KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Technology

First cycle, 15 credits

Prompt engineering and its usability to improve modern psychology chatbots

School of Electrical Engineering and Computer Science

ISAK NORDGREN
GUSTAF SVENSSON

Stockholm, Sweden 2023

PROMPT ENGINEERING AND ITS USABILITY TO IMPROVE MODERN PSYCHOLOGY CHATBOTS

PROMPT ENGINEERING OCH DESS ANVÄNDBARHET FÖR ATT FÖRBÄTTRA PSYKOLOGICHATBOTTAR

Isak J. Nordgren, L. Gustaf E. Svensson

***Abstract*—As advancements in chatbots and Large Language Models (LLMs) such as GPT-3.5 and GPT-4 continue, their applications in diverse fields, including psychology, expand. This study investigates the effectiveness of LLMs optimized through prompt engineering, aiming to enhance their performance in psychological applications. To this end, two distinct versions of a GPT-3.5-based chatbot were developed: a version similar to the base model, and a version equipped with a more extensive system prompt detailing expected behavior.**

A panel of professional psychologists evaluated these models based on a predetermined set of questions, providing insight into their potential future use as psychological tools. Our results indicate that an overly prescriptive system prompt can unintentionally limit the versatility of the chatbot, making a careful balance in instruction specificity essential. Furthermore, while our study suggests that current LLMs such as GPT-3.5 are not capable of fully replacing human psychologists, they can provide valuable assistance in tasks such as basic question answering, consolation and validation, and triage.

These findings provide a foundation for future research into the effective integration of LLMs in psychology and contribute valuable insights into the promising field of AI-assisted psychological services.

***Sammanfattning*—I takt med att framstegen inom chatbots och stora språkmodeller (LLMs) som GPT-3.5 och GPT-4 fortsätter utvidgas deras potentiella tillämpningar inom olika områden, inklusive psykologi. Denna studie undersöker effektiviteten av LLMs optimerade genom prompt engineering, med målet att förbättra deras prestanda inom psykologiska tillämpningar. I detta syfte utvecklades två distinkta versioner av en chatbot baserad på GPT-3.5: en version som liknar bas-modellen, och en version utrustad med en mer omfattande systemprompt som detaljerar förväntat beteende.**

En panel av professionella psykologer utvärderade dessa modeller baserat på en förbestämd uppsättning frågor, vilket ger inblick i deras potentiella framtida användning som psykologiska verktyg. Våra resultat tyder på att en överdrivet beskrivande systemprompt kan ofrivilligt begränsa chatbotens mångsidighet, vilket kräver en noggrann balans i specificiteten av prompten. Vidare antyder vår studie att nuvarande LLMs som GPT-3.5 inte kan ersätta mänskliga psykologer helt och hållet, men att de kan ge värdefull hjälp i uppgifter som grundläggande frågebesvaring, tröst och bekräftelse, samt triage.

Dessa resultat ger en grund för framtida forskning om effektiv integration av LLMs inom psykologi och bidrar med värdefulla insikter till det lovande fältet av AI-assisterade psykologtjänster.

***Index Terms*—Large Language Models, LLM, GPT, GPT-3.5, GPT-4, chatbots, psychology, prompt engineering**

I. INTRODUCTION

MENTAL ILLNESS is estimated to be one of, if not the leading global burden in regards to the health of the public. The disease burden has been estimated to make up 32.4% of years lived with disability, having overtaken both cardiovascular and circulatory disease [1].

According to Bendig et al. [2], studies have shown that a contributing factor to this problem is that many people diagnosed with a mental disorder are reluctant to seek professional help. According to the same study, some reasons for this might be the stigmatization of mental illness and professional help, bad experiences and negative attitudes against professionals and the treatment they offer, lack of insight into their illness, as well as shortcomings regarding the availability of time and place.

Chatbots are a tool that can be used to make therapy more available [2], and, hence, could be used to reduce the number of people who do not seek help because of the inaccessibility of mental health services.

