American University of Central Asia

Software Engineering Program

Application of Artificial Intelligence in Mental Healthcare: Generative Pre-Trained Transformer 3 (GPT-3) and Cognitive Distortions

Submitted to Software Engineering Program

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Arts

at the

American University of Central Asia

May 2023

Author
Deniz Nazarova
Software Engineering Program

Certified by
Thesis Supervisor
Almaz Bakenov

Accepted by
Almaz Bakenov
Head of Software Engineering Program

ABSTRACT

This project explores the phenomenon of AI technology in therapy and aims to develop software that can be used in mitigating mental health issues among young people. The product of this research is TeaBot, an AI bot that applied GPT-3 models and uses methods of Cognitive Behavioral Therapy (CBT) to help users recognize and challenge distorted thoughts. Taking into account the limitations of GPT-3 in context understanding, technology was fine-tuned on the self-gathered dataset and all models were tested to find out the most accurate and cost-effective model. Curie model demonstrated the highest performance for the recognition task, and davinci showed the best results in generating a response to user's distortions. In addition, the bot was validated through the 8-week experiment on 68 AUCA students and interviews with practicing psychologists. Research findings revealed that TeaBot is an instrument that can be used for the prevention and intervention of mental disorders with the group assigned to communicate with the bot exhibiting statistically significant difference compared to the control group. Experts also shared positive views on the quality of TeaBot and were eager to implement it as a tool in their therapeutic practice. In addition to the existing version of the bot, a manual was developed to practice the application of the bot in therapy as well as inform the users who are new to counseling what therapeutic methods were utilized. The development process also included a manual for the future practice of augmented therapy. In conclusion, TeaBot continues to be one of the earliest known bots in Central Asia that applied OpenAI's models and were tested using ethical research methods.

Keywords: artificial intelligence, chatbot, GPT-3, mental health, CBT, students, youth

Contents

ACKNOWLEDGEMENTS

This project was inspired by Sweaty Machines' song "Blue Jeans and Bloody Tears", a Eurovision song created by AI. I would like to thank the creators for raising the thought about whether we can make an AI depressed based on the data it was fed.

No gratitude would be enough to thank all faculty and supervisors that participated in reviewing this big and long project. I would like to recognize the invaluable contributions of Google, Stack Overflow, and programmers who analyzed the code and suggested optimization methods.

This project was generously supported by the Center for Civic Engagement at the American University of Central Asia.

With love and gratitude to my partner.

CHAPTER 1: AN INTRODUCTION AND OVERVIEW

The domain of Artificial Intelligence capabilities has been expanding dramatically for several decades. Nowadays, it is unattainable to encounter a field that was not affected by the presence of new technology. Modern AI algorithms can analyze enormous amounts of data to suggest what to watch over dinner. Some models already excel at games created by humans and win world-class champions in a variety of games – from chess to StarCraft (Team, 2019; Haenlein & Kaplan, 2019). The programs developed during the MegaFace Challenge, a competition between developers on the topic of face recognition, demonstrated extremely high rates that exceeded the human ones (Kemelmacher-Shlizerman et al., 2016). Interestingly, the current progress made in this branch of science let GitHub implement GitHub co-pilot, a code that can code (Yetiştiren et al., 2022).

While all these achievements remain impressive, the question "Can machines think?" once asked by Turing is still unanswered (Turing, 1950, p. 435). His hypothesis was that at some point, technology might persuade a human that it is another human communicating to them, i.e., winning the imitation game (Turing, 1950). These discussions led to the definition of AI as a machine simulating human behavior (McCarthy et al., 2006). This implies that our discourse about technology is often shaped from the perspective of how well computers can persuade us of their 'humanity'. One of the branches of AI studies that was strongly inspired by this approach is Natural Language Processing, or NLP, a branch of AI science studying the engineering of algorithms that would allow a computer to better understand and generate human language (Luxton, 2016). Language being one of the most definitive aspects of human intelligence is still hard to apply for contemporary innovations. Nevertheless, recent discoveries might surprise the skeptics as OpenAI developed generations of Generative Pre-trained Transformers, or GPTs, language models that are

famous for the size of the data they used for analysis and a wide range of tasks from translation to text generation (OpenAI, n.d.-a). The recent success of the newest sub-model called ChatGPT, a conversational AI that is able to produce human-like responses, raises new topics for research in human-computer interaction (OpenAI, 2022).

The ability of AI to form a conversation can be applied in multiple spheres, far beyond mere chit-chat. There is a growing interest in implementing AI in psychotherapy due to its accessibility, affordability, scalability, and security in terms of preventing existing stigma. These factors became beneficial after the COVID-19 outbreak worsened mental health globally and demonstrated the existing gap in healthcare (World Health Organization [WHO], 2022). Young people, especially students, were strongly affected on a multifactorial basis inducing financial issues, trouble with focus, and feeling of being alienated due to the lockdown (Active Minds, 2020; WHO, 2022). Besides, Kretzschmar et al. (2019) report that 75% of mental health diseases tend to form around the period of young adulthood, and mitigating them might decrease the long-term economic and personal consequences. Meanwhile, in Kyrgyzstan, the global pandemic affected the youths' mental health in the rising number of suicides (UNICEF, 2020). This sets an urgent need to enhance the current mental health facilities available for students. However, the current state of healthcare in Kyrgyzstan does not inspire hope as even before the pandemic mental health field did not meet 87% of yearly plans (Orlova, 2019). Pinchuk et al. (2021) also emphasize the neglected status of mental healthcare in Kyrgyzstan by stating such problems as "insufficient financing, lack of equipment, underregulation, a deficiency of qualified specialists" (p. 4). Overall, there is a clear lack of resources to work with the consequences of the pandemic, recession, and conflicts on the population's mental well-being.

Living in a post-pandemic world might be challenging for one's mental well-being too. The recent recession is leading to people not being able to get the necessary help due to financial

difficulties. Considering these circumstances, implementing AI-chatbot might be even more effective in developing countries due to their cheapness compared with in-person therapy (Abd-alrazaq et al., 2019). Aside from the monetary struggles, there is growing political instability in the post-soviet region that requires fast mental support for the affected civilians. During the border conflict in Batken, Kyrgyzstan, around 140 thousand people had to flee to other regions, and the shock received during the event required both short-term and long-term help from professional psychologists (gov.kg, 2022; Larionov, 2021). In the Russia-Ukraine war, the need for emergency help is often highlighted (Kucenko, 2022). Due to the focus on urgency and the ability to help such a large number of people, AI technology might reduce mental unrest effectively. However, the phenomenon of AI-supported mental health remains unexplored in Central Asia, thus the main purpose of this research is to develop a therapeutic AI chatbot in Kyrgyzstan, measure its efficiency by conducting an experiment on the sample of the target audience and discussing the potential for augmenting existing in-person therapy to ensure the supervision of the new technology while also preparing people for the new epoch in mental healthcare.

Structure of the Paper

Chapter 2 of the paper focuses on the review of the existing literature including the history of the application of AI in psychotherapy, its advantages and disadvantages while also identifying gaps in the research that this project aims to fulfill. Chapter 3 explains the methodology used with a detailed description of procedures implemented for therapeutical aspects of the bot and its interface; this section also informs about the processes that took place for the evaluation of the effectiveness of the bot. Chapter 4 states the results of the project both from the software and user sides. Chapter 5 suggests further work and recommendations for improving mental healthcare in Kyrgyzstan by applying an augmented therapy approach.

CHAPTER 2: LITERATURE REVIEW

GPT: Under the Hood of Hype

Generative Pre-trained Transformer 3, or GPT-3, might be one of the most discussed topics in technology right now. The recent variation of the model called ChatGPT has already inspired multiple news reports that predict both wonders and nightmares about the future (Shankland, 2023). This excitement might be explained by several discoveries that this invention brought with its launch which this segment explains by deeply analyzing the given name and context. First, the word "transformer" refers to the neural network architecture that overperforms the preexisting neural networks in terms of parallelization and training time while also demonstrating efficiency in terms of context understanding (Vaswani et al., 2017). Besides that, the application of transformers led to the expansion of the functionality where the increase of parameters allowed to "capture polysemous disambiguation, lexical and syntactic structures, as well as factual knowledge from the text" (Han et al., 2021, p.226). "Pre-trained" is an innovative two-step learning process that allowed to train the model on a large amount of unlabelled data to identify linguistic patterns and then fine-tuned, e.g., trained on smaller task-specific data (Han et al., 2021). Combining unsupervised and supervised approaches, GPT-3 is capable of conducting a variety of tasks from creating poetry to writing code without overwhelmingly large datasets of labeled data which is a common roadblock in the Natural Language Processing (NLP) field of AI research (Dale, 2021; Han et al., 2021). "Generative" is defined as the model that can predict, i.e., generate new tokens based on the prompt given (Google, 2022). The number three stands for the version of the model that considerably exceeds the power of the ancestors, with GPT-3 trained on the 45 TB of data gathered from books and the Internet and enlarging the number of parameters to 175 billion making it "the largest language model ever created" (Zhang & Li, 2021, p. 832).

Despite such impressive numbers, GPT-3 has limitations that might lower enthusiasm about its language comprehension skills. Often model's outputs are criticized for turning into gibberish after a certain word threshold (Dale, 2021). This might be explained as GPT-3 does not understand the language in the way a human being would, rather it remembers how one is statistically likely to reply (Floridi & Chiriatti, 2020; Zhang & Li, 2021). The quality of output is very data-dependant which implies that any errors and biases in the data might significantly harm the end result despite the computational power (Zhang & Li, 2021). In addition, there are certain dangers of misuse from the user side, e.g., the recent incident with users being able to control the information ChatGPT delivers without having access to the code by simply writing a cunning prompt (Floridi & Chiriatti, 2020). This suggests that although the model revolutionized the perception of AI, it is still raw and requires a lot of extra work from the developer's side.

Application of AI in Psychotherapy through Time and Tpace

While the talent for computations was not a big surprise, the success of computers in communications was groundbreaking when the research on ELIZA was published. ELIZA is the first chatbot that applied therapeutic techniques when communicating with people (Weizenbaum, 1966). Ironically, the main developer did not expect ELIZA to succeed in therapy, however, participants often commented on the efficiency of the chatbot (Zeavin & Peters, 2021). Weizenbaum (1966) programmed ELIZA to be based on the Rogerian approach which allowed her to avoid the need for explicit knowledge about the outside world. Although being heavily scripted made its effectiveness quite sensitive to the input, this still made some of the participants believe that she was a human (Weizenbaum, 1966). ELIZA became a foundation for the later AI chatbots with some of her successors being constructed using more sophisticated design.

One of the modern tendencies became implementing Machine Learning in chatbot's architecture. Machine Learning, or ML, is defined as AI models that are capable of acting intelligently based on the training data their algorithm processed (Samuel, 2000). In the context of communication, the application of ML algorithms strongly relies on the dataset in order to build a more natural communication that requires Big Data, or extremely large yet unstructured information that needs interpretation (Oracle, n.d.). Using ML leads to the more sentient technology that is able to reflect on the previous sentences and build its own sentences that are not based on heavy coding while Big Data gives the machine possibility for diversification (Abd-alrazaq et al., 2019). The models developed by this tandem are very precise in recognizing the patterns in thinking which made them the best fit for Cognitive Behavioral Therapy. One of the most successful AI applications that combined the aforementioned design might be Woebot, a chatbot developed in the United States to treat symptoms of depression and anxiety. It is capable of learning from the data given by the user while applying different therapeutic methods starting from mindful exercises to therapeutic journaling (Fitzpatrick et al., 2017). Similar in popularity products that also passed clinical tests successfully include bots like Tess and Wysa, also developed in the US (Fulmer et al., 2018; Inkster et al., 2018). At the moment of writing, Koko bot, one of the most recently developed mental health chatbots, openly declared using a novel ChatGPT AI model to help people deal with their unrest, however, the lack of empirical and ethical research to validate the technology led to the huge critique of the application (Xiang, 2023).

