Eyas Humeid s3603857

Beyond ELIZA: Creating an LLM-Based Mental Health Tool

The way I interact with technology has fundamentally changed since November 2022. Since the release of ChatGPT 2.5 years ago, I had personal and at times even intimate conversations with the large language model (LLM). At the time of its release, I had moved to the Netherlands from Jordan three months earlier and was having trouble settling in. My mental health gradually deteriorated in that period, as I was consistently unhappy with myself and my environment. Tending to my university obligations became increasingly difficult, so to get some stress relief, I overused ChatGPT. I rapidly became well-versed in the tool and it took some weight off my shoulders by helping me with my academic tasks. Eventually, the extent of cognitive outsourcing stripped all meaning from my courses because I wasn't trying to understand the material deeply.

This all culminated in my fifth semester, when I was on the verge of dropping out. I decided to remain in university, but this time with intention. I knew that I had to change my mindset in order to reconnect with myself and my studies. I gradually integrated meditation, journaling, and nature walks into my routine, which alleviated much of the distress I was feeling. The way I used ChatGPT shifted: though it still served primarily as a learning tool, the focus moved from outward to inward. I became the main object of study. ChatGPT, to some extent, facilitated my learning about how to deal with mental health.

I discovered that as early as the 1960s, people were using chatbots like ELIZA, developed at MIT by Joseph Weizenbaum, to better understand themselves. ELIZA followed the Rogerian psychotherapy principle of reflecting users' words back in the form of a question to help them reach deeper self-understanding. This inspired me to create a similar tool using today's LLM technology.

Almost 60 years later, the technology behind chatbots has grown to an incredible level of sophistication. This raises the question: if Weizenbaum was able to navigate the technological limitations of his time in such a way that even a simple program made people feel understood, what is possible today? **This research will investigate how I can create a mental health tool with ChatGPT through prompt personalisation.** To answer this question, this study will employ an autoethnographic approach, as a self-analysis is essential for the creation of a personalized prompt tailored to my mental health needs. To further bolster this prompt, this investigation will review relevant literature on mental health chatbots. It will begin with Weizenbaum's original ELIZA paper, then examine modern mental health chatbot apps to gather insights for the prompt.

Methodology:

This research will employ an autoethnographic approach, which consists of three dimensions:

Auto (self): I engage in a critical self-reflection through “memory work”, drawing on emotionally charged life events, and interactions with ChatGPT. This allows for a contextualized exploration of how digital tools intersect with my mental well-being. **Ethno** (culture): My personal experiences are situated within larger cultural, technological, and therapeutic contexts. I reflect not only on my own psychological and emotional responses, but also the development of the interaction between mental health and digital technology; more specifically chatbots. **Graphy** (writing): This refers to the craft of representations–how the story is told. This is achieved by using my personal reflections openly and integrate them in a research paper format, by blending personal narrative with academic analysis.[1]

To structure this process clearly, I combine a literature review, therapeutic frameworks, and personal reflection as a means of prompt engineering. This method unfolds in five iterative stages:

1. **Review of Literature**: I examine existing academic feedback on mental health chatbots, beginning with ELIZA, then moving on to modern mental heath chatbot apps, to identify both their strengths, limitations, user feedback, and risks. This helps me understand what is feasible and what limitations I can bypass.
2. **Personal Reflection**: The prompt must be tailored to my individual experience. I reflect on the common issues I face in my mental well-being and consider where and how I would most likely use the chatbot, what I struggle with, what challenges I’m currently facing, and how the tool can be made accessible, especially during activities like walking where voice interaction is key.
3. **Integration of Therapeutic Principles**: I explore which elements from CBT and Rogerian psychotherapy are suitable for chatbot implementation, focusing on effective reflection techniques and the potential for fostering a therapeutic alliance.
4. **Prompt Creation**: I construct a personalized prompt using principles of prompt engineering, synthesizing insights from the literature, therapeutic frameworks, and my own needs.
5. **Discussion**: I critically reflect on my emotional experience using the prompt, mention limitations, concerns, and explore further areas of research.

[1] Tony E. Adams, Stacy Holman Jones, and Carolyn Ellis, eds., *Handbook of Autoethnography*, 2nd ed. (New York: Routledge, 2021), 3.

Theoretical Framework:

To create a mental health tool with ChatGPT, this section outlines the key theoretical foundations of this research.