Another study by Cameron et al. [3] found that people find it easy and safe to talk with chatbots. This indicates that AI therapists can be used to lower the threshold for seeking help for mental illness.

Bending et al.'s [2] work concludes that there are indications that chatbots can be useful in clinical psychology, but there is yet more research that needs to be made before we can draw definitive conclusions about their effectiveness and suitability as a clinical tool.

A. Recent developments in conversational agents

With the introduction of ChatGPT, chatbots have attracted considerable attention globally, impressing many with their ability to respond intelligently to a broad range of topics. [4].

This capability arises from ChatGPT being an open-domain chatbots, a category of chatbot capable of generating responses on virtually any subject [5]. This capability is made possible by training these chatbots on extensive text datasets, thus equipping the neural network with wide-ranging knowledge. Subsequently, the chatbot uses language models to generate responses based on this acquired knowledge [6], [7].

To effectively describe how chatbots, like ChatGPT, perform so well, it is best to understand the key technical components of their architecture. These include the Encoder-Decoder architecture, attention mechanism, and a specific model architecture known as the Transformer [8]. This Transformer-based

architecture forms the foundation of the Generative Pretrained Transformer (GPT) series of Large Language Models (LLMs) upon which ChatGPT is built.

This will be detailed in Section II, but briefly, the Transformer model uses an encoder to interpret the input text and a decoder to generate responses, with an attention mechanism to determine which parts of the input are most relevant to each step of the response. This model, combined with extensive training data, enables the chatbot to produce human-like responses which appear to be conscious and aware of the context provided by the user.

B. Implications of the developments

The introduction of the Transformer model and its self-attention mechanism marked a significant milestone in natural language processing, laying the groundwork for the impressive performance of chatbots like ChatGPT. Enabled by self-attention, GPT can generate contextually relevant and grammatically consistent responses [8], hardening its position as a powerful tool for diverse natural language tasks.

Nonetheless, training these models includes some problems. It requires not only a substantial amount of data but also deep technical expertise and significant computational power, all of which translates into a considerable investment of time and resources [9]. This investment can act as a barrier for businesses, institutions, and other organizations wishing to create their own chatbot.

However, the rise of pre-trained LLMs like GPT, on which ChatGPT is based, has mitigated these challenges considerably. Because of their open-domain nature, these models can be repurposed for a wide range of applications [9]. This versatility has made the utilization of chatbots a more cost-effective and accessible solution for a multitude of tasks, ranging from customer service and e-commerce to health support and therapy.

C. Leveraging System Prompts

Language models like ChatGPT can be guided in their behavior using *system prompts* or *system messages*, which are provided at the onset of the conversation [10]. These instructions could direct the model to emulate a Linux terminal, mimic a toddler, or in our context, behave as an AI Psychologist.

Despite ChatGPT being an advanced and capable system [4], it remains uncertain whether the base version of GPT-based LLMs, or one that includes a simple system prompt like ”You are an AI psychologist”, is the optimal configuration for these models. Moreover, it is unclear whether a more detailed or rigorous system prompt would result in a more effective AI Psychologist.

This study seeks to explore whether LLMs such as GPT-4 can be enhanced beyond their base capabilities for specific tasks. Our focus is on the realm of online psychology, but we also consider the implications for broader applications. We aim to investigate the effectiveness of different system prompts and their potential to augment the performance of these models in specified roles.

D. Project context

Our investigation is part of a larger project, which aims to create a chatbot specialized in Cognitive Behavioral Therapy (CBT). The goal is to use this chatbot in conjunction with online CBT programs. This project is headed by Birger Moëll, a licensed psychologist and Ph.D. student at KTH. The project team includes us and two other groups, who are also working on their bachelor’s theses within the same subject. Additionally, a group of psychologists and psychology students are involved, providing assistance on psychology-related matters.