Although chatbots remain mostly a Western feature of mental health, there are also a few products developed in other parts of the world. One of the most prominent examples might be Elomia, a mental health AI bot that operates in English, that was developed in Ukraine and clinically tested (Romanovskyi et al., 2021). Because of the current war in Ukraine, it is not clear how the development process of the bot will go further. However, the market in other post-soviet countries

stays lacking appropriate research on efficiency. Russian developers recently announced the launch of SabinaAI without publishing any data openly on its testing process except for the mention that it was developed in cooperation with psychologists (AvatarMachine, LLC, n.d.). Kyrgyzstan is currently at the beginning stage of digital mental health with products like a game about bride kidnapping and scripted telegram bots informing on sexual education and well-being (*Oilobot*, n.d.; Open Line, 2020). Although these products usually cause a lot of interest from the media news, there is also no clinical research conducted on the efficiency of these innovations.

AI as Support

The current research agrees that there are several advantages of technology compared to human specialists. First of all, mental health applications remain more accessible source of support as it requires only a device with an Internet connection (Kretzschmar et al., 2019). Moreover, unlike human therapists, chatbots are available 24/7 which makes them especially helpful in situations when a psychotherapist is not available to answer, e.g., during other sessions or at nighttime. This might be especially helpful when there is a lack of human psychotherapists available. For example, Abd-alrazaq et al. (2019) stress that in developing countries chatbots can be an additional help as there is a severe disbalance in the specialist-to-patient ratio. Due to the high demand for mental health services, the unavailability of specialists and putting people on long waiting lists might be alleviated by delivering an e-support via chatbots (Kretzschmar et al., 2019).

The scalability advantage of technology might be especially helpful considering the aforementioned lack of resources. Torous et al. (2020) inform that unless the patient is against the application of telehealth, “hybrid solutions that offer a blend of face-to-face and online or app-based treatment will be the most effective solution” (p. 2). Moreover, this flexibility of the format where the participation of technology is easily regulated from solely human interactions to the integration

of different applications can significantly enrich the repertoire of therapeutic services (Torous et al., 2020). For example, the implementation of bots as data gatherers can assist practitioners in better preparation for the first therapy session (D'Alfonso et al., 2017; Damij & Bhattacharya, 2022). Luxton (2016), on the other hand, suggests that AI technology can be used to both reduce the existing pressure on the healthcare system and reduce costs by implementing "the stepped-care approach" that distributes mental health services according to the patient's needs, e.g., "care seekers can take self-assessments with a virtual care provider and be transferred, if necessary, to full therapy with a human care provider" (p. 17). These cases demonstrate that technology has a great potential to enhance the quality of therapy in a variety of ways depending on the level of involvement a specialist would like to apply. However, as the study by Torous et al. (2020) claims there is a vital need to educate professionals on the opportunities this tandem provides.

The next feature of using chatbots in mental healthcare is their affordability. Being a low-cost or even free mental health service, technology often removes the existing barriers that people experience when reaching out for help (Fitzpatrick et al., 2017; Gabrielli et al., 2021; Watts et al., 2013). Luxton (2016) highlights that producing a mental health specialist is more-time consuming compared to software that can be replicated quickly. This implies that chatbots are also less costly in terms of production and deployment compared to human specialists. In the study by Fulmer et al. (2018), the cost-efficiency of computer-assisted therapy is especially emphasized considering that this type of intervention is less-intensive than an in-person one. Taking into account the financial struggles that students might experience in their new stage of life, it is also crucial to provide them with evidence-based sources of self-care. Grové (2021) stresses that besides clinical intervention, chatbots might be used as a tool to deliver free and high-quality advocacy for youth about their mental health. Indeed, when implementing technology into this field, technology might be especially

beneficial due to low-cost and less time-intensive which might be especially crucial for young people who struggle with both.

Chatbots as well as other non-person mental health services can also be gateway support for people who struggle to start therapy because of the existing stigma (Abd-alrazaq et al., 2019; Fitzpatrick et al., 2017). According to Boucher et al. (2021), this very stigma might be one of the leading causes for young adults avoiding therapy in the first place. Due to the anonymity that chatbots provide, patients might feel more secure and more eager to share information that they might hide in traditional therapeutical sessions (Fulmer et al., 2018; Kretzschmar et al., 2019). Another reason why artificial companions might be so stigma-reducing is a perception of them being non-judgmental (Kumar et al., 2022). Thus, chatbots might be an especially helpful treatment in Asia where people with mental health issues face discrimination as society often sees them as dangerous and needed to be isolated (Lauber & Rössler, 2007; Sulaiman et al., 2022). In the study done by Pinchuk et al. (2021), stigma is perceived as one of the crucial reasons why people avoid treatment.

In general, evaluating the benefits of implementing chatbots in mental healthcare, it is evident that they display qualities that are especially needed considering the current crisis in service provision. While the inclusion might have certain disadvantages that would be described in the next segment, the strengths demonstrate that this technology might serve as a good starting point for a lot of people who cannot afford to attend therapy fully or are not ready due to the existing assumptions about individuals with mental issues.

AI as a Threat

As with any intervention, the application of technology for treating mental health issues has negative outcomes as well that should be taken into consideration. One of the most mentioned in the

study remains the problem of bots being primitive and limited in terms of context understanding (Vilaza & McCashin, 2021; Yang, 2020). This might be especially problematic as chatbots can repeat the same bias that already exists in therapy. For example, Brown (2018) criticizes traditional therapy for being not suitable for the problems of queer people, ethnical minorities, people of color, and women who are especially vulnerable when using therapy that is not focused on recognizing the social injustice and reinforces the stigma. If a bot is designed without knowledge of this, then it will be able to see all cases shared by a client as irrational, this might make the client adjust to the traumatizing environment. An infamous example of this is the case of Woebot replying to the 12-year-old boy reporting the sexual harassment with "Sorry you're going through this, but it also shows me how much you care about connexion and that's really kind of beautiful" (Vilaza & McCashin, 2021, p.2). Clearly, this context blindness is a significant issue that undermines the unsupervised usage of any chatbot.

When implementing artificial intelligence in chatbots, it is very crucial to check the data used for the training. AI in such technology allows to avoid dry answers compared to the hard scripted programs, however, this requires a large dataset to be applied to guarantee the variance of response and understanding of the user's message (Luxton, 2016). In order to generate an answer that is both similar yet unique for each user, the machine is fed a dataset that contains possible inputs and outputs (Zhang & Li, 2021). This requires the application of Big Data which helps the machine-generated speech look more sentient but often lacks diversity in the sample despite the size which results in replication of the existing biases (Bender et al., 2021). For example, GPT-3, or third-generation Generative Pre-trained Transformer, is a model that is claimed to be trained on the whole Internet often criticized for gathering data mainly from white and male-dominated spaces (Bender et al., 2021). This suggests that although modern AI programs are trained on data that exceed human understanding when designing a bot, there is a need for data to be checked.

Despite chatbots being cheaper than traditional therapy, they remain a product of a large business industry that might aim to gain benefits rather than help a person in need. Taking into account the aforementioned cost advantage of this technology, it is often explained by the data being sold to third-party companies (Gratzer & Goldbloom, 2020). Other bots are designed in a new way to extort money from people by abusing their loneliness. Depounti et al. (2022) view Replika as one of the most extreme examples of how commercial interest might lead to psychological harm to clients due to its advertisement as an AI friend to improve mental health, yet often suggesting upgrades that include sexual content.

Considering the abovementioned limitations, chatbots regardless of whether they are heavy-scripted or apply Artificial Intelligence continue to be non-sufficient for becoming a fully independent mental health intervention. However, research suggests that this should not be used as a point for abandonment of the idea but rather investigate the options for the integrated therapy, i.e., an approach that includes both therapist and bot where drawbacks of one technique are mitigated by the pros of another (Luxton, 2016).

CHAPTER 3: PROCESS AND METHODS

The Present Study

The application of AI in psychotherapy remains quite an ambivalent topic. The current ethical disputes require developers to be extra careful when designing a program considering both the efficiency and safety aspects of the product. The first one can be achieved by integrating stronger technologies, for example, Abd-alrazaq et al. (2019) report that only 7.5% of AI chatbots use a machine-learning approach and suggest implementing it more in the future (p. 5). D'Alfonso et al., (2017) persist in the incorporation of machine learning to ensure the eagerness of users to share information with a bot. If realized, this technology will boost the abilities of the bot as well as increase the resemblance of the conversation with the human speech considering that data used for training is carefully gathered. As GPT-3 is a very novel invention in the field, there is a lack of research on how well it can be incorporated into mental healthcare as well as therapeutic programs based on it. At the same time, considering the aforementioned problems with complex context understanding, GPT-3 requires a special design to experience the high quality of services with OpenAI (2021) openly stating its interest in the implementation of the technology in healthcare.

Regarding the safety considerations, Gratzer and Goldbloom (2020) advise that bot engineers should be also mindful of the ethical principles to prevent data leaks and explicitly inform the user about the way how their data is used in a concise manner. It is strongly recommended to invite specialists and discuss with them the process of bot development as well as conduct testing with them due to some bots being developed by people outside of the mental health field (Gratzer & Goldbloom, 2020). AI technology that aims to work with human subjects needs to go through clinical testing to ensure its validity that, as an intervention, it does not bring any harm. While there is proof of the efficiency of mental health chatbots in the short term, they are usually tested for 2

weeks which is shorter than the expected duration of the CBT course (Fitzpatrick et al., 2017; Gabrielli et al., 2021). Gabrielli et al. (2021) also highlight that the studies need to include a control group to compare the change in behavior with the baseline. Additionally, most of the research is conducted in the US and using a privileged minority as a sample, current findings cannot be claimed as universal and need to be tested on people outside of that scope to diversify findings (Abd-alrazaq et al., 2019; Macleod et al., 2020). Therapy approaches are not ever-changing and need to be tuned according to the region.

Another problem is the common intention to prove the independency of AI rather than exploring the phenomenon of integrated therapy and how it can be built. Although products warn that they are not a substitution for a mental health specialist, they rarely suggest other sources of help or suggest how a program can be supervised. Elomia is one of few examples of software that suggested using itself as a self-reflection tool between actual therapeutic sessions (Romanovskyi et al., 2022).

This project aims to close the gap found in the research by experimenting with the application of large AI language models like GPT-3 in psychotherapy, overcoming the limits of this technology, developing a psychotherapeutic chatbot that is applicable for augmented therapy in Kyrgyzstan, and estimating its efficiency according to the requirements of modern psychological research. Moreover, the work includes the creation of a manual that can be further used by clinicians to practice integrated therapy to delegate some of the therapeutic tasks to a bot. Overall, this project shows originality by being the first (of known) interdisciplinary research on capabilities of GPT-3 in post-soviet space while complying with the academic standards. As the end product, this research offers specialists a ready-to-use tool for integrated therapy that might be one of the key services for mental health services to become more affordable and accessible for the youth.

