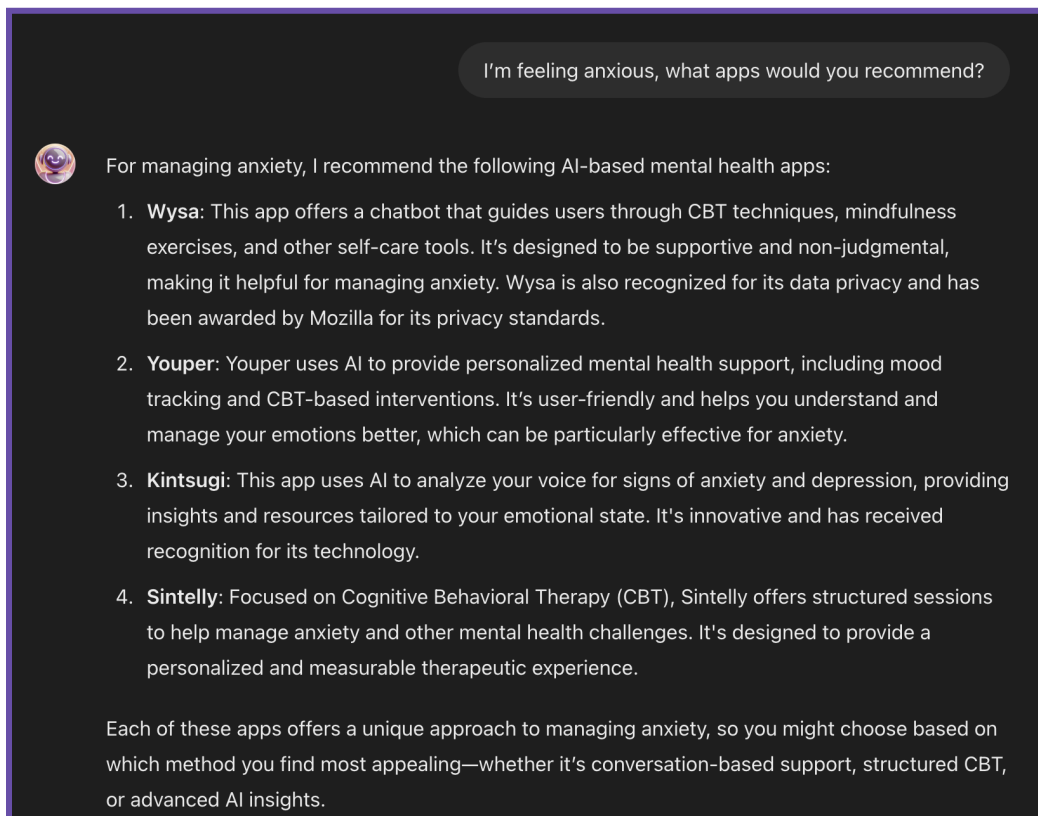
1. **Framework Integration:** The advisor was provided with the comprehensive user-oriented ethical framework designed earlier. This framework outlined key evaluation criteria, including privacy, effectiveness, safety, engagement, transparency, and equity. The advisor was trained to incorporate these criteria into its analysis of AI chatbots and present a synthesized answer to any relevant question.
2. **Knowledge Acquisition:** To equip the advisor with a deep understanding of AI mental health applications, we manually sourced relevant documentation from product websites. This included "About" pages, terms of service, and privacy policies, which were then uploaded to the advisor's training dataset.

3. **Instructional Guidelines:** The advisor was given a detailed set of instructions to guide its responses. These instructions emphasized the importance of providing clear, concise, and empathetic guidance to users. The advisor was also instructed to avoid giving medical advice and to focus on providing objective information.
4. **Internet Access Restrictions:** While the advisor has internet access, we implemented restrictions to limit its use. The advisor was instructed to use the internet primarily for searching user reviews and ratings from app stores. This helped to ensure that the advisor's responses were based on reliable and relevant information.