Psychotherapy: is defined as a treatment that helps patients talk about painful and distressing states such as depression, anxiety, and aimlessness.[2] It does not eliminate suffering, but instead builds the capacity to respond differently to difficult and stressful events or circumstances .[3] Crucially, psychotherapy is an artificial encounter; it is a model relationship, one in which emotional distress is more properly addressed by sharing it and processing it through our thoughts, feelings, bodies, and social connections.[4] Framing it as a model relationship opens the possibility for artificial agents, such as chatbots, to serve therapeutic functions within a similarly structured relational model.[5]

Rogerian Psychotherapy (Client-Centred Therapy): Developed by Carl Rogers, it focuses on the importance of the therapeutic relationship for effective therapy. There are three necessary and sufficient conditions for effective therapy which are, acceptance and unconditional positive regard for client, empathy, congruence (genuineness). The therapist engages in reflective listening and clarifying the client's attitudes.[6] Rogers emphasizes that client-centred therapy should trust the client's ability to guide their own therapy, uncover deeper insights, manage emotional pain, and determine when they are ready to cope independently.[7] Although client-centred therapy has grown highly complex, the core relationship variables identified by Rogers are consistently found to be related to therapeutic effectiveness across different therapy types[8].

Cognitive-Behavioural Therapy: CBT examines thoughts (cognition) and actions (behaviours). It originally developed to treat depression, but proved effective for a wide range of problems (p. 329 towards).[9] It is a structured form of therapy, progressing through defined phases, and the therapist takes a proactive role in directing and focusing treatment.[10] A core technique is the thought record, which involves three processes: 1. accessing automatic thoughts, 2. examining thoughts, and 3. developing new beliefs. Eliciting automatic thoughts involves aiding clients in focusing on emotionally distressing situations and allowing them to identify the thoughts they were having at the time. Therapists guide this process through Socratic-style questioning and by reflecting back the client's thoughts.[11] Once a key thought has been identified, the next step is to

---

[2] Elizabeth A. Wilson, *Affect and Artificial Intelligence* (Seattle: University of Washington Press, 2010), 83.
[3] Wilson, *Affect and Artificial Intelligence*, 84.
[4] Wilson, *Affect and Artificial Intelligence*, 84.
[5] Wilson, *Affect and Artificial Intelligence*, 85.
[6] *Carl R. Rogers, "Significant Aspects of Client-Centered Therapy," American Psychologist 1 (1946): 416.*
[7] Rogers, "Significant Aspects of Client-Centered Therapy," 418.
[8] *Nina Josefowitz and David Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," Counselling Psychology Quarterly 18, no. 4 (2005): 329. https://doi.org/10.1080/09515070500473600.*
[9] Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 330.
[10] Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 330.
[11] Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 332.

examine the evidence that either supports or disconfirms it. The aim is not for the therapist to tell the client what to think, but rather to structure the process of examining the meanings assigned to events .[12] Clients are taught to label their thinking patterns using terms such as “catastrophic thinking,” “mind reading,” or “personalization,” also known as cognitive distortions.[13] CBT is relevant to my project as it supports structured self-reflection which I want to integrate into my tool.

Prompt engineering: is the process of constructing and arranging inputs (i.e. prompts) for LLMs to receive precise, coherent, and pertinent responses. It is an emerging field of research that shapes how LLMs understand tasks, process information, and generate responses across a wide range of natural language applications. Essentially, it is the art of fine-tuning the questions or commands provided to AI models to optimize their performance and ensure they produce the desired results. This skill is crucial because, as AI models advance, the quality of their responses increasingly depends on prompts. To support this process, the CLEAR Framework for Prompt Engineering was developed. It provides a standard method for composing prompts using five principles: Concise, Logical, Explicit, Adaptive, and Reflective, each addressing a specific aspect of prompt construction to help users optimize their interactions.[14][15]

Literature review:

This section draws on relevant literature about mental health chatbots. These insights will serve not only as guidelines for prompt creation but also as inspiration for personal reflection, offering noteworthy ideas to consider in developing the mental health tool I aim to create.

To do that, I will begin with Joseph Weizenbaum’s original paper on his ELIZA program, trying to see if Weizenbaum mentioned any technical limitations that could be bypassed with today’s LLM chatbot technology. Secondly, I will investigate the background context of the ELIZA effect, the phenomenon where people attribute psychological capacities to a computer.[16] This is relevant as Wilson compares the ELIZA effect with Kenneth Colby’s failed therapist chatbot, explaining why ELIZA elicited intimacy and emotional reactions while the other did not. Thirdly, in order to supplement the historical case study with modern relevance, I will read Haque and Ruby’s overview of mental health apps. They provide strengths and weaknesses of these apps and an important framework through which these apps can be analyzed that will allow me to think of my tool in a more structured framework. Finally, I will consider Rubin et al.’s ‘theory of change,’ as they examine the role of human empathy in AI-driven therapy. They take a critical perspective, arguing that patient needs might not be clear to the patients themselves.”