Methods

Before starting coding, one needs to decide on the evidence-based therapeutic approach to build a bot on. Cognitive behavioral therapy was used as an inspiration for the bot as it focuses on changing a client's well-being by learning to recognize and challenge negative thoughts and behaviors (Beck, 1995). These tasks are always practiced as a part of the homework that the patient is expected to practice (Beck, 1995). CBT is also very structured and requires psychoeducational elements to be present, and this allows to imagine how technology might fit into the therapeutic agenda. Implementing such activities would require an AI that is capable of recognizing natural (human) language on a very high level, thus, applying a neural network. This is a machine-learning model inspired by the human brain that employs neurons, i.e., interconnected nodes, to detect patterns and correlations in data, while also learning from examples (Luxton, 2016). Following the literature review, the bot was determined to be executing specific tasks only and serve as a CBT exercise rather than a friend-interlocutor. The bot was expected to perform the following tasks: educate and remind the patient about certain aspects of CBT, recognize healthy thoughts from unhealthy ones, classify the detected distortion, identify the most appropriate common question to challenge thinking, and create the most context-fitting version of it. Meanwhile, all of these should be hosted in a user-friendly environment, like an app or website. The first two tasks were determined to be written by the principal researcher and directly incorporated in the user side rather than designed by AI due to the ethical complications that might arise in case of the model's inability to explain complex notions to the patient. Thus, the process of bot engineering consisted of several stages that required different technologies. The first segment describes the aspects of integrating artificial intelligence part in the bot's conversational logic while the second one indicates how it was incorporated into the user interface and also informs about additional features that were added for the validity research and user convenience. The final part is dedicated to the procedures the first

version has undergone to ensure that the participants are well-informed and safe. Figure 1 shows the graphic overview of the process.

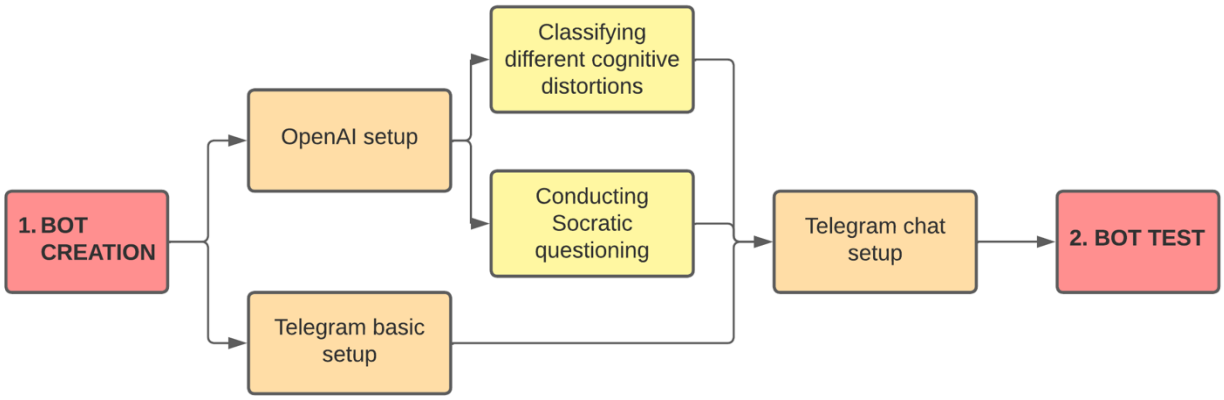

Figure 1. *Bot's Development Scheme*

Back logic: GPT

The end goal of AI-part was classifying the thought, identifying the right base question, and creating the response to the user. The first engineering task was to choose to either experiment with the existing models or build one from scratch. Developing own neural network that specializes in natural language understanding requires tremendous material and temporal resources (Brown et al., 2020; Zhang & Li, 2021). In-use models, on the other hand, are often cheap and accessible. OpenAI created Generative Pre-trained Transformer 3, or GPT-3, a language model that can be used by anyone for a considerably small amount of money and does not require significant computational power (Han et al., 2021). One of the strongest sides of GPT-3 is being trained on 45 TB of Internet and book data making it "the largest language model ever created" (Zhang & Li, 2021, p. 832). This allows technology to generate texts that are very diverse and human-like while requiring much less additional training from developers. Nevertheless, even these advantages did not fully remove the common weakness between language models such as primitive context understanding and

generating gibberish from time to time (Dale, 2021; Floridi & Chiriatti, 2020). Moreover, GPT-3 is often criticized for its dependency on the meticulously designed prompt, the frailty that was already documented as a way to hack the most recent version, ChatGPT (Floridi & Chiriatti, 2020;). Zhang and Li (2021) see these drawbacks as a demonstration that "[GPT-3's] application value is primarily reflected in intelligent auxiliary tasks, and it cannot directly interface with the end-user" suggesting further supervision (p. 833). Despite these limitations, the language models developed by OpenAI persists in being the strongest in the market, especially when compared to the ones crafted by enthusiasts with limited sources. Consequently, OpenAI products were chosen as candidates for a base for AI logic.

On the other hand, these models are not designed to be a CBT psychologist, thus, the mere pasting of the problem often resulted in misunderstanding and giving advice (see Figure 2 for more details). The latter is avoided in therapy as this often leads to developing reliance on external resources which is the opposite of goal therapy pursuits (Corey et al., 2011). Hence, even in some cases when GPT-3 might generate good advice, this approach cannot be considered therapeutic. Moreover, what is more dangerous is that because GPT-3 is not sentient and often judged for its lack of context understanding, it often fails to detect subtle messaging. For example, while a therapist might understand that a client demonstrated minimizing distortion, i.e., "when people devalue their positive role in the event that happened to them", as the phrase conveys the meaning that one's achievement of getting a job was *just* luckiness (Covin et al., 2011, p. 321). Both GPT-3 and ChatGPT, do not track any distorted thoughts. This might be explained either by the features of the training process (GPT models are trained mostly on unlabelled data, and develops their own patterns of language recognizing) or mere complexity of human emotions and metamessaging that are not available for language models at the moment of writing. In this case, the danger is that this pattern of communication might harm by reinforcing the already existing negative self-talk, the very

opposite of what a CBT specialist would do. Henceforth, unsupervised usage of GPT-3 or ChatGPT is strongly unadvised. In addition, both systems partially demonstrate “know-it-all” bias assuming that the answer is present in the data (Heikkilä, 2022). Developers might constraints the model manually, e.g., by disabling ChatGPT’s ability to comment on politics, not only these cases have been already hacked, but this also leaves areas of GPT’s expertise that might not be properly assessed, like mental health advising that GPT-3 and ChatGPT do uncontrollably (Oremus, 2023). To eliminate the already described drawbacks, it was decided to limit the model’s agency by applying fine-tuning, i.e., adding dataset to specify the task and add specific patterns of questions that AI is allowed to use. This feature would allow OpenAI’s model to differentiate between the different distortions. However, among the recently developed OpenAI products, only GPT-3 enables this function (OpenAI, n.d.-b). At the moment of writing, there was no official statement on when ChatGPT’s and GPT-4’s APIs would be available for fine-tuning making GPT-3 the only model that can be used for ethical development. Interestingly, this approach does not affect the model’s ability to create human-like content yet narrows the model’s understanding of the context.

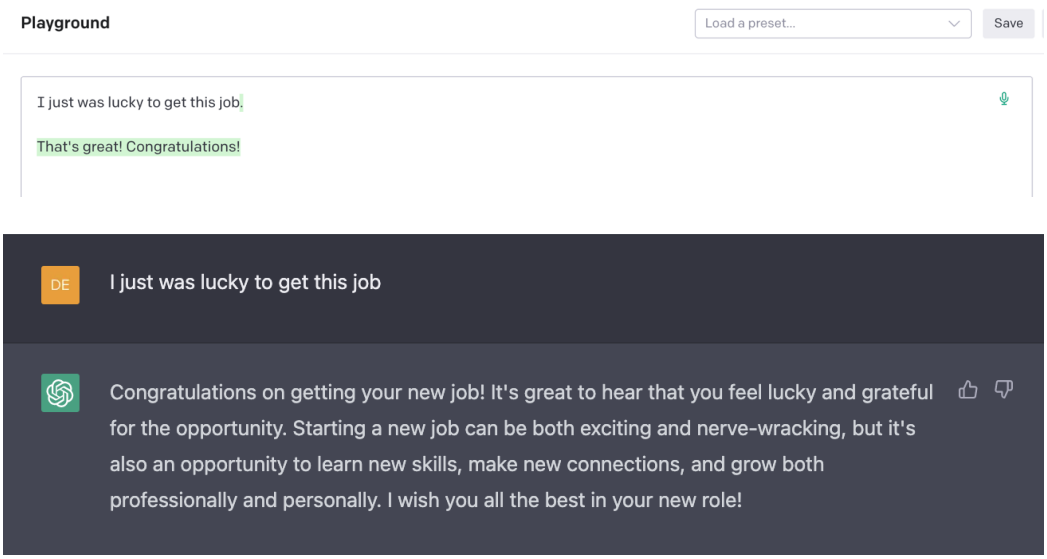

Figure 2. *Two Images illustrate how GPT-3 (above) and ChatGPT (below) react to a distorted thought.*

In CBT, cognitive distortions are described as errors in reasoning (Beck, 1963). Burns and Beck (1999) identified 10 types of these, and a bot needs to successfully differentiate between them as each type requires a specific approach to work with. GPT's fine-tuning option allows the creation of a new sub-model that is specifically trained for classifying distortions. To complete it, a developer is required to create a labelled dataset, e.g., "thought" – "type of distortion" (see Figure 3 for a sample). In addition, to reduce the know-ot-all effect of GPT-3, the 'unclassified' category was added in which case, the program would openly say to the user about the limitations of the software and suggest better ways to phrase the thought.

```
{"prompt": "I was lucky I got this job.", "completion" : "Minimizing"}
{"prompt": "They said that I am looking nice but they are just nice to me.", "completion" : "Minimizing"}
{"prompt": "I passed this exam because it was just too easy.", "completion" : "Minimizing"}
{"prompt": "Everyone can do that.", "completion" : "Minimizing"}
{"prompt": "Everyone has the qualities that I have, there's nothing special about them.", "completion" : "Minimizing"}
{"prompt": "I'm not as talented as people think, I was there at the right time and in the right place.", "completion" : "Minimizing"}
{"prompt": "My life and my achievements are nothing compared to other successful people.", "completion" : "Minimizing"}
{"prompt": "I don't think that I'm smart, it's just common sense.", "completion" : "Minimizing"}
```

Figure 3. *Excerpts from TeaBot's Training Set*

The data collection for this labeled dataset was gathered by a combination of examples from CBT literature and anonymous submissions made by the principal investigator and AUCA psychology students who got access to the document by a link that was shared in student group chats. This sample was chosen due to them both being part of the population that will go through the experiment and study the CBT concepts, thus, knowing how to label data. They were asked to write 10 cognitive distortions in total. After the completion of a dataset, the whole dataset was checked and edited by a principal investigator, the project's supervisor, and a practicing CBT psychologist. In total, 240 examples of cognitive distortions were accumulated and divided into training and test sets in a ratio of 3 to 1.

After finishing the dataset, a version of GPT-3 was chosen by comparing the results of the test dataset classification which is needed for balancing between the price and quality of the output. A more detailed review of the model's performances might be found in Chapter 4. The process of

training requires a different amount of time based on the model chosen. As the output, the fine-tuning offers the model's ID that can be activated by a personal OpenAI token. More details can be found in Figure 4.