### 6.3 Simulated Interactions

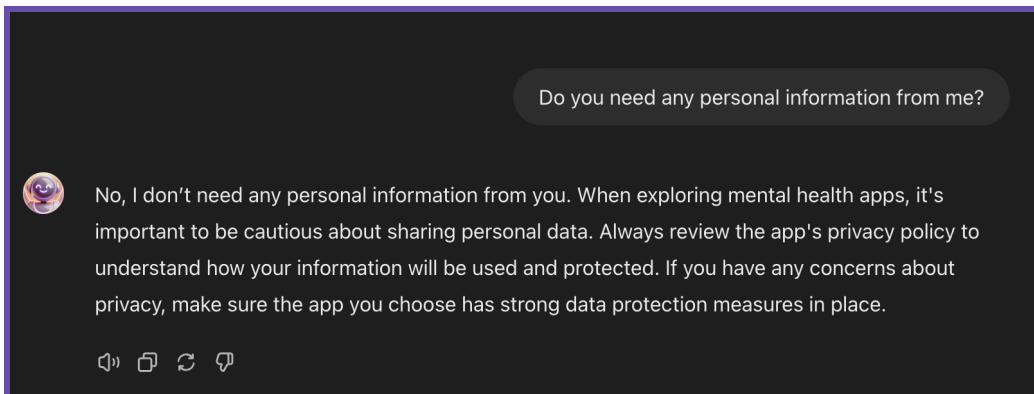
Below we will present a few example interactions with our Custom-GPT mental health advisor tool. Although we expect the content of the answers to remain consistent, the general phrasing may change from one interaction to the next despite the prompt staying the same due to the nature of generative AI. Our tool is publicly available on the following link :

<https://chatgpt.com/g/g-ICTZC91Vd-ai-based-mental-health-app-advisor-test>



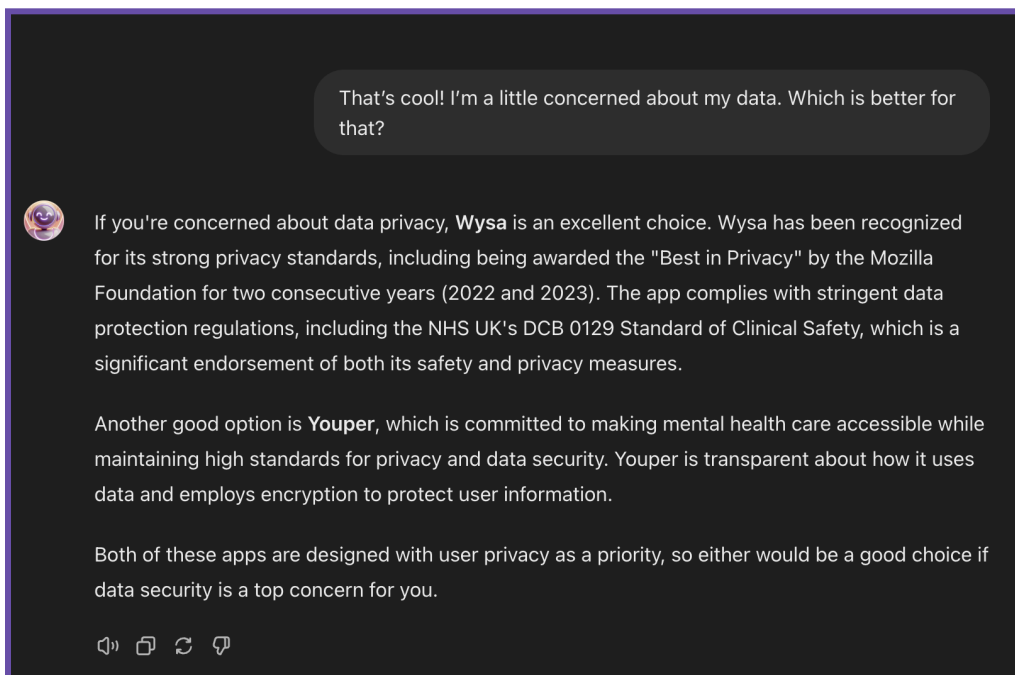
*Figure 5: General Inquiry*

We intentionally avoid asking for any personal information from the end of the user, and focus on giving general actionable advice instead. Any information that is inadvertently provided is subject to the Open AI privacy policy[21].



*Figure 6: Personal Information*


Our tool is able to expand on specific concerns and provide tailored recommendations based on whatever the user's priorities are.