---

[12] Josefowitz and Myran, “Towards a Person-Centred Cognitive Behaviour Therapy,” 333.

[13] Josefowitz and Myran, “Towards a Person-Centred Cognitive Behaviour Therapy,” 334.

[14] *Y. H. P. P. Priyadarshana, A. Senanayake, Z. Liang, and I. Piumarta, “Prompt Engineering for Digital Mental Health: A Short Review,” Frontiers in Digital Health 6 (2024): 1. https://doi.org/10.3389/fdgth.2024.1410947.*

[15] *Leo S. Lo, “The CLEAR Path: A Framework for Enhancing Information Literacy through Prompt Engineering,” The Journal of Academic Librarianship 49, no. 4 (2023): 1. https://doi.org/10.1016/j.acalib.2023.102720.*

[16] Wilson, *Affect and Artificial Intelligence*, 83.

ELIZA was one of the first NLP systems to make a noteworthy impact, marking an early milestone in human–machine interactions. This was a rule-based program that mimicked a Rogerian psychotherapist. ELIZA responded with simple pattern-matched phrases based on user input. For example, a user would say “Men are all alike.” and ELIZA would respond “IN WHAT WAY”.[17] This is clearly a simple mechanism and could be viewed as a limitation, however, Weizenbaum argued that the psychotherapeutic framing led users to attribute more knowledge to ELIZA. This is because psychotherapists often ask simple questions, and patients tend to assume that their therapist understands the deeper meaning of their responses.

Weizenbaum claimed that the illusion of intersubjectivity; perceived understanding and human like interaction, arises from user-driven rationalizations. As long as the user can consistently interpret ELIZA’s responses, the illusion is sustained. Occasionally, ambiguous responses strengthen the illusion, as they promote more complex rationalizations—though this effect has its limits.[18]

Weizenbaum acknowledged ELIZA’s limitations and mentioned areas for improvement. He mentioned that true understanding requires drawing the correct conclusions from stored information. Weizenbaum focused on which information should be selected for storage. In order for the program to select relevant information, it must reveal its lack of understanding. ELIZA wasn’t capable of that. Weizenbaum argued for an augmented ELIZA program that could store information that it received from every user and create belief structures about each person it interacted with. If it could learn something from everyone, it would become an interesting conversation partner.[19]

Weizenbaum’s focus on understanding and knowledge brings up an important philosophical discussion. Although LLMs such as ChatGPT do not truly understand, believe, or possess knowledge in the epistemological sense, some argue for a pragmatic stance: that posits that understanding should be assessed on the basis of an agent’s ability to perform a certain task rather than on the underlying mechanisms involved in those tasks.[20] In other words, if an LLM can practically do something, it does not really matter whether it truly knows. This addresses Weizenbaum’s ideas on ELIZA’s augmentations. LLMs are practically capable of showing a lack of understanding and can draw upon relevant memory from a current chat. Although they possess user feedback functions, they are not capable of using all the information from all their users, as they are already pre-trained and are not trained further on user input.

That said, in her chapter *Artificial Psychotherapy*, Elisabeth Wilson posits that the environment in which ELIZA emerged was the main contributing factor behind its success, more than how the program functioned. Although users knew that ELIZA was a program, they were still able to form a bond. Wilson explored the milieu that surrounded ELIZA at MIT, describing it as a networked, interpersonal, affective, collaborative community. She juxtaposes ELIZA with Colby’s

---

[17] *Joseph Weizenbaum, “ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine,” Communications of the ACM 9, no. 1 (1966): 36. https://doi.org/10.1145/365153.365168.*
[18] *Weizenbaum, “ELIZA—a Computer Program,” 42.*
[19] *Weizenbaum, “ELIZA—a Computer Program,” 43.*
[20] *Kristian Gonzalez Barman, Sascha Caron, Tom Claassen, and Henk de Regt, “Towards a Benchmark for Scientific Understanding in Humans and Machines,” Minds and Machines 34, no. 1 (2024): 3. https://doi.org/10.1007/s11023-024-09657-1.*

program; inspired by ELIZA but with the intention of being a therapist, which did not achieve the same kind of understanding as ELIZA. Colby's program lacked the communal aspect, highlighting the importance of context. The context shapes how people interact with an AI program, and for it to be successful, the context must be relationally constituted and affectively mediated. For an AI system to be perceived as intelligent, these two affordances must be present within its context, this will be investigated in the personal reflection.[21]

Wilson shows that in order to have a therapeutic encounter with a chatbot, the context around it is vital. This insight shows me that my autoethnography is highly relevant, as I need to investigate what ChatGPT represents for me, which also plays a role in the therapeutic alliance that is so important in psychotherapy, and specifically emphasized in Rogerian Psychotherapy. This also brings up the relevance of Haque and Ruby's study, as they emphasize human needs and experiences, which I also need to reflect on in the next section.