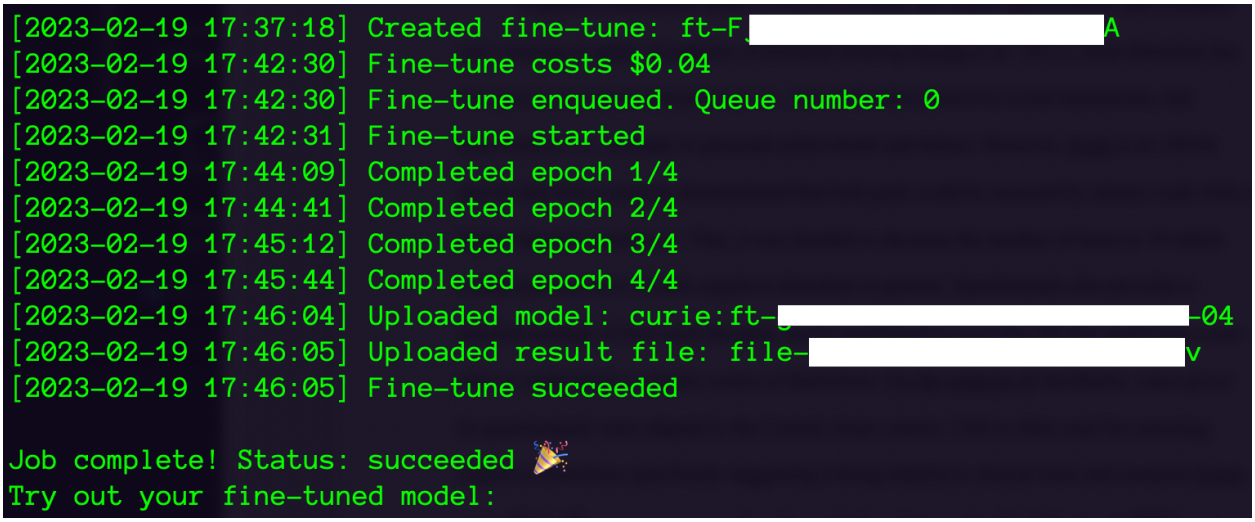

Figure 4. *GPT-3 Fine-tuning Process*

Note: The screenshot above describes fine-tuning of the curie version. Spots were marked for security reasons

The next stage is applying Socratic questioning, or guided discovery, a process of asking open-ended questions that stimulate the process of patient's reflection and critical examination of own beliefs (Beck, 1995). Each distortion represents different unhealthy tendencies in thinking which might require a variety of approaches. For instance, a person says that his friend does not care about him anymore. This is a mindreading, or a tendency to "assume that others are thinking negatively about them even though the other person has not said anything negative" (Covin et al., 2011, p. 317). This type of distortions is often battled with evidence questions that would ask a client to try to come up with evidence against this assumption (Beck, 1995). However, some questions might be less relevant for this case while being very helpful for others. Therapists often use defining questions when working with labeling, or when people attribute their one action to their

whole personality (Beck, 1995). Digging into the semantics of the used word might be helpful when a person states that they are not smart because they did not get a good grade, nevertheless, in the case of mindreading might not be effective. Consequently, identifying distortion does not solely solve the issue, but rather only the first step. Due to GPT-3 limitations, instead of asking the questions from scratch, a separate list of base questions, i.e., appropriate to use when challenging particular distortion, was created. Whenever a model recognizes a distorted thought, it also accesses that list and chooses the most appropriate one to challenge the thought. When a question is chosen, a model is asked to come up with a version of a base question that is most appropriate to the problem user described. This was achieved by using the completion function of GPT-3 and all language versions were also tested on the ability to ask a question that triggers the reflection process most eloquently. The AI was given a prompt that asked to compare the user input and ask a question using the base question as an example. The outcome of the whole GPT-3 process is the user receiving both the category of cognitive distortion that the program caught and a question that helps to challenge the thought (see Figure 5).

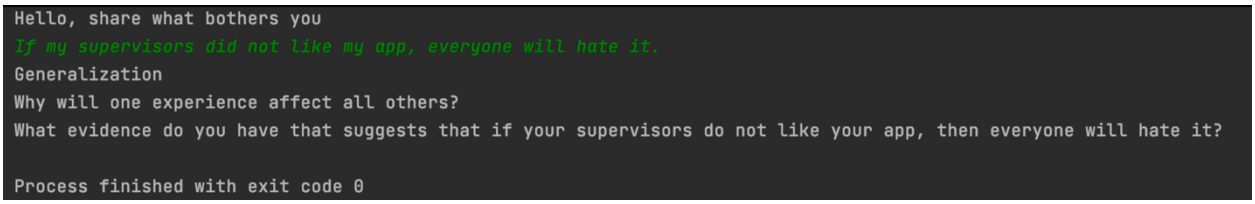

Figure 5. *Fine-tuned and Modified GPT-3's Response in Terminal*

User-Interface: Telegram

After the core AI logic was assembled, it needed to be wrapped in a user-friendly interface. This requires a decision on the format of the application, however, there is no unambiguous answer to this. Standalone applications give the advantage of being able to send user notifications that are often reported by users as an uplifting feature that eases the bonding process with the bot

(Fitzpatrick et al., 2017). Adamopoulou and Moussiades (2020) report that websites can be seen as more liberating ways of building a program with developers not limited by the policy of host application. However, both mediums require significant data protection as any case of private information leak would be damaging to users (Adamopoulou and Moussiades, 2020). An alternative to that might be an integration of a chatbot into the environment of an existing messenger. Kretzschmar et al. (2019) disclose that this approach might help with engagement and adjustment to technology too as they already exist in a familiar environment. Taking into account Telegram's security measures and friendly environment for chatbot implementation, it was chosen as a platform (Albrecht et al., 2022).

The python-telegram-bot library was used to access Telegram Bot API due to its rich documentation, asynchronous programming, and customizability (Toledo, n.d.). The library is well-explained for those who only started working with bot engineering and addresses all the issues one might face during the development. One of the most unique aspects of using python-telegram-bot was the ability to use handlers. This feature eased the process of integrating assessment in the bot and organized the flow of the user experience in more naturally. Asynchronous programming allowed to work with multiple requests at the same time. From the UI perspective, working with this library gives developers flexibility and allows to easily set up the necessary UI requirements while still having control over the Telegram environment.

The bot was given the name TeaBot. There is a long tendency for artificial interlocutors to have female names and perform female behavior during communication (Depounti et al., 2022; Fulmer et al., 2018; Gabrielli et al., 2021; Weizenbaum, 1966). Depounti et al. (2022) criticize this approach for reinforcing the existing bias of serving and caring role that woman is expected to have in society. Fitzpatrick et al. (2017) inform that even a specifically chosen robotic name does not interfere with the human tendency to build empathy towards nonhuman agents. Nevertheless, to

avoid any reproduction of existing gender stereotypes and reduce the effect of association with a human being on results, the TeaBot name was picked to remind a client that they communicate with a non-human. In addition, this name resembles the word “teapot” which can attract certain users due to the intended pun.

It was also decided to add a profile picture for TeaBot as this feature often allows to feel users more comfortable talking to a bot compared to the absence of it or a company logo (Palosaari, 2022). Kim et al. (2021) notify that there is no causality between the level of anthropomorphism of the bot’s profile image and user engagement. Combined with the previous literature on the effects of bots exhibiting human-like appearance, TeaBot’s profile picture was designed to be mascot-like reminding users that they contacted the bot. Nevertheless, to avoid blending with other bot-like profile images, the teapot wordplay was applied again, and the final version of the TeaBot profile picture exhibit a teapot in pixel aesthetic to connect it to its digital origin (see Figure 6).

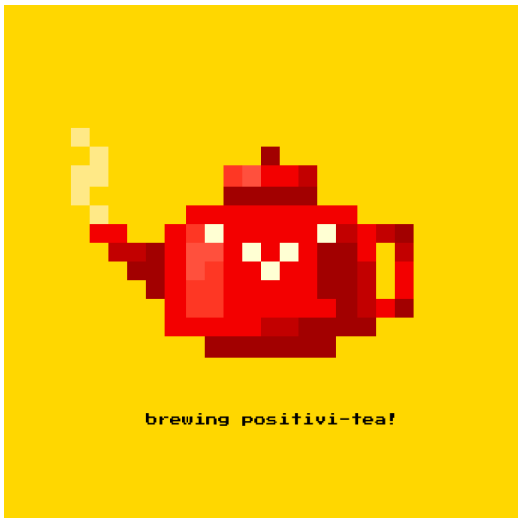

Figure 6. *TeaBot’s Mascot*

Regarding the communication style, a bot does not intend to sound machine-like while still admitting that it is not a human being communicating with a user. For example, when a bot is asked to communicate about things it is not trained to work with, TeaBot states, “I am sorry, I can work

only with distorted thoughts for now. Yet who knows how I will be able to help in the future!?". First, this tone was applied to avoid reinforcing any gender stereotypes that can affect how a user perceives the interaction (Depounti et al., 2022). Unintuitively, this clear presentation of TeaBot as a bot reduces the risk of users being not eager to interact with a bot as Kim et al. (2021) highlight that a high level of anthropomorphism can actually backfire on the bot if it does not self-disclose. At this stage of development, TeaBot is not planned to express or share any feelings, therefore, it was decided to maintain the friendly robot attitude.

Before Deployment

For TeaBot to be able to participate in the experiment to test its validity big sample of human subjects, there are several requirements that it needs to pass.

Due to the interdisciplinarity of the project, the pre-deployment review was done by experts in both psychology and software engineering. Two practicing psychologists were interviewed to gather feedback on the bot's work. In general, both experts expressed their satisfaction with the overall work of the bot, although the first tester expressed their concerns about the bot not always guessing the distortions correctly and suggested emphasizing that the bot is in its test mode. The second participant, on the other hand, was very impressed and enthusiastic about future applications of the bot and it being a probable addition to CBT for young people. Other comments from both experts mostly focused on the improvement of the communication of the bot, e.g., adding a tutorial to explain how the bot works, implementing more empathetic language to the bot, and reducing the number of items in the questionnaire.

After this, necessary changes were applied and interviews with two software engineers were conducted to discuss what issues might arise after deployment and how they can be prevented. The first was a back-end developer who focused on hosting the bot on the cloud server and suggested

options for more convenient data collection and analysis. The feedback from another programmer specialized in bot development included applying more Telegram features, e.g., pinning messages and adding the "/help" command for easier navigation. Both requested to reduce the size of the text and use more day-to-day language.

Technically, the bot needed to be hosted on the server to ensure reliable flow even when a large number of people try to access the program, secure data storage, and accessibility to the program for support from any place (Krissaane et al., 2020; Mathew & Varia, 2014).). For this Amazon Web Services (AWS), one of the most popular cloud resources, was picked due to its low price and high quality of maintenance (Krissaane et al., 2020). Following the initial account setup, after which the EC2 instance was configured. Amazon EC2, or Amazon Elastic Compute Cloud, is a service provided by Amazon that allows users to create a cost-effective and flexible renting of a virtual server (Amazon, n.d.). The project was cloned to the server and the required dependencies were installed to run both bots.

The nohup command was applied to endure the whole project was able to run uninterrupted in the background mode. Considering the importance of user data preservation and availability for preliminary research, a Cron scheduling system was used to guarantee regular and up-to-date data backup.