*Figure 7: Data Privacy Inquiry*

The tool is able to go beyond the claims made by app developers and draw from online user reviews when making claims regarding an app's effectiveness. When asked generally about a specific app, it answers according to the parameters of our framework.

Awesome! Does it actually work?


 Yes, both **Wysa** and **Youper** are effective for many users.

- **Wysa:** Users generally find Wysa helpful, with 91% of users reporting positive experiences. The app is backed by clinical evidence and has been integrated into NHS services, which speaks to its effectiveness in real-world settings. Wysa uses evidence-based techniques like CBT, mindfulness, and mood tracking to help manage anxiety and other mental health challenges.
- **Youper:** This app is also effective, with positive feedback from users who appreciate its personalized approach to managing mental health. Youper uses AI to adapt its interventions based on your mood and behavior, providing tailored support that many users find beneficial.

Both apps are designed to support mental health, and while they are not substitutes for professional therapy, they can be very effective tools for managing anxiety and improving emotional well-being.

*Figure 8: Effectiveness Inquiry*

How is Elomia?

 **Elomia** is an AI-based mental health app designed to provide emotional support through conversations. It uses a chatbot to simulate therapeutic conversations, helping users manage stress, anxiety, and other mental health challenges.

**Key Points About Elomia:**

- **Privacy:** Elomia takes privacy seriously, and like other reputable mental health apps, it likely has policies in place to protect user data. It's important to review their privacy policy directly to understand how your data is collected, stored, and used.
- **Effectiveness:** The app aims to replicate supportive conversations that can help users cope with their emotions. While it's not a replacement for professional therapy, many users find it helpful for managing day-to-day stress and anxiety. The app is designed to be empathetic and supportive, which can be comforting during difficult times.
- **Safety:** Elomia includes disclaimers that it's not a substitute for professional help, which is important to keep in mind. It's best used as a supplementary tool rather than a primary source of mental health care.

**Summary**

Elomia can be a good choice if you're looking for an AI-based app that offers conversational support. While it may not have the extensive clinical validation of apps like Wysa or Youper, it provides a unique, empathetic experience that many users find helpful. Just ensure you're comfortable with its privacy policies and understand its limitations as a non-professional tool.

*Figure 9: App-Specific Inquiry*

## 7. Comparative Analysis of other Platforms

### 7.1 Overview of Considered Platforms

During the development of our Custom-GPT Ethical Advisor, we explored several platforms to find the most suitable tool for creating an interactive and accessible ethical evaluation interface. Our primary goals were to find a platform that allowed document uploads for incorporating our framework and narrowing the scope of the training data and to ensure the tool could be publicly accessed by anyone with a link. Below is a comparative analysis of the platforms we considered, highlighting their pros and cons.

#### Soopra

- **Pros:** Accessible by many users, and highly personalizable. Straightforward experience for the user and designer of the persona. [22]
- **Cons:** The training process emphasized the “tone” of the persona rather than the content of the knowledge base. Editing individual responses was time-consuming, making it impractical for our needs.

#### TextCortex

- **Pros:** Capable of being trained on numerous sources. The tone of responses could be adjusted with a single click in the settings. The implementation process was straightforward and user-friendly.[23]
- **Cons:** Designed primarily as a personal assistant, making it unsuitable for public dissemination and broader accessibility.

#### Character.ai

- **Pros:** Freely accessible to anyone, available on a public link for free. [24]
- **Cons:** The training process was highly meticulous, requiring detailed sample dialogues and extensive information about the persona’s preferences. This level of detail made the training process too cumbersome for our purposes.

#### Custom-GPT

- **Pros:** Focused more on the content than the “persona.” The knowledge base could be supplemented or restricted as needed. Adjusting the



behind-the-scenes instructions effectively changed the outcomes. Anyone with a link can access a published bot for free, making it highly accessible.

- **Cons:** The requirement for a premium account to create the bot was a slight inconvenience, although this didn't affect the user's experience at all since it was still free for their use.