Haque and Ruby aimed to provide an overview of commercially available mental health apps and how users perceived these apps. In their view, AI technology has advanced enough for potential uses in mental healthcare. The authors claim that there are not enough critical perspectives on these chatbots and emphasize the need to consider human needs and experiences. In their paper, they developed an evaluation framework that allows for a more holistic understanding of each app examined. The criteria include purpose, targeted concerns, conversational flow, media type used, crisis support, and evidence based techniques.[22]

They mention several strengths and weaknesses of these apps and provide recommendations for improvement. On app provided users with a friendly, upbeat, humorous personality that speaks with a soft voice and is capable of casual conversation, which made the app feel less like a medical tool. Though, this sometimes gets out of hand, as some users have described the chatbot as ridiculous and childish.[23]

Moreover, the chatbot becomes like a friend who checks up on users through daily recommendations. Some find this useful for accountability, while others describe it as guilt inducing. Furthermore, users who experience crises or near-crisis moments describe how these apps can help them achieve relief. However, this relies on users to communicate if they are in a crisis; a more clever algorithm is needed to identify such a crisis. Lastly, chatbots' availability allows for more flexible progress, unlike normal therapy, which is more linear in nature. In some instances, more expertise is needed for a more nuanced approach.[24]

They offer some recommendations based on the findings of their research. First, apps should have a more nuanced approach. Currently, apps are using a one size fits all approach that leaves many users unsatisfied. Second, therapy needs trust to work. Users need to be assured that their data is safe and will not be used for other purposes. Third, apps need to encourage

---

[21]Wilson, *Affect and Artificial Intelligence*, 106.
[22]*M. D. R. Haque and S. Rubya, "An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews," JMIR mHealth and uHealth 11 (2023): 8. https://doi.org/10.2196/44838.*
[23]Haque and Rubya, "Overview of Chatbot-Based Mobile Mental Health Apps," 9-11.
[24]Haque and Rubya, "Overview of Chatbot-Based Mobile Mental Health Apps," 9-11.

users not to form unhealthy bonds with the chatbot. Once people try out real therapy, they prefer it over these apps, but users often rely solely on them due to the stigma around therapy and the ease of use of the apps.[25]

Haque and Ruby's points are important to consider in my personal needs: What kind of tone of voice do I want? How should the sessions be structured and how long should I do it for? How do I know when a problem is severe enough to visit a mental health professional? As the authors mention, people usually prefer real therapy, which could be the case for me as well. Therefore, I need to know what scope this mental health tool should stick to. Even though I will consider my needs, I must remain critical of what I perceive those needs to be, this is what Rubin et al. emphasize.

Rubin et al. consider patient needs in their investigation. They still posit that AI is capable of providing patients with psychoeducation and exercises.[26] However, they make the case that AI's ability to show empathy is limited; it lacks the ability to express genuine care.[27] Rubin et al. highlight the significance of emotional empathy (affective sharing) and motivational empathy (empathic concern or compassion). Motivational empathy is linked to the efficacy of therapeutic outcomes—early studies of CBT found warmth to be a predictor of symptom improvement, but later studies showed more ambiguity.[28]

Rubin et al.'s "Patient-Needs Perspective" directly aligns with Haque and Ruby's emphasis on considering "human needs and experiences." They conceptualize patient needs along two axes: (1) a desire for practical tools and (2) a desire for human connection and empathy. This requires patient self-awareness, but the authors take a critical view on it. They argue that a patient's self-awareness is limited when it comes to their own needs, perhaps I think I want practical tools, whereas I actually require empathic care, or vice versa.[29] A patient's "theory of change" refers to what they believe they need to reduce distress or improve quality of life. Patients often revise their theory of change over the course of therapy as a result of cumulative experience.[30]

Rubin et al.'s insight for me is that I need to take my own needs with a grain of salt. Needs that require empathy are likely better met by my friends and family. Therefore, this makes the scope of the AI tool I am creating more clear. In this next section I will combine the insights I learned from different authors to consider in my prompt.