By the end of this stage, TeaBot was to be fully deployed and start the experiment to validate its efficiency. The details of this process are described in the next part.

CHAPTER 4: RESULTS

This chapter specifies technical results during each stage of TeaBot development. The first segment describes the performance of GPT-3 models on recognizing and challenging distorted thoughts. This section also focuses on the price of building a GPT-3 based project. Then, the final version of the bot is presented with all functions explained in detail. The results of the psychological testing on human subjects and interviews with psychotherapists can be found in Chapter 5.

GPT-3

All four versions of GPT-3 (ada, babbage, curie, and davinci) were trained and examined with the test set to define the model that would best fit the needs of the project (see Table 1). For the purposes of comparison, the random assortment was also applied. After all tests, the curie model was chosen for the task of type recognition due to its balance between price per token and efficiency at text categorizing.

Table 1. *Analysis of GPT-3's Versions Performances on Test Set*

Version	Price per 1000 tokens, USD	Performance[b], n (%)
Random[a]	0	8 (13)
Ada	0.0004	32 (53)
Babbage	0.0005	26 (43)
Curie	0.0020	41 (68,30)
Davinci	0.0200	37 (61,67)

[a] Random is not a version of GPT-3, rather an algorithm that randomly assigned a type to the case
[b] Performance demonstrates a number of correctly recognized distortions from the test set, out of 60

Figure 7 demonstrates the performance of each model according to each distortion. In general, all models recognized catastrophizing, overgeneralization, and should statements better than others with the curie model guessing all items in the test set correctly. This can be explained due to the more structured nature of these distortions with key language hints that can be spotted when a client states their issue. For example, overgeneralization often includes "all", "always", "every", and

other words that could have been traced by the model as markers for identification. On the other hand, more complex distorted thoughts such as mental filter were in poor-recognized for all 4 versions of GPT-3. This might have happened due to the mental filter not having particular keywords and often understood via the context that transformer-based systems are reported to lack (Floridi & Chiriatti, 2020; Zhang & Li, 2021).

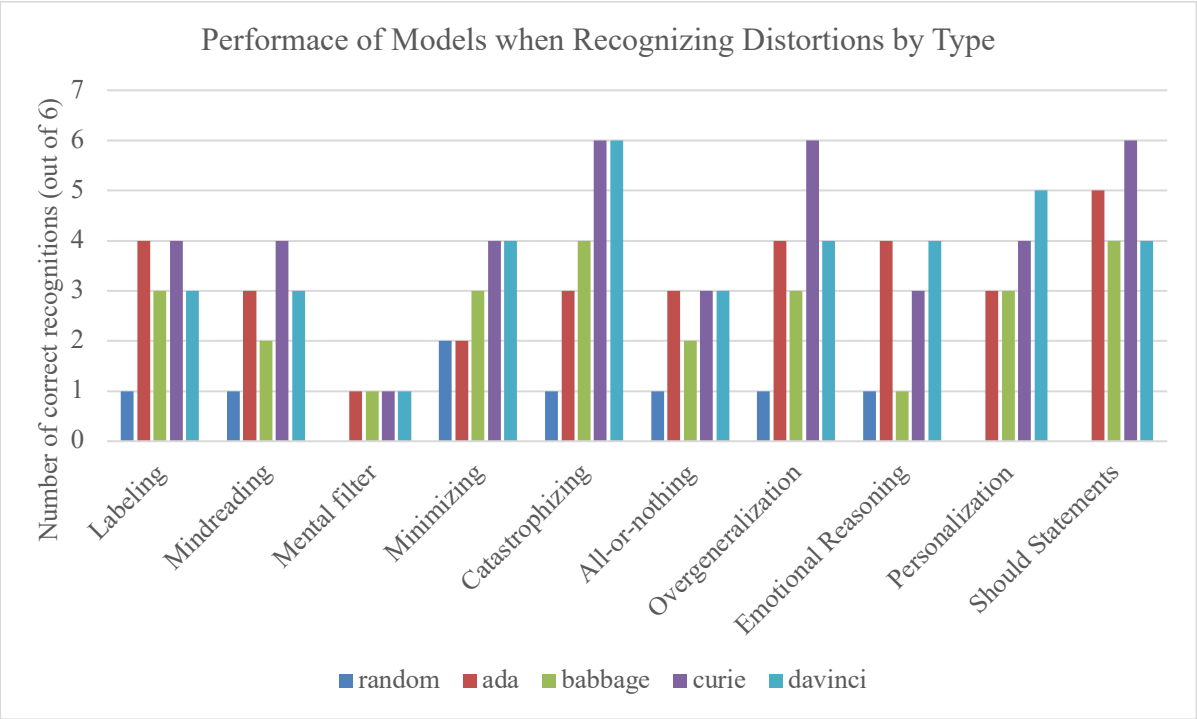

Figure 7. *Performance of Models when Recognizing Distortions by Type*

When choosing the model for applying Socratic questioning, another test was applied. All examples used for measuring performance with recognition were sent to the model and the output in a form of a question was analyzed manually by the principal investigator with some cases being documented in a more detailed manner. All models performed well when asking a challenging question if they recognized the distortion correctly (see Table 2). This might have happened due to the existing sample question that all models applied as a base for their answer and combined with the generative abilities of OpenAI models. Replies usually differ in terms of the length and depth of

questions. The simpler models like aba and babbage tend to ask 1 question that is similar to the base one with the addition of context details. This might hide from the user the existing structure of common questions that the therapist might ask and start the reflection process. Curie and davinchi models, on the contrary, tend to ask longer questions that apply more details although also using one base question. Overall, due to the ability of davinci to generate very detailed questions combined with OpenAI's (n.d.-a) statement on its quality of responses, the davinci model was chosen for generating text questions for the user.

Table 2. *Comparative Review of Model's Output (excerpt)*

User's message	Model	TeaBot's question	Comments
I don't feel that I'm ready for the future and I'm anxious about that	ada	What in particular makes you feel like you're not ready for the future?	Correct identification. Reply targets specificity of anxious behavior neutralizing pattern of generalization
	babbage	How likely is it that this person's anxiety about the future is an emotional assumption instead of an objective truth?	Correct identification. Reply targets sense of confidence about unreadiness giving space to think about objectivity of the claim. Tone is impersonal.
	curie	What makes you feel like you're not ready for the future? Do you think that there's only two outcomes for the future - either you're ready or you're not?	Correct identification. Reply targets the all-or-nothing pattern which promotes the emotional reasoning.
	davinchi	What makes you feel anxious about the future? Do you feel like you're not ready for what's to come, or do you have specific concerns about what might happen? Are there other people who share your concerns, or do you feel like you're the only one who feels this way?	Correct identification. Model asks several questions, each targeting different aspects of distortion: (1) focus on specific factor, rather than generalizing, (2) specifying the source of anxiety, and (3) examining source of evidence

Price of Implementation

Utilizing OpenAI's APIs proved to be a very cost-effective solution. OpenAI uses a token model of payment meaning that the company charges clients based on the size of their text (OpenAI, n.d.-a). Overall, developing and testing the GPT side of the project cost 12,85 USD. Preparation of GPT-3

required payment of 5,34 USD including setting up all models on the training set and examining each on a test set for both recognition and Socratic questioning. The remaining 7,51 USD were spent on experimental testing on 34 users that lasted for two months and expert interviews that required counselors to have access to the bot before and during the procedures. The single-user request was often worth much less than one cent, however, at times when user interaction peaked, the total sum reached around 9 cents per day, or about 40 messages sent to TeaBot. Figure 8 describes how user activity affected the payments. One of the most expensive parts of the project was fine-tuning the systems as it was a single intake of a large sample of data.

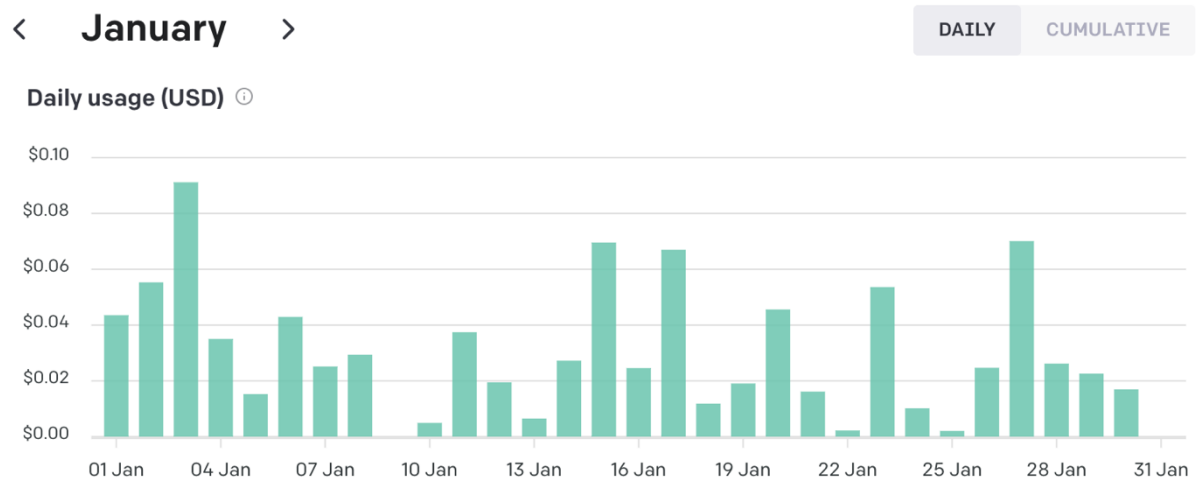

Figure 8. *OpenAI report on cost of using API in January*

AI Bot

TeaBot is the final result of this research that combines the functionality of a Telegram chatbot with an AI system based on GPT-3. Telegram as a medium helps to maintain security considering the sensitivity of the data while the library applied allows serving multiple users asynchronously., This also eased the UI/UX of the bot as users can use the bot as a part of their messaging experience. Telegram remains one of the richest messengers in terms of functionality

offering users pinning messages, and using commands when interacting with bots for easier communication, and customization. The only downfall of establishing TeaBot in Telegram remains the inability of the bot to contact users first which might decrease their activity with the technology. Application of the python-telegram-bot library allowed to add the element of additional data collection in the bot that was applied for conducting experimental research on users' interaction with the bot. Gratzer and Goldbloom (2020) stress that the unethicality of data policies remains one of the key issues when implementing AI in mental healthcare, therefore, TeaBot informs that it does not gather any content of the messages. Instead, users were tagged as "check-in" whenever they contacted a TeaBot to analyze user engagement.