Moëll conducted a pilot study before our study, which tested the feasibility of the project. The findings suggested that chatbots based on GPT have significant potential in the field of psychology. The majority of participants found the bot’s interactions engaging and human-like, despite some criticisms of it lacking empathy. Participants greatly appreciated the bot’s expertise in the field, and its ability to deliver quality responses. In summary, this pilot study demonstrated that it is possible to develop a conversational agent that is both engaging and useful in the field of psychology.

E. Project goal

The goal of the project is to learn if existing language models can be modified to perform better and be more suitable in psychology, as well as compare how these language models compare to a human psychologist. We also try to find out what the future role chatbots can take within the field of psychology.

This goal is an extension of the broader goal of improving internet-based CBT services, automating counseling, and making it accessible to more people.

Furthermore, there have been many general advancements in research concerning chatbots, including those specifically made for psychology. However, more research is needed if they are to be used efficiently in practice, especially on the conceptual side [2]. We argue that this is why this paper is of interest since it instead of investigating *if* chatbots are viable investigates how a modified version compares to a base version as well as to a human, and how these can be used within psychology.

F. Impact on Society and Ethics

According to WHO, one in seven adolescents suffer from some form of mental illness [11], and as already said, an identified problem with how we treat mental illness today is that there is a deficit in psychologists that can treat those who need treatment [2]. By developing chatbots specifically for the treatment of mental illness the psychologists could be relieved of some of their patients reducing their stress as well as making mental health treatment more accessible for all patients.

By making treatment for mental health problems more accessible we work towards the third of the Sustainable Development Goals (SDGs), *Good health and well-being*. More specifically it is the fourth sub-part of the goal *Reduce mortality from non-communicable diseases and promote mental health* that is relevant to our project. The mental health part of SDG 3.4 is measured through suicide rate per 100.000 population.

There have been many research projects in the last decade within different areas of chatbots. The conclusion we reached was that despite a multitude of technological aspects of chatbots being researched, the societal or business aspect of chatbots seemed to lack research. We will hence back our primary research question regarding chatbots within therapy and psychology with an analysis and discussion about the future of LLM-based psychologists.

According to the Swedish Ethical Review Act (2003:460) §4, an ethical review of a research project is required if the research aims at affecting the subject either physically or psychologically [12]. In order to not be required to apply for a grant by the Swedish Ethical Review Authority, only students in psychology or legitimized psychologists will be used to test our chatbots. The study will focus on the perceived experience of using each chatbot, not their capabilities within the field of psychology.

G. Scientific Question

1) Computer Science: The main focus of this project is computer science and thus the main research question of the project will be anchored in computer science. As stated above the project is also included in a larger union of research projects focused on the application of chatbots within psychology and therapy. We focus on how existing LLMs can be modified using prompt engineering of system prompts, and have therefore formulated these research questions:

a How does an extended version of a Large Language Model like GPT-3.5 and GPT-4 compare to a base version of the models when designed to be used within therapy and psychology?
b What conclusions can be drawn from this comparison?

2) Industrial Engineering and Management: As stated above, the chatbot applications within CBT are rather unexplored outside the scope of Computer Science. This project will thus contain research within Industrial Engineering and Management (IEM) in parallel to the research questions within the computer science scope.

The two research questions we pose in regard to this are:

c What could be the future role of chatbots and LLMs within the field of psychology?
d How does the integration of chatbots impact the work system of psychological therapy in terms of its key components?

Research question d is built upon the System Theory-framework presented by Alter in [13].

II. Theory

This segment serves the purpose to give the reader an introduction to the theoretical aspects of Large Language Models, the technological aspects deemed critical for our study, the frameworks used to evaluate the project, as well as an introduction to the System Theory Approach to analyzing organizations and workflows.

A. Large Language Models

A Large Language Model is a type of artificial intelligence that utilizes Natural Language Processing, as well as other machine learning techniques, to understand, generate, and interact using human language.

Large Language Models are typically built using an artificial neural network, something called the Transformer [8], which is described thoroughly below.