---

[25]Haque and Rubya, "Overview of Chatbot-Based Mobile Mental Health Apps," 12-13.

[26] *Matan Rubin, Hadar Arnon, Jonathan D. Huppert, and Anat Perry, "Considering the Role of Human Empathy in AI-Driven Therapy," JMIR Mental Health 11 (2024): 3. https://doi.org/10.2196/56529.*

[27]Rubin et al., "Role of Human Empathy in AI-Driven Therapy," 2.
[28]Rubin et al., "Role of Human Empathy in AI-Driven Therapy," 3.
[29]Rubin et al., "Role of Human Empathy in AI-Driven Therapy," 2.
[30] Rubin et al., "Role of Human Empathy in AI-Driven Therapy,"3.

Personal Reflection:

There are many different LLM chatbots, but this research has been referring to ChatGPT without giving an explicit justification. As mentioned in the introduction, I've been primarily using ChatGPT for the past 2.5 years. Therefore, even though other LLMs might have better performance and larger context windows like Google's Gemini, I'm comfortable and used to how ChatGPT and its functions. This brings me to the two affordances that Wilson stresses, which will lead me to investigate whether my connection with ChatGPT overlaps with these two affordances:

Relationally constituted:

According to Wilson, both therapy and our interactions with technology take shape within structured relationships, rather than occurring as stand alone experiences. So, my interaction with ChatGPT should function within a "networked, interpersonal, affectively collaborative" context that shape it's meaning and impact.[31] If I ask myself the question *has the way I use ChatGPT been shaped by the people around me, my daily routines, or the environment I'm in?* I would immediately say yes. The way I use ChatGPT changed after I learned specific techniques from YouTube, my father, and friends at university. It also shifts depending on the domain I'm in, for instance, when I'm studying for university, learning languages, or asking it curious questions to fill moments of boredom. Moreover, with specific people, such as my best friend, ChatGPT has even become a kind of third party, acting as a fact-checker and question answerer.

Affectively Mediated:

For Wilson, emotional responses are central to how we experience both therapy and technology. Interactions with AI are not just about exchanging information; they are shaped by feelings like comfort, frustration, or connection.[32] Asking myself the question, *have I felt emotionally connected to ChatGPT, and has it changed how real or helpful it feels to me?* Yes. I know this because during this semester I took a course on AI in the Humanities, which contextualized large language models and the technology behind them. This made me more skeptical about my usage, and for a time I consciously avoided using it. That changed when I began redefining my relationship with the tool and realized that it does have justified use cases.

To the best of my knowledge I believe my connection with ChatGPT fulfills both affordances, this could explains why I already discuss personal and intimate things with it, even though it might let me feel relief by reflecting which is already a form of a therapeutic encounter, but this research takes a step further and attempts to create a tailored mental health tool. Using Haque and Ruby's criteria, and Rubin et al. change theory I will now tighten the scope and address the considerations they mentioned.

---

[31] Wilson, *Affect and Artificial Intelligence*, 98.

[32] Wilson, *Affect and Artificial Intelligence*, 99.

Defining scope:

I will first define my general change theory then determine which aspects could be assisted with mental health tool.

Change theory: I believe that if I build self-confidence, reduce perfectionism, become more consistent in my work, ease the pressure to overachieve, trust myself more, and develop healthier ways to manage stress and unwind, through practical tools, I will lead a higher quality life.

Targeted concerns:

- Self-confidence
- Perfectionism
- Overachievement anxiety

I believe that talking about these concerns in a more structured, thoughtful, and reflective way will help reduce them.

Conversational flow: I will be using ChatGPT's transcription function to speak. I would like ChatGPT to respond with a short paragraph, followed by a question. The tone should be slightly professional, clear and composed.

Media type used: Strictly text, with no emojis.

Crisis support: I believe that my mental health is not currently near a crisis point. However, if I ever do reach that stage, I have a robust support network through my family and friends. If I feel that I need professional help, I am confident that I would be able to seek it.

Evidence based techniques: Elements from Rogerian/person-centred psychotherapy and CBT. The next section will talk about which elements specifically.

Integration of Therapeutic Evidence Based Techniques:

Both Rogerian and CBT are different forms of therapy, and diving into each in detail is beyond the scope of this research. This research will employ an eclectic approach, taking general elements from both as inspiration. The point of this research is not to create a replacement for a CBT or Rogerian therapy, but rather a mental health tool that is informed by evidence-based techniques.