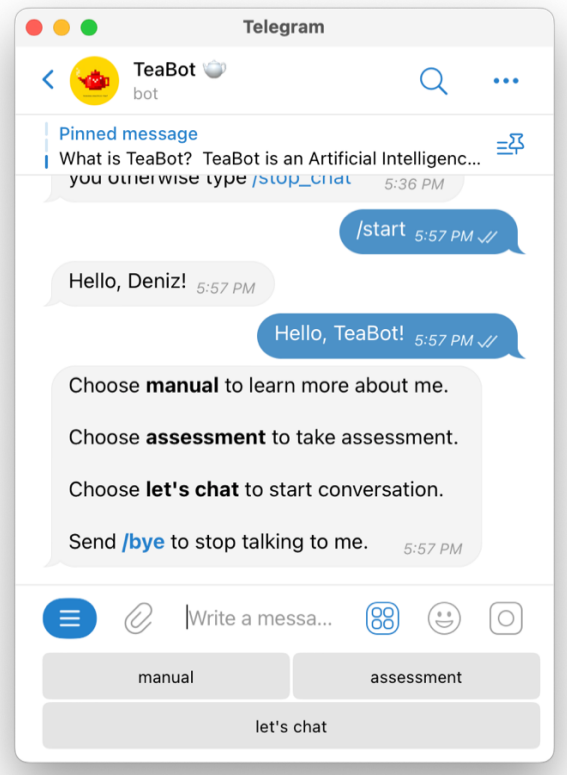

Figure 9. *TeaBot's Interface*

The user interacted with TeaBot in the following manner: informed consent, greeting, and access to three functions: manual, assessment, or let's chat (see Figure 9). During the very first meeting with the program, one receives informed consent and a tutorial on how to use the chatbot.

This is needed to ensure that the user is well aware of what the bot can and cannot do. The user is obligated to give consent in order to continue using the application. After that, the greeting function was added to break the initial discomfort (see Figure 10 for a detailed diagram). The bot also asks the user to come up with the name if the username in Telegram is not specified. After that, TeaBot suggests its functionality in a continuous flow manner until the user chooses to end the interaction by typing "/bye". The new conversation starts with a greeting without asking for consent again.

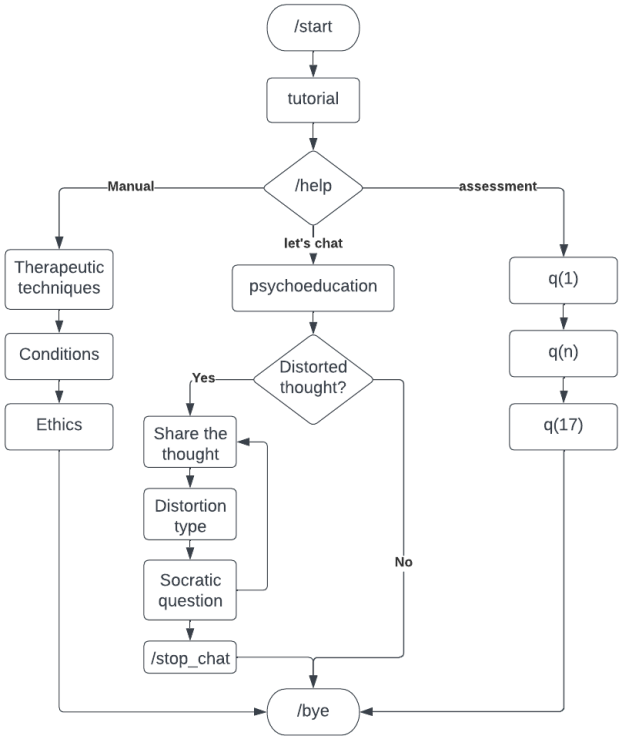

Figure 10. *TeaBot's scheme of work*

The manual function delivers to the user information about the Bot and acts as a therapist who would introduce the bot to the patients. The Manual consists of three sections: "therapeutic techniques used" which mainly explains CBT, the concept of distorted thoughts, their types, and how to battle them; "conditions for participation" which is used to get a copy of the informed consent user agreed before, and "ethics and safety" that informs what institutions approved the bot

and how data is collected. In the future, this manual can be used as training for practitioners who would like to include the bot in their practice.

The assessment function was developed to test the process of data collection by the bot. The results of all assessments are recorded on a separate file which allows to store data more efficiently compared to the traditional paper structure. This function was mainly used for completing the validating part of the research as the digital assessment allows gathering information from participants much faster, from different places anonymously without breaking contact with the application compared to the traditional Google form approach. This segment can also be later redesigned to add additional assessments depending on the needs of the client.

The let's chat is an AI-powered function that is a fundamental feature of TeaBot. The dialogue starts with a psychoeducational question that aims to filter whether the bot can be actually helpful to a person. As the bot is designed to work only with cognitive distortions and not mere chit-chat, the program explains to the user what a cognitive distortion is and asks whether a person has them. When receiving a positive reply, TeaBot invites a user to share this thought in a cause-effect format. After that, the GPT-3 functions are turned on. Users can talk to TeaBot for an unlimited amount of time until sends "/stop_chat" which ends the process. Figure 11 depicts how the final version of TeaBot communicates with users in AI mode.

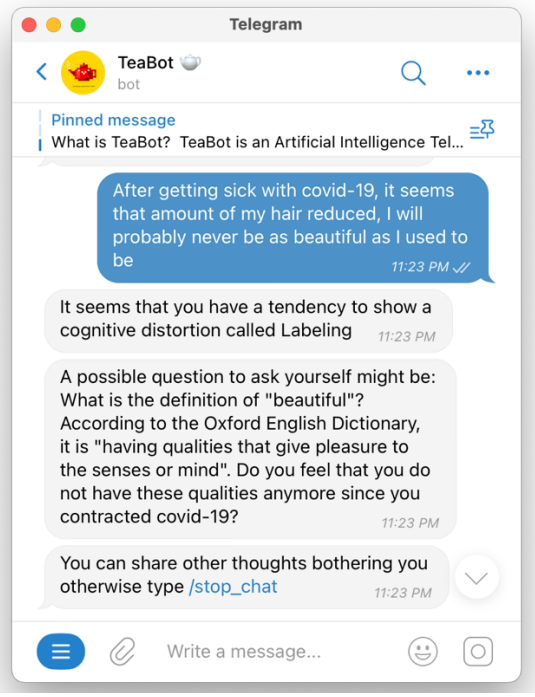

Figure 11. *TeaBot's Let's Chat Function*

In conclusion, TeaBot, an AI Telegram chatbot based on GPT-3 technology is a ready-to-use tool for reducing cognitive distortions. The GPT-3 models were fine-tuned on a self-developed dataset and tested to identify the curie model for recognition purposes and davinci for purposes of generating responses. Both tests demonstrated that OpenAI APIs perform very well in both tasks with slight differences depending on the needs of the project. The program can execute the following functions: taking assessments and saving them in a database, giving psychoeducational content, and chatting using an AI model fine-tuned on the data gathered for this project specifically. The software also includes the manual that aims to serve as a mediator for the future implementation of augmented therapy where the bot is used in tandem with a psychologist.

The next chapter explains the quantitative and qualitative results of the testing of the TeaBot both on users and experts who might be interested in integrating augmented therapy into their practice.

CHAPTER 5: PSYCHOLOGICAL TESTING

Experimental Testing on Users

Quantitative analysis of TeaBot's performance consisted of an 8-week experiment done on 68 AUCA students who were recruited using convenient sampling based on a social media campaign and universitywide newsletter. After the recruitment was done, participants were randomly divided into two groups: control and experimental. The experimental group received intervention in the form of interaction with TeaBot and was asked to use a manual for learning more about the therapeutic approach used. The control group received no intervention with only a manual available to learn more about distortions. Each group's progress was measured by two questionnaires one of which evaluated the test taker's relationship with their thoughts, and the other estimated their level of cognitive distortions. These assessments were conducted three times: twice before the intervention itself and once after the experiment was over. After that, the gathered data was analyzed using the statistical software JASP. Figure 12 depicts the schematic process of how this part of the research was executed.

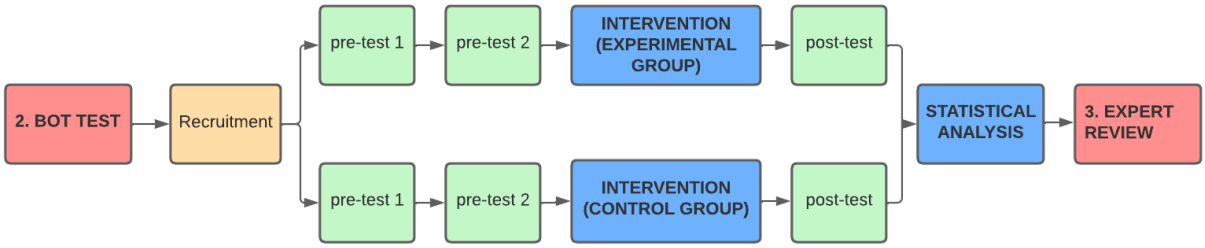

Figure 12. *Experimental Testing Scheme*

Regarding ethical standards, the experiment was approved by AUCA Institutional Review Board. All participants gave informed consent, had the right to withdraw at any moment, and were notified about what group they joined. All participants were compensated with 500 soms.

Statistical Findings

The quantitative data analysis to measure TeaBot efficiency included multiple vectors for research: normality check, general efficiency, attrition, use, influences of demographic factors, and factor analysis. A normality check was applied to test whether the results are normally distributed and can be generalized. All efficiency tests passed the normality check. When estimating the efficiency of TeaBot, the difference between pretests and posttests was compared, and the experimental group demonstrated statistically significant improvement in scores of both assessments (see Figure 13). Talking to TeaBot on average reduced the scores of the first questionnaire to the one below the cutoff level assisting the participants to leave the risk zone for developing mental disorders.

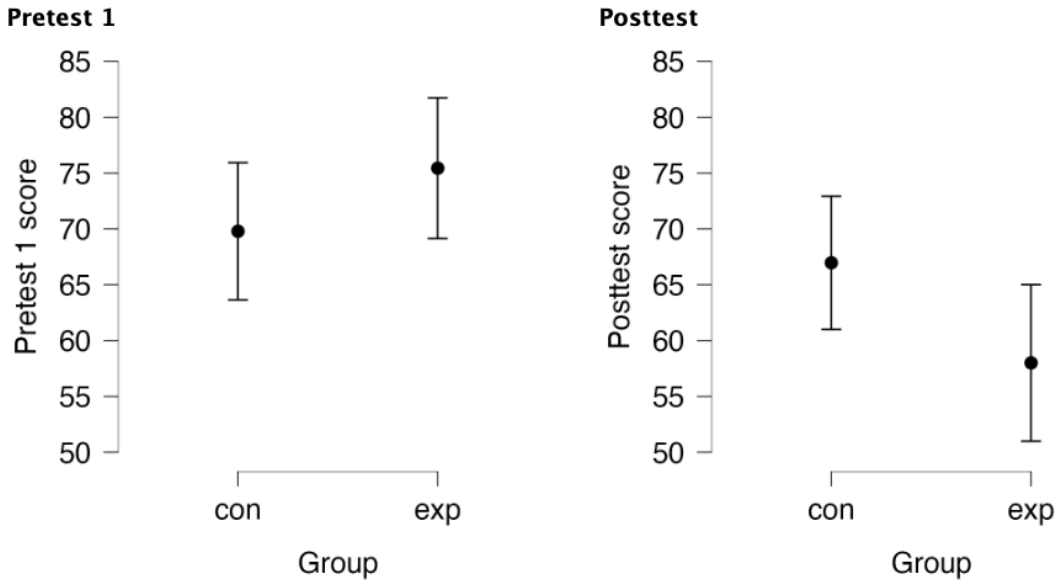

Figure 13. *Change in Mean Results (AAQ+CDS) for Both Groups over the period of the experiment*

Attrition

Figure 14 demonstrates the flow of the participants during the project. Out of 110 participants who contacted the researcher interested in the experiment, 81% (35/54) submitted

partial data, e.g., at least one pretest, and 62% (34/34) completed the experiment. Attrition rates differ between the control and experimental group with 27,66% for the control group and 46,03% for the experimental. Nevertheless, there was no difference in pretest-1 results detected between those who completed the experiment and the ones who withdrew.