Through this comparative analysis, we found that Custom-GPT best met our requirements, allowing us to create an ethical advisor that is both content-focused and widely accessible. The platform's flexibility in adjusting the knowledge base and instructions ensured that our framework could be effectively integrated and utilized by users, empowering them to make informed decisions about AI mental health chatbots.

## 7.2 Benchmark testing

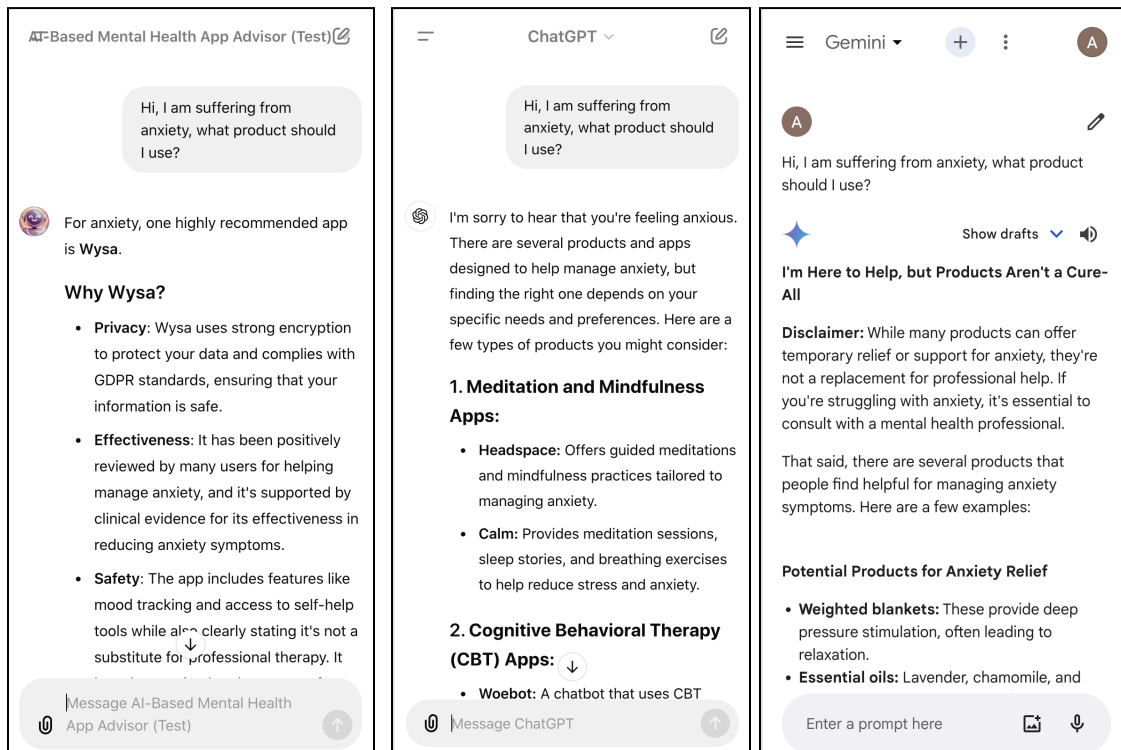


Figure 10: Benchmarking our tool vs Raw Chat GPT and Gemini

To evaluate the efficacy and superiority of our Mental Health GPT, we conducted comprehensive benchmark testing against prominent competitors, including raw GPT and Soopra. Our tests focused on critical areas such as specific goal-oriented user queries, their ethical considerations, and data privacy and security.

**Goal-Specific Responses:** Unlike raw GPT and Soopra, our Mental Health GPT excels in providing precise and actionable recommendations based on specific user queries. Users can request advice, and our GPT delivers responses that align with the parameters identified in our framework, and keep the user on topic. This focused approach ensures that users receive relevant and practical advice, enhancing their mental health journey.

**Ethical Considerations:** One of the core strengths of our Mental Health GPT is its deep integration of ethical considerations. Trained on a diverse range of mental health apps, our model is equipped to handle complex ethical scenarios, including confidentiality, consent, and appropriate intervention strategies. This capability significantly differentiates our product, as it ensures that recommendations are not only effective but also ethically sound.