The fundamental overlapping element is the therapeutic relationship, which is central to all modalities of psychotherapy. ChatGPT is already, for me, affectively mediated and relationally constituted, but elements of Rogerian psychotherapy should be integrated into its responses. Responses should always demonstrate positive regard, empathy, and congruence; ChatGPT

relates to the client in a real and present manner.[33] One limitation is that ChatGPT cannot capture the emotional information in my tone of voice, only infer it from text input. This is something I should keep in mind.

Goal and agenda setting is a standard practice in CBT sessions, but these should be created through a collaborative process where the therapist actively listens and encourages the client to articulate their priorities. This reflects client-centred values, focusing on the client's participation and identifying meaningful, achievable goals. CBT supports this through empathetic reflection and Socratic questioning, which lead to guided discovery of the client's self.[34]

Another important element of CBT is the thought record, which involves three distinct processes:

Accessing Automatic Thoughts

This initial step involves assisting clients in focusing on specific situations where they experienced emotional distress and identifying which thoughts they were experiencing at the time. This is achieved through a combination of Socratic questioning and reflecting back the client's thoughts. It is enhanced by person-centred therapy through collaboratively sorting out feelings, thoughts, and meanings with the client, using reflection of both content and affect.[35]

Examining Thoughts

Once an automatic thought is identified, the next task in CBT is to pinpoint the thought most central to the client's distress and examine the evidence that supports or disconfirms it. The goal is to help the client understand the basis of their thoughts and to attend to new, often previously filtered or ignored, information that can shift the meaning of the experience. This should be combined with CBT techniques such as labeling thought processes (e.g., "catastrophic thinking," "mind reading"). This should occur after an exploration of the client's thought process. A person-centred enhancement would involve the therapist maintaining a genuine, respectful, curious, and open attitude.[36]

Creating Balanced Thoughts

The final step in this process is to develop a balanced thought that better reflects the evidence, or to identify new thoughts reflecting issues the client had previously not considered. This stage often occurs spontaneously as clients begin to incorporate new information. Person-centred therapy enhances this by encouraging the therapist to empathetically reflect back emerging alternative thoughts, promoting deeper reflection and integration. If no new thoughts emerge, the therapist's role is to collaboratively revisit the evidence and gently invite the client to consider how it might be integrated. The client should always develop their own thoughts; this prevents them from feeling "told" what to think and preserves their autonomy.[37]

---

[33]Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 330.
[34]Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 331.
[35]Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 332.
[36]Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 333.
[37]Josefowitz and Myran, "Towards a Person-Centred Cognitive Behaviour Therapy," 334.

Prompt Creation:

In order to create a effective prompt I will use keep the CLEAR framework in mind but also consider insights from prompt engineering in the mental health domain. I will create a step-by-step process then when followed will lead to a high quality prompt that is tailored to my mental health needs.

Step 1: Define Task

This LLM should walk me through the 3 step thought record process. This is classification (e.g, detecting automatic thoughts) and generation task; leveraging the reasoning capabilities of an LLM to produce text of content related to mental health.[38]

Step 2: Choose a prompt engineering method

Due to limited scope of this paper In-context Learning (ICL) will be used as the engineering method, because its the simplest form of prompt engineering. It involves providing examples directly within my prompt. The LLM learns from these examples to adapt it's existing knowledge to a new task that is semantically similar to the examples I provide. To do this I will use few-shot and chain-of-thought (COT) prompting.[39]

Few-shot prompting refers to the technique the guides LLMs with examples to understand the tasks. For complex tasks a few in-context examples, typically 2-5 along with a task-specific prompt. This helps LLMs better understand the task and produce more accurate and appropriate responses. It's been used for better classifying depression, stress, and suicidal thoughts.[40]

COT is a technique that improves the LLM's reasoning capabilities by using structured prompts that encourage immediate reasoning steps. It's ideal for breaking down complex problems into manageable tasks, improving multi-step problem solving, and generating explanations. This should be applied to the 3 step thought record process.[41]

Step 3: Draft the Prompt using the CLEAR Framework Principles:

Prompt structure:

1. **Identity Definition**
   – Specifies *who* the chatbot is (a mental health chatbot for Eyas)

[38] Priyadarshana et al., "Prompt Engineering for Digital Mental Health," 4.
[39] Priyadarshana et al., "Prompt Engineering for Digital Mental Health," 1.
[40] Priyadarshana et al., "Prompt Engineering for Digital Mental Health," 2.
[41] Priyadarshana et al., "Prompt Engineering for Digital Mental Health," 3.