Figure 14. *Participants' Flow*

Use

On average, 29 out of 46 users who completed the second assessment used TeaBot every week. The maximal number of users was observed during the first week (43 users), and the minimum was detected during the sixth week (18 users). The average duration of bot usage was 6 weeks, with 12 people (35,29%) interacting for all 8 weeks. In general, users contacted the bot on particular dates that coincide with the dates when notifying emails were sent. The maximum number of messages sent by one user in a week was 47, minimum – 0. Reflecting on this variance and the presence of outliers, the median was calculated for each week and, consequently, the mean of these values was equal to 6.125 suggesting that, on average, users send 6 messages per week. In total, the mean number of check-ins throughout the 8-week period was 51.

Regarding the timing of the messaging, bots are often referred to as highly accessible tools that can be used when human therapists are out of reach (Kretzschmar et al., 2019). Consequently, the hours when users contacted TeaBot were divided into three categories: night (from 01:00 to

9:00) when most specialists are unavailable, working day (from 09:00 to 17:00) when a client can approach counselors for a session, and post-working hours (from 17:00 to 01:00) when most psychotherapists do not conduct sessions (see Table 3). Interestingly, the difference between the number of messages sent during the working ($N = 2593$) and past-working ($N = 2350$) hours is small suggesting that people used TeaBot both when therapy was available and not. The number of messages peaked after 19:00 ($N = 450$) with the second maximum being after 11:00 ($N = 409$), with the lowest point of activity being after 5:00 ($N = 0$). A breakdown of activity by the hour is depicted in Figure 15.

Table 3. *User Engagement According to Time Periods*

Time Period	Number of Messages Sent
01:00 – 09:00	317
09:00 – 17:00	2593
17:00 – 01:00	2350

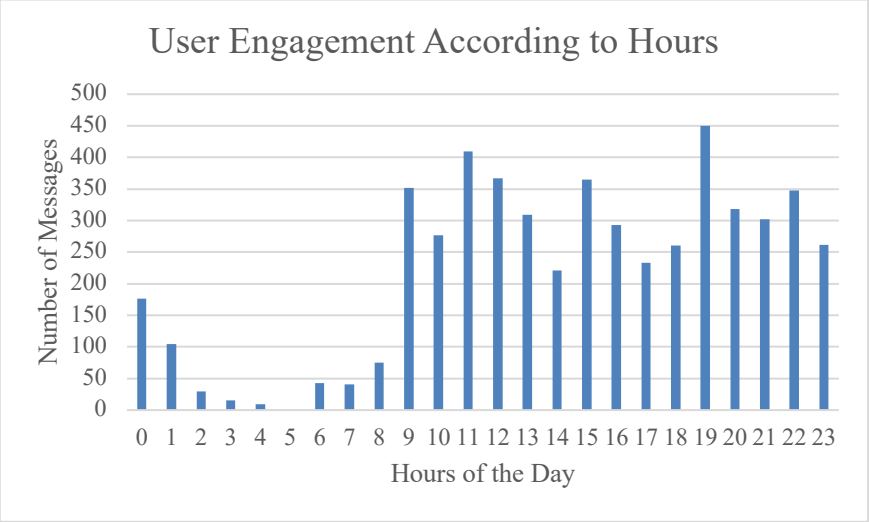

Figure 15. *User Engagement According to Hours*

Expert Interview

The quantitative side of the research focused on receiving feedback on TeaBot's performance from practicing psychologists while also exploring the topic of augmented therapy and learning how TeaBot can be included as a part of therapy rather than a standalone intervention. Participants were 7 counseling psychologists who studied or practice CBT and speak English and were recruited based on the suggestions from the psychology-side supervisor as well as snowball sampling when participants suggested their colleagues who might be interested in participating in the project. The data was collected through the interviews which included questions on specialist's contentment with TeaBot's quality, views on implementing it in their practice, and recommendations for the next version of the bot. Before the interview, counselors had access to the bot to ensure they had enough time to examine TeaBot's features. Interviews lasted around 40 minutes. Thematic analysis was applied to find out the common topics between the interviews and divide them into smaller subtopics. After this stage, recommendations were applied to the final version of the bot. Figure 16 depicts how the process went schematically.

Figure 16. *TeaBot's Let's Chat Function*

Interviews demonstrated concordance with the following topics: TeaBot as an application, in-therapy tool, out-therapy one, and recommendations. In general, counselors gave positive feedback on TeaBot's functionality and performance. As an application, it was reported to have high accuracy when predicting distortions, the format of the chatbot was seen as a more immediate instrument compared to paper-based tests and homework as well as having strong potential to work

well with young people. Regarding the in-therapy use of TeaBot, the majority of counselors ($N = 6$) expressed a high probability of integrating TeaBot into their therapy. TeaBot was seen as equipment for CBT homework, diary, and assessment while also promising delegation of labor as a follow-up intervention between therapy sessions or after therapy ends. Using TeaBot can assist with creating a more therapeutic environment when a client is immersed in reflective processes even after the session ends. For those who did not attend therapy, counselors see TeaBot as a psychoeducational aid demystifying certain aspects of therapy that might also serve as a gateway application that can be especially helpful considering the present stigma about attending therapy.

Recommendations mainly included suggestions about improving the accuracy of TeaBot when recognizing distortions. Almost all counselors suggested expanding the rationale, and explanation of why they should use TeaBot to ensure clients know when and how they need to use TeaBot. Some of the participants suggested developing training for psychotherapists who are interested in using the bot in their practice. To better integrate the bot with therapeutic sessions, it was advised to add an element of statistics, so a client can see their progress and share it with the counseling psychologist. These suggestions were implemented in the third version of TeaBot.

CHAPTER 6: LIMITATIONS AND FURTHER WORK

Limitations

As with any language-based technology at its initial stage, TeaBot operates only in one language. The decision to focus on English was made due to the complexities of translating OpenAI's system into Kyrgyz or Russian which might have affected the quality of software performance. Considering that TeaBot needed not only to recognize distortions but also to clearly communicate in an understandable manner, any challenge with communication in Kyrgyz or Russian might have significantly affected the perception of the technology. Another disadvantage of the project might have been the absence of a notification function for Telegram bots to avoid spamming. This challenge was mitigated via notifications via email, however, Fitzpatrick et al. (2017) inform about the significance of a bot's initiating the chat. The integrated reminders to talk might decrease attrition levels and user engagement. However, one of the counselors highlighted that because they were able to contact the bot only when they wanted might help with wanting to write to the bot more as the bot does not.

Artificial Intelligence remains one of the most fast-growing fields in the domain of Software Engineering. During the development of the project, two major events happened: the release of ChatGPT and GPT-4. While the first one was partially tested when analyzing the capability of AI properly recognize and communicate with a person displaying distorted thoughts, the gpt-3.5-turbo model on which ChatGPT is based was released for public use at the end of the user testing process. GPT-4 was announced at the moment of finalizing the research data with APIs being available only through granted access by OpenAI. Nevertheless, at the moment of writing, both products lack access to the fine-tuning function reducing the autonomy of developers when building applications. This feature was especially important for ethical considerations, thus, combined with the notion that

the project consisted not only of the engineering stage that included evaluation of different models and price analysis but also of both 8-week-long experimental testing on users and qualitative interviews with practicing psychologists, the decision was made to focus on the GPT-3 version of TeaBot for preserving the integrity of the research. As this project was made by one software developer, the preference was made for the quality of the product. The world of AI is everchanging, yet the research methods applied in this research remain replicable for the new versions once fine-tuning is available while still presenting a product that was tested and modified according to the feedback.

Further Work

The next generation of chatbots can be developed based on the aforementioned gpt-3.5-turbo and GPT-4 models. Besides a quantitative analysis of the size of data and performance in recognition, future research might attempt to conduct an intergenerational analysis of OpenAI models to determine the key characteristics that helped to improve Human-Computer Interaction.

Similar projects should apply larger datasets for the recognition of distorted thoughts and find the optimal size of data for fine-tuning addressing the rising critique that larger datasets do not always guarantee better performance (Bender et al., 2021). As Macleod et al. (2020) state that studies are often done for privileged segments of the population, the researchers might be interested in developing bots that are trained on datasets developed specifically for marginalized groups and test the ability of technology to improve the mental state of the oppressed when such datasets used.

If remaining in the medium of Telegram, developers might implement more Telegram functions such as reactions which can be used for improving the quality of conversation based on immediate users' feedback. This technology is already utilized in ChatGPT and can help developers to better track the pitfalls of the application.

Future products might implement more human-like behavior into the bot's conversation with users. Due to several challenges with the proper integration of emojis into the Telegram reply system as well as GPT-3 feedback, it was decided to minimize the number of cases when they are employed. Kim et al. (2021) report the positive effect of facial expressions on users' willingness to self-disclose when communicating with a bot. Consequently, developers might experiment with GPT's ability to generate human-like text with emojis or similar colloquial methods of communication, e.g. ":)".

Being mindful of the pace with which AI develops nowadays, the main suggestion would be to be open to any experiments with the technology. At the moment when this project was initiated, GPT-3 was one of the novel discoveries known in small circles of AI enthusiasts.

CHAPTER 6: CONCLUSIONS

The impact of previous years on youth's mental health and the possible consequences of untreated traumas on future development urge to emphasize the importance of this topic (Kretzschmar et al., 2019; WHO, 2022). Considering the existing lack of sources, especially in developing countries like Kyrgyzstan, AI technology can be helpful due to its affordability, accessibility, and security when addressing stigmatized topics like mental health (Abd-alrazaq et al., 2019; Orlova, 2019; Pinchuk et al., 2021).

Consequently, this research focused on the exploration of the phenomenon of AI technology in therapy. The goal was to create software that can be used to help close the gap in providing mental healthcare for young people. The main outcome of this research is TeaBot, an AI chatbot that tested the efficiency of the GPT-3 models for different therapeutic tasks and can be used for working with cognitive distortions. The OpenAI model was chosen due to the complexity of the natural language processing and the lack of resources for the development of the independent model trained on a reliable size of data. As modern AI technologies still demonstrate problems with context understanding, fine-tuning was applied using a self-developed dataset to reduce the chances of a machine responding in a way that can harm a user. Furthermore, a price analysis was conducted to determine the cost-effective solution to the problem. TeaBot was able to recognize accurately between 10 cognitive distortions and provide the user with thought-provoking questions to challenge the negative thoughts.

TeaBot was validated quantitatively by users and qualitatively by practicing psychologists. The user testing lasted for 8 weeks which is quite rare in the field with most research being limited to 2 weeks (Reardon et al., 2002; Sankar et al., 2021). Research data suggest that TeaBot is indeed an efficient tool when working with distorted thoughts implying possibilities for using it both as a

mechanism for the prevention and intervention of mental disorders. Moreover, there was a positive correlation between the number of weeks users spent communicating with TeaBot and a decrease in their level of cognitive distortions that allow assuming that regular exposure to the technology leads to better results. The participants of the experiment were AUCA students to estimate the effect of AI technology on a young and diverse group of people. Feedback from experts was positive proposing possibilities for augmented therapy where TeaBot is applied as an extension when a therapist is out of reach. In addition to the existing version of the bot, a manual was developed to practice the application of the bot in therapy as well as inform the users who are new to counseling what therapeutic methods were utilized. The development process also included a manual for the future practice of augmented therapy.