An example of an implementation of a transformer-based LLM is the already-mentioned GPT series, which GPT-4, as of the writing of this paper, is the latest in the series[14]. These models are trained on large corpora of text data from the internet, allowing them to generate human-like text based on given prompts [14].

Large Language Models excel at various tasks, including but not limited to question-answering, translation, and summarizing long texts [15]. Due to their exposure to diverse text data, these models exhibit a degree of 'understanding' of context, style, and content. However, LLMs have limitations and risks, since they lack an inherent understanding of the world and rely only on patterns learned from data. They can sometimes produce plausible-sounding but incorrect or nonsensical answers [15], and can therefore be considered to have low reliability [16] and struggle with complex tasks that require deep understanding and nuanced judgment [17]. In addition, because of their training on data from the internet [14], they can perpetuate biases present in their training data. Further, LLMs can sometimes, instead of making conversation, follow what it interprets as instructions by the user blindly, making their responses misaligned with what in fact is true and factual [15]. Thus, while they are powerful tools, they must be used with understanding and responsibility.

These limitations underline the need for carefulness and continual improvements in the deployment of chatbots, especially in sensitive areas like psychology or healthcare.

However, despite the challenges, the potential of Large Language Models in a wide array of applications, from education to entertainment to psychology, is significant and the subject of ongoing research.

B. Recent progress in text generation for LLMs

Advancements in the field of Natural Language Processing, specifically Large Language Models like the GPT series, have impacted various domains, including programming, education, and customer service. As society gets more digital, understanding these AI systems become increasingly relevant. This section aims to describe the core architectures and mechanisms that enable these kinds of models to function, bringing light to strengths, limitations, and the role they play in current research.

In our view, it is essential to understand the different components in order to be able to design a well-functioning chatbot. It also shows how complex it can be to create your own chatbot, which connects to the mission of this project: Being able to create a chatbot that is well-adapted to be used in psychology using system prompts to alter a GPT model.

1) Encode-Decoder Architecture: A major development within Large Language Models is the Encoder-Decoder Architecture. This architecture is commonly used in natural language processing applications and is a key feature of chatbots and Large Language Models like ChatGPT. It includes two primary parts - the encoder and the decoder.

The encoder's role is to interpret the input, in this case, plain text. It processes the input data and creates a contextual representation of the text in vector form, which can be thought of as a description of the input in a way that the model can understand.

The decoder uses this abstract representation to generate a response. The connection between the encoder and decoder allows the decoder to leverage the contextual information that the encoder's output contains to produce its output [18].

2) The Attention Mechanism: A crucial advancement within neural network design is Attention since it allows models to focus on relevant parts of the input when generating the output. In the context of language models, it allows the model to pay more 'attention' to important words or phrases when generating a response.

In an Encoder-Decoder architecture, attention helps the decoder focus on the most important parts of the encoder's output when generating its response. This is because the decoder does not just use the output of the encoder blindly, but also *attends* to different intermediate states of the encoder's process of encoding the input. This means that it can better capture all the necessary information, which can be especially useful for long input sequences [8], [19].

C. The Transformer Architecture

Transformer models, like GPT, take the concept of attention even further with a design known as *self-attention* or *transformer attention*. These models enable for the removal of the traditional separation of encoders and decoders since they process the input and output sequences concurrently in a series of self-attention layers.

In a Transformer model, each token in the input sequence can attend to every other token, using a mechanism called self-attention, to capture complex relationships within the text. This design effectively aids many language tasks by grasping the context of a token within a sentence, paragraph, or a sequence of a certain length [8]. However, while a transformer can handle long-range dependencies, it lacks an explicit model of temporal sequencing. Its "understanding" is pattern-based, driven by the learned training data, and not the same as human comprehension.

D. Prompt Engineering

To begin with, a *prompt* is a text string with natural language instructions that is given to an LLM to customize, specialize or improve its abilities [20]. They can for example be used to set rules and boundaries for an LLM, tell it to only answer in a specific format, or mimic the style