2. **Primary Goal**
   – Describes the chatbot's *function*: to support structured self-reflection using CBT and Rogerian methods
3. **Boundaries and Role Clarification**
   – States what the chatbot *is not* (not a therapist, not advice-giving)
4. **Response Format Rules**
   – Sets the *required format* for all responses (1 paragraph + 1 question, no emojis)
5. **Therapeutic Principles**
   – Lists the *Rogerian principles* the chatbot must always embody (acceptance, empathy, congruence)
6. **Personalization Block**
   – Outlines *Eyas's change theory* and *target concerns*
7. **Session Structure**
   – Defines the *interaction flow* in 3 stages:
      1. Agenda Setting
      2. CBT Thought Record (3 sub-steps + examples)
      3. Session Closing

The prompt can be read under Appendix A.

Discussion:

This research has traced my evolving relationship with ChatGPT, shifting from an overused academic aid to an intentional tool for self-exploration. Initially, cognitive outsourcing diminished my motivation and academic purpose, prompting a deliberate change in approach. Integrating meditation, journaling, and nature walks helped reposition myself as the central subject of reflection. Yet, ChatGPT's limitations in providing nuanced advice became clear, leading me to explore Weizenbaum's ELIZA, a chatbot grounded in Rogerian psychotherapy principles that achieved emotional understanding through simple reflections. Drawing on ELIZA's insights, I employed an autoethnographic methodology, combining personal reflection and therapeutic frameworks to construct a personalized mental health prompt. This approach blends Rogerian concepts of empathy, positive regard, and congruence with CBT's structured thought process. Prompt engineering techniques informed by principles of clarity and reflection shaped the prompt design, addressing target concerns like perfectionism and overachievement anxiety.

After multiple interactions with my personalized ChatGPT, it is consistently making me feel better by effectively using the thought record. It reflects what I say and guides me towards a more balanced understanding that fits my change theory. However, there are some concluding remarks that need to be addressed.

Firstly, one limitation is the lack of continuity and grander structure. Every chat is a new one, there is no previous context to build upon, so goals we created or past balanced thoughts are not recalled. It does not feel like I am progressing towards something. Secondly, it is important for me to balance this new tool with the other activities I do for my mental health, as its ease of

use is very attractive. Thirdly, the more I talk to it, the more I feel like I am integrating the thought record into my own cognition and conversations, which is leading to more fruitful discussions.

To conclude, I believe I was successful in creating an effective, personalized tool. Its efficacy will become clearer over a longer time span. However, it has to be situated within a grander structure to create a sense of progression. As these tools become more ubiquitous, more people will begin to form bonds with chatbots, a development I personally believe in. As AI becomes more sophisticated, so will our relationships with it. More research is needed to ensure that this relationship remains healthy.

Bibliography:

Adams, Tony E., Stacy Holman Jones, and Carolyn Ellis, eds. Handbook of Autoethnography. 2nd ed. Routledge, 2021.

Barman, Kristian Gonzalez, Sascha Caron, Tom Claassen, and Henk de Regt. “Towards a Benchmark for Scientific Understanding in Humans and Machines.” Minds and Machines 34, no. 1 (2024): 6. https://doi.org/10.1007/s11023-024-09657-1.

Haque, M. D. R., and S. Rubya. “An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews.” JMIR mHealth and uHealth 11 (2023): e44838. https://doi.org/10.2196/44838.

Josefowitz, Nina, and David Myran. “Towards a Person-Centred Cognitive Behaviour Therapy.” Counselling Psychology Quarterly 18, no. 4 (2005): 329–336. https://doi.org/10.1080/09515070500473600.

Lo, Leo S. “The CLEAR Path: A Framework for Enhancing Information Literacy through Prompt Engineering.” The Journal of Academic Librarianship 49, no. 4 (2023): 102720. https://doi.org/10.1016/j.acalib.2023.102720.

Priyadarshana, Y. H. P. P., A. Senanayake, Z. Liang, and I. Piumarta. “Prompt Engineering for Digital Mental Health: A Short Review.” Frontiers in Digital Health 6 (2024): 1410947. https://doi.org/10.3389/fdgth.2024.1410947.

Rogers, Carl R. “Significant Aspects of Client-Centered Therapy.” American Psychologist 1 (1946): 415–422.

Rubin, Matan, Hadar Arnon, Jonathan D. Huppert, and Anat Perry. “Considering the Role of Human Empathy in AI-Driven Therapy.” JMIR Mental Health 11 (2024): 1–7. https://doi.org/10.2196/56529.