Despite the TeaBot being in non-native users' language and the continuous release of new GPT versions, TeaBot remains one of the first known bots in Central Asia developed in collaboration with OpenAI's models and tested in accordance with ethical research methods.

Overall, TeaBot's first steps demonstrate the huge potential of technology.

References

Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*, Article 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Active Minds. (2020). *April 2020 Survey Data*. Retrieved from Active Minds: https://www.activeminds.org/studentsurvey/

Amazon. n.d. *Amazon EC2*. AWS. https://aws.amazon.com/ec2/

AvatarMachine, LLC. (n.d.). *Сабина Ai – Ваш ИИ друг*. Sabina-Ai. https://sabina-ai.com/

Beck, J. S. (1995). *Cognitive therapy: Basics and beyond*. New York: Guilford.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices*, *18*(sup1), 37-49.

Brown, L. S. (2018). Introduction: Feminist therapy—Not for cisgender women only. In L. S. Brown, *Feminist therapy* (pp. 3–10). American Psychological Association. https://doi.org/10.1037/0000092-001

Covin, R., Dozois, D. J., Ogniewicz, A., & Seeds, P. M. (2011). Measuring cognitive errors: Initial development of the Cognitive Distortions Scale (CDS). *International journal of cognitive therapy*, *4*(3), 297-322.

Dale, R. (2021). GPT-3: What's it good for?. *Natural Language Engineering*, *27*(1), 113-118.

D'Alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., Gleeson, J., & Alvarez-Jimenez, M. (2017). Artificial intelligence-assisted online social therapy for youth mental health. *Frontiers in psychology*, *8*, 796.

Damij, N., & Bhattacharya, S. (2022, April). The Role of AI Chatbots in Mental Health Related Public Services in a (Post) Pandemic World: A Review and Future Research Agenda. In *2022 IEEE Technology and Engineering Management Conference (TEMSCON EUROPE)* (pp. 152-159). IEEE.

Dekker, I., De Jong, E. M., Schippers, M. C., De Bruijn-Smolders, M., Alexiou, A., & Giesbers, B. (2020). Optimizing students' mental health and academic performance: AI-enhanced life crafting. *Frontiers in Psychology*, *11*, 1063.

Depounti, I., Saukko, P., & Natale, S. (2022). Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media, Culture & Society*, 01634437221119021.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), Article e19. https://doi.org/10.2196/mental.7785

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health*, *5*(4), e9782.

Gabrielli, S., Rizzi, S., Bassi, G., Carbone, S., Maimone, R., Marchesoni, M., & Forti, S. (2021). Engagement and effectiveness of a healthy coping intervention via chatbot for university students: Proof-of-concept study during the COVID-19 pandemic. *JMIR MHealth and UHealth, 9*(5), Article e27965. https://doi.org/10.2196/27965

Google. (2022, July 18). *Background: What is a Generative Model?*. Google Developers. https://developers.google.com/machine-learning/gan/generative

gov.kg. (2022, September 20). *Информация по работе по восстановлению и помощи Баткенской области в связи с событиями 14-17 сентября*. Кабинет Министров Кыргызской Республики. https://www.gov.kg/ru/post/s/21944-14-17-sentyabrdagy-okuyalarga-baylanyshtuu-batken-oblusun-kalybyna-keltir-ishteri-zhana-zhardam-krst-boyuncha-maalymat.

Gratzer, D., & Goldbloom, D. (2020). Therapy and e-therapy—preparing future psychiatrists in the era of apps and chatbots. *Academic Psychiatry*, *44*(2), 231-234.

Grové, C. (2021). Co-developing a mental health and wellbeing chatbot with and for young people. *Frontiers in psychiatry*, *11*, 606041.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, *61*(4), 5-14.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J., Yuan, J., Zhao, W., X., & Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225-250.

Heikkilä, M. (2022, August 31). *What does GPT-3 "know" about me?*. MIT Technology Review. https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, *6*(11), e12106.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4873-4882)

Kim, M., Park, J., & Lee, M. (2021). The Effects of Chatbot Anthropomorphism and Self-disclosure on Mobile Fashion Consumers' Intention to Use Chatbot Services. *Journal of Fashion Business*, *25*(6), 119-130.

Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People's Advisory Group. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical informatics insights*, *11*, 1178222619829083.

Kucenko, V. (2022, March 10). *Война травмировала психику десяткам миллионов украинцев, и просто жить дальше не выйдет - психолог*. Telegraf. https://telegraf.com.ua/ukraina/2022-03-10/5698950-viyna-travmuvala-psikhiku-desyatkam-milyoniv-ukraintsiv-i-prosto-zhiti-dali-ne-viyde-psikholog.

Kumar, H., Musabirov, I., Shi, J., Lauzon, A., Choy, K. K., Gross, O., Kulzhabayeva, D., & Williams, J. J. (2022). Exploring The Design of Prompts For Applying GPT-3 based Chatbots: A Mental Wellbeing Case Study on Mechanical Turk. *arXiv preprint arXiv:2209.11344*.

Larionov, K. (2021, June 4). *Проблемы и возможности быстрого социального и экономического роста Баткенской области*. *ЦППИ.* https://center.kg/article/398.

Lauber, C., & Rössler, W. (2007). Stigma towards people with mental illness in developing countries in Asia. *International review of psychiatry*, *19*(2), 157-178.

Luxton, D. D. (2016). An Introduction to Artificial Intelligence in Behavioral and Mental Health Care. *Artificial Intelligence in Behavioral and Mental Health Care*, (pp. 1–26). Academic Press. https://doi.org/10.1016/b978-0-12-420248-1.00001-5

Macleod, C. I., Bhatia, S., & Liu, W. (2020). Feminisms and decolonising psychology: Possibilities and challenges. *Feminism & Psychology*, *30*(3), 287-305.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, *27*(4), 12-14.

Oilobot. (n.d.). Oilobot. https://oilobot.info/

OpenAI. (2022, November 30*). ChatGPT: Optimizing Language Models for Dialogue*. OpenAI. https://openai.com/blog/chatgpt/.

OpenAI. (n.d.-a). *Models*. OpenAI. https://platform.openai.com/docs/models

OpenAI. (n.d.-b). *Fine-Tuning*. OpenAI. https://platform.openai.com/docs/guides/fine-tuning

Open Line. (2020, June 22). *В Кыргызстане появилась первая игра-сериал про ала качуу для смартфонов*. Retrieved from Open Line: https://openline.kg/new/%D0%B2-%D0%BA%D1%8B%D1%80%D0%B3%D1%8B%D0%B7%D1%81%D1%82%D0%B0%D0%BD%D0%B5-%D0%BF%D0%BE%D1%8F%D0%B2%D0%B8%D0%BB%D0%B0%D1%81%D1%8C-%D0%BF%D0%B5%D1%80%D0%B2%D0%B0%D1%8F-%D0%B8%D0%B3%D1%80%D0%B0-%D1%81/

Oracle. (n.d.). *What is Big Data?* https://www.oracle.com/big-data/what-is-big-data/

Oremus, W. (2023, February 14). *The clever trick that turns ChatGPT into its evil twin*. The Washington Post. https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak/

Orlova, M. (2019, February 21). *Mental Health in Kyrgyzstan. Governmental program is not executed*. 24kg. https://24.kg/obschestvo/109829_psihicheskoe_zdorove_vkyirgyizstane_pravitelstvennaya_programma_neispolnyaetsya/

Palosaari, K. (2022). Bots for everyone: designing an individualised chatbot avatar for optimised customer engagement with an eye on diversity.

Pinchuk, I., Yachnik, Y., Kopchak, O., Avetisyan, K., Gasparyan, K., Ghazaryan, G., Chkonia, E., Panteleeva, L., Guerrero, A., & Skokauskas, N. (2021). The implementation of the WHO mental health gap intervention guide (mhGAP-IG) in Ukraine, Armenia, Georgia and Kyrgyz Republic. *International Journal of Environmental Research and Public Health*, *18*(9), 4391.

Reardon, M. L., Cukrowicz, K. C., Reeves, M. D., & Joiner, T. E. (2002). Duration and regularity of therapy attendance as predictors of treatment outcome in an adult outpatient population. *Psychotherapy Research*, *12*(3), 273-285.

Romanovskyi, O., Pidbutska, N., & Knysh, A. (2021). Elomia Chatbot: The Effectiveness of Artificial Intelligence in the Fight for Mental Health. In *COLINS* (pp. 1215-1224).

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44, 206-226. https://doi.org/10.1147/rd.441.0206

Sankar, A., Panchal, P., Goldman, D. A., Colic, L., Villa, L. M., Kim, J. A., Lebowitz, E. R., Carrubba E., Lecza B., Silverman, W. K., Swartz, H. A., & Blumberg, H. P. (2021). Telehealth social rhythm therapy to reduce mood symptoms and suicide risk among

adolescents and young adults with bipolar disorder. *American journal of psychotherapy*, *74*(4), 172-177.

Shankland, S. (2023, February 10). *Why the ChatGPT AI Chatbot Is Blowing Everybody's Mind*. CNET. https://www.cnet.com/tech/computing/why-the-chatgpt-ai-chatbot-is-blowing-everybodys-mind/.

Sulaiman, S., Mansor, M., Wahid, R. A., & Azhar, N. A. A. N. (2022). Anxiety Assistance Mobile Apps Chatbot Using Cognitive Behavioural Therapy. *International Journal of Artificial Intelligence*, *9*(1), 17-23.

Team, A. (2019). AlphaStar: Mastering the real-time strategy game StarCraft II. *DeepMind blog*, *24*.

Toledo, L. n.d. *Documentation*. python-telegram-bot. https://docs.python-telegram-bot.org/en/stable/

Torous, J., Myrick, K. J., Rauseo-Ricupero, N., & Firth, J. (2020). Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR mental health*, *7*(3), e18848.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433-460.

UNICEF. (2020). *Анализ ситуации в области сущила и суицидальных попыток среди подростков и молодёжи в Кыргызстане.* Бишкек.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaizer, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vilaza, G. N., & McCashin, D. (2021). Is the Automation of Digital Mental Health Ethical? Applying an Ethical Framework to Chatbots for Cognitive Behaviour Therapy. *Frontiers in Digital Health*, *3*. https://doi.org/10.3389/fdgth.2021.689736

Watts, S., Mackenzie, A., Thomas, C., Griskaitis, A., Mewton, L., Williams, A., & Andrews, G. (2013). CBT for depression: a pilot RCT comparing mobile phone vs. computer. *BMC psychiatry*, *13*(1), 1-9.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

World Health Organization. (2022, March 2). *COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide*. WHO: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

Xiang, C. (2023, January 10). *Startup Uses AI Chatbot to Provide Mental Health Counseling and Then Realizes It 'Feels Weird'*. Vice: https://www.vice.com/en/article/4ax9yw/startup-uses-ai-chatbot-to-provide-mental-health-counseling-and-then-realizes-it-feels-weird

Yang, M. (2020). <? covid19?> Painful conversations: Therapeutic chatbots and public capacities. *Communication and the Public*, *5*(1-2), 35-44.

Yetistiren, B., Ozsoy, I., & Tuzun, E. (2022, November). Assessing the quality of GitHub copilot's code generation. In *Proceedings of the 18th International Conference on Predictive Models and Data Analytics in Software Engineering* (pp. 62-71).

Zeavin, H., & Peters, J. D. (2021). *The Distance Cure: A History of Teletherapy*. The MIT Press.

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831-833.