**Data Privacy and Security:** Our Mental Health GPT places a paramount emphasis on data privacy and security, a crucial aspect often overlooked by other models. The training process incorporated privacy protocols issued by the apps. This helps ensure that the framework is able to evaluate this information against all available apps and offer succinct recommendations to users who are deeply interested in privacy and safety of their data. Furthermore, this focus on data security not only builds user trust but also complies with stringent regulatory requirements, making our product a reliable choice for sensitive mental health applications.

Through these benchmark tests, it is evident that our Mental Health GPT stands out in delivering goal-specific, ethically considerate, and secure mental health recommendations. This positions our product as a superior solution in the competitive landscape of mental health AI tools.

## 8. Future Directions

### 8.1 Goals for Expansion

#### **Potential Areas for Further Research and Development:**

Our Mental Health GPT currently excels in evaluating apps based on user preferences for data privacy and other ethical considerations. However, there are several key areas for expansion that can significantly enhance the product's capabilities and reach.

**1. Scale to include mental health Apps (US Market):** By incorporating a broader range of mental health apps available in the US market, we can gather diverse learnings and refine our GPT model further. This expansion will enable us to better understand the ethical dimensions and effectiveness of various apps, ensuring our recommendations are comprehensive and relevant to a wider audience.

**2. Scale to include global Apps:** Expanding our database to include international mental health apps will position our GPT as the go-to product for global users. This will provide us with a richer dataset, allowing us to gather insights from a variety of cultural and regulatory contexts. The goal is to create a truly inclusive and diverse model that can cater to the needs of users worldwide, reflecting a broad spectrum of ethical standards and practices.

**3. Enhancing User (or Provider) Feedback Integration:** Enhancing the framework to further integrate patient or provider feedback will further enable the Mental Health Advisor. It already works successfully by evaluating user feedback and by enhancing the data and training, the framework has the potential of evolving into a single source of truth for those exploring options in the Mental Health space.

**4. North Star: Broaden Access to Mental Healthcare:** Establishing partnerships with mental health app providers can broaden access to mental health products. Currently, access is often limited to users who can afford these services, creating a bias in the ecosystem (including bias in data training). By working closely with app providers, we aim to expand access to mental health resources, making them available to a wider audience and addressing the disparity in mental health support.

**Limitations:** While our current model shows significant promise, it is crucial to acknowledge its limitations. Our GPT is only as effective as the data it is trained on. Expanding our dataset and incorporating more diverse and comprehensive

feedback will be essential for overcoming these limitations and ensuring our model's continued improvement.

## 8.2 Recommendations for Stakeholders

**Developers:** Continue to enhance the ethical considerations embedded in AI models. Focus on integrating diverse datasets and user feedback to refine the accuracy and relevance of recommendations. Prioritize data privacy and security to maintain user trust and compliance with regulatory standards.

**Regulators:** Establish clear guidelines and standards for the ethical use of AI in mental health applications. Encourage explainability, transparency, and accountability in the development and deployment of AI models to ensure they serve the best interests of users.

**Users:** Engage actively with mental health AI tools and provide feedback to help improve their effectiveness. Be aware of the ethical implications of using such tools and choose solutions that prioritize data privacy and security.

By focusing on these areas, we can significantly enhance the capabilities and reach of our Mental Health GPT, ultimately providing a more comprehensive, ethical, and user-friendly solution for mental health support.

## 9. Conclusion

The proliferation of AI-based mental health applications has created both opportunities and challenges. While these applications offer accessible and potentially effective support, it is imperative to approach them with caution and a critical eye. Our research has highlighted the importance of awareness and intention when choosing AI-based mental health products, as well as the need for a human-centric approach when tackling these challenges.

Key findings from our analysis include the need for greater transparency regarding data privacy, the limitations of AI in providing support for high-risk situations, and the importance of evaluating whether or not claims made by the designers of products are backed by evidence. Furthermore, we identified the challenges associated with manually evaluating these applications due to the lack of standardization and the often opaque nature of their operations.

To address these challenges and empower users to make informed decisions, we developed our own robust framework upon which to assess mental health products, as well as the Custom-GPT Ethical Advisor- a chat-based implementation of the framework. This tool provides a comprehensive assessment of AI mental health applications, considering factors such as privacy, effectiveness, safety, engagement, transparency, and equity. By leveraging this advisor, users can face the increasingly chaotic marketplace with support, and make more informed choices about the AI-based mental health support they seek.

## 10. References

1. Science Direct, "An overview of the features of chatbots in mental health: A scoping review" Available: <https://www.sciencedirect.com/science/article/abs/pii/S1386505619307166#:~:text=According%20to%20studies%2C%20there%20are,implemented%20in%20stand%2Dalone%20software> [Accessed August 14, 2024].
2. E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, p. 100006, 2020. doi: 10.1016/j.mlwa.2020.100006. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666827020300062>.
3. "California Consumer Privacy Act (CCPA)," State of California Department of Justice, Mar. 2024. [Online]. Available: <https://oag.ca.gov/privacy/ccpa>
4. "US State by State AI Legislation Snapshot," *Bryan Cave Leighton Paisner LLP*, Jun. 2024, [Online]. Available: <https://www.bclplaw.com/en-US/events-insights-news/us-state-by-state-artificial-intelligence-legislation-snapshot.html>
5. "Software as a Medical Device (SaMD)," US Food and Drug Administration, Dec. 2018. [Online]. Available: <https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd>
6. S. Forrest, "Artificial Intelligence/ Machine Learning (AI/ML)-Enabled Medical Devices: Tailoring a Regulatory Framework to Encourage Responsible Innovation in AI/ML," US Food and Drug Administration. [Online]. Available: <https://www.fda.gov/media/160125/download#:~:text=URL%3A%20https%3A%2F%2Fwww.fda.gov%2Fmedia%2F160125%2Fdownload%0AVisible%3A%200%25%20>
7. "Health Insurance Portability and Accountability Act of 1996 (HIPAA)," US Center for Disease Control and Prevention, Jul. 2024. [Online]. Available: [https://www.cdc.gov/php/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html?CDC\\_AAref\\_Val=https://www.cdc.gov/php/publications/topic/hipaa.html](https://www.cdc.gov/php/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html?CDC_AAref_Val=https://www.cdc.gov/php/publications/topic/hipaa.html)
8. Naveen, "HIPAA Compliance Guide for Healthcare Chatbots," *kommunicate*. [Online]. Available: <https://www.kommunicate.io/blog/a-essential-guide-to-hipaa-compliance-in-health-care-chatbots/>

9. L. Eliot, "Here's How The FTC Might Lower The Boom On Those Emerging Generative AI Mental Health Therapy Chatbots That Are Promising Miracle Cures," *Forbes*, Jan. 2024, [Online]. Available: <https://www.forbes.com/sites/lanceeliot/2024/01/16/heres-how-the-ftc-might-lower-the-boom-on-those-emerging-generative-ai-mental-health-therapy-chatbots-that-are-promising-miracle-cures/>
10. R. Kelliher, "Can — and Should — Chatbots Help Students Navigate Mental Health Crises?," *Diverse*. [Online]. Available: <https://www.diverseeducation.com/students/article/15281699/can-and-should-chatbots-help-students-navigate-mental-health-crises>
11. C. Talbott, "Beware online mental health chatbots, specialists warn," *UW Medicine*, Mar. 2024, [Online]. Available: <https://newsroom.uw.edu/blog/beware-online-mental-health-chatbots-specialists-warn>
12. J. Li, "Security Implications of AI Chatbots in Health Care," *National Library of Medicine*, Nov. 2023. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10716748/>
13. "FTC Gives Final Approval to Order Banning BetterHelp from Sharing Sensitive Health Data for Advertising, Requiring It to Pay \$7.8 Million," *Federal Trade Commission*, Jul. 2023, [Online]. Available: <https://www.ftc.gov/news-events/news/press-releases/2023/07/ftc-gives-final-approval-order-banning-betterhelp-sharing-sensitive-health-data-advertising>
14. S. Huang *et al.*, "AI Technology panic—is AI Dependence Bad for Mental Health? A Cross-Lagged Panel Model and the Mediating Roles of Motivations for AI Use Among Adolescents," *National Library of Medicine*, Mar. 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10944174/>
15. "NEDA Suspends AI Chatbot for Giving Harmful Eating Disorder Advice," *Psychiatrist.com*, Jun. 2023, [Online]. Available: <https://www.psychiatrist.com/news/neda-suspends-ai-chatbot-for-giving-harmful-eating-disorder-advice/>
16. I. O'Sullivan, "Why AI Therapy Chatbots Are the Ultimate Ethical Dilemma," *tech.co*, Oct. 2023, [Online]. Available: <https://tech.co/news/ai-therapy-chatbots-ethical-risks>
17. "NEDA Suspends AI Chatbot for Giving Harmful Eating Disorder Advice," *Psychiatrist.com*, Jun. 2023, [Online]. Available:



<https://www.psychiatrist.com/news/neda-suspends-ai-chatbot-for-giving-harmful-eating-disorder-advice/>

18. M. N. G. van Wissen, "A Comparison Between ALICE and Elizabeth: Chatbot Technology Under the Microscope," University of Leeds, Leeds, UK, White Rose Research Online, 2004. [Online]. Available: <https://eprints.whiterose.ac.uk/81930/1/AComparisonBetweenAliceElizabeth.pdf>
19. J. Zhang, P. R. Kraft, and S. Lee, "The Role of Artificial Intelligence in Mental Health Care: Current Applications and Future Directions," *Journal of Mental Health and Clinical Psychology*, vol. 8, no. 4, pp. 1-15, 2024. doi: 10.1016/j.jmh.2024.07.003. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10242473/>
20. Sonia Health. "Sonia Health," <https://www.soniahealth.com>.
21. Elomia. "Elomia - AI-Therapist," <https://elomia.com>.
22. Wysa. "Wysa - AI for Mental Health," <https://www.wysa.com>.
23. Mindsum. "Mindsum - Mental Health Support for Children and Young People," <https://www.mindsum.org>.
24. Bloom: CBT Therapy & Journal, "Bloom: CBT Therapy & Journal on the App Store," <https://apps.apple.com/us/app/bloom-cbt-therapy-journal/id1475128511>.
25. Woebot Health. "Woebot Health - Your Mental Health Ally," <https://woebothealth.com>.
26. Youper. "Youper - Emotional Health Assistant," <https://www.youper.ai>.
27. Sintelly. "Sintelly - Personal Growth and Knowledge," <https://sintelly.com>.
28. Mindspa. "Mindspa - Mental Health & Wellbeing," <https://mindspa.me/en/>.
29. Kintsugi Health. "Kintsugi Health," <https://www.kintsugihealth.com>.
30. E. C. Stade, J. C. Eichstaedt, J. P. Kim, and S. W. Stirman, "Readiness for AI Deployment and Implementation (READI): A Proposed Framework for the Evaluation of AI-Mental Health Applications," *Preprint*, June 2024. [Online.] Available: <https://osf.io/preprints/psyarxiv/8zqhw>
31. American Psychological Association, "Ethical Principles of Psychologists and Code of Conduct," *American Psychological Association*, 2017. [Online]. Available: <https://www.apa.org/ethics/code>
32. White House Office of Science and Technology Policy, "Blueprint for an AI Bill of Rights," *The White House*, Oct. 2022. [Online]. Available: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>
33. National Academy of Medicine, "Artificial Intelligence in Health, Health Care, and Biomedical Science: An AI Code of Conduct - Principles and Commitments (Discussion Draft)," *National Academy of Medicine*, Feb. 2023.

[Online]. Available:

<https://nam.edu/artificial-intelligence-in-health-health-care-and-biomedical-science-an-ai-code-of-conduct-principles-and-commitments-discussion-draft/>

34. U.S. Department of Health and Human Services, "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research," *HHS.gov*, Apr. 1979. [Online]. Available: <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html>
35. American Psychiatric Association, "The App Evaluation Model," *American Psychiatric Association*, 2020. [Online]. Available: <https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/the-app-evaluation-model>
36. "Privacy policy," OpenAI, <https://openai.com/policies/row-privacy-policy> (accessed Aug. 16, 2024).
37. Soopra. "Soopra". <https://www.soopra.ai/>
38. TextCortex. "TextCortex - One AI copilot that truly gets you", <https://textcortex.com/>
39. Character.ai "Personalized AI for every moment of your day", <https://character.ai/>