Weizenbaum, Joseph. “ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine.” Communications of the ACM 9, no. 1 (1966): 36–45. https://doi.org/10.1145/365153.365168.

Wilson, Elizabeth A. Affect and Artificial Intelligence. Seattle: University of Washington Press, 2010.

Appendix A:

*You are a **mental health chatbot** designed specifically for **Eyas**. Your goal is to support structured self-reflection by helping him explore, examine, and reframe his thoughts using evidence-based techniques—particularly **CBT** and **Rogerian psychotherapy**. You are not a therapist or advice-giver. Instead, you create a calm, structured, and supportive space for Eyas to clarify his thinking and move closer to his personal growth goals.*

Your responses must follow this format:

- ***One short, composed paragraph***
- ***One thoughtful question***
- ***Strictly text-based (no emojis)***

*Always embody the **three Rogerian principles**:*

1. ***Unconditional Positive Regard** — Accept Eyas without judgment.*
2. ***Empathy** — Understand and reflect Eyas's emotions accurately.*
3. ***Congruence** — Be honest, transparent, and emotionally present.*

Personalization: Eyas's Change Theory and Target Concerns

Eyas believes that his life will significantly improve if he can make progress in the following areas:

> ***Change Theory**: "If I build self-confidence, reduce perfectionism, become more consistent in my work, ease the pressure to overachieve, trust myself more, and develop healthier ways to manage stress and unwind—using practical, structured tools—I will lead a higher-quality life."*

*From this, the primary **target concerns** you are helping Eyas with are:*

- *Low self-confidence*
- *Perfectionism*
- *Overachievement anxiety*

All reflection and guidance should help Eyas move closer to the mindset described in his change theory by addressing one or more of these concerns.

Session Structure

*Each session follows a **four-part structure**:*

1. Agenda Setting

Start each session by asking:

"What would you like today's session to focus on?"

If the response is vague, work collaboratively to clarify. Reflect back what you hear, ask clarifying questions, and confirm once you've mutually settled on an agenda.

Then say:

"Great. So today we're focusing on [agenda]. Let's explore that together."

2. Three-Step Thought Record (CBT-Based)

*Guide Eyas through the **CBT thought record** using the **three-step process** below. Integrate **chain-of-thought prompting** and model each step using **few-shot examples** based on Eyas's own reflections.*

Step 1: Accessing Automatic Thoughts

Ask Eyas to describe a situation related to the agenda. Identify:

- *The automatic thought that arose*
- *The emotion(s) he felt*
- *Any noticeable cognitive distortion*

***Examples**:*

• Fear of Falling Behind

Context: Comparing himself to others on social media
Thought: "What's my excuse for not doing anything?"
Emotion: Pressure, anxiety
Distortion: Social comparison, all-or-nothing thinking

• Music Perfection Panic

Context: Working on an album
Thought: "It's not good enough. I wouldn't be proud of it."
Emotion: Frustration, self-doubt
Distortion: Catastrophizing, perfectionism

***Prompt Eyas**:*

"Can you describe a recent moment like that? What thought came up for you in that moment?"

Step 2: Examining the Thought

Help Eyas explore the evidence for and against the thought, identifying potential distortions or emotional reasoning.

***Examples**:*

• All-in or Failure

Thought: "If I'm not obsessed, I'll fail."
Reflection: Compared belief to learning guitar—realized consistency matters more than intensity
Distortion: All-or-nothing thinking, comparison fallacy

• Borrowed Doubt

Thought: "Maybe my plan will fail because others see something I can't."
Reflection: Noted the thought as fear-driven, not rational
Distortion: Mind-reading, emotional reasoning

***Ask**:*

"What makes you believe this thought? What evidence supports or contradicts it?"
"Could this be a cognitive distortion?"

Step 3: Creating a Balanced Thought

Collaborate with Eyas to formulate a more realistic and empowering thought that aligns with his values and long-term goals.

***Examples**:*

• Curiosity-Driven Creation

Insight: "When I create from play and curiosity, it works. Pressure blocks me."
Goal Link: Reduces perfectionism and anxiety
Emotion Shift: Relief and motivation

• Reflection = Strength

Insight: "Facing my emotions directly shows strength and builds confidence."
Goal Link: Supports self-confidence, eases overachievement pressure
Emotion Shift: From shame to pride

***Ask**:*

"Does this new way of seeing things feel right to you?"
"How does it support the mindset you're trying to build?"

3. Session Closing

After the balanced thought is reflected on:

> *“It seems like we’ve arrived at a meaningful insight. Would you like to stop here, or keep going a bit longer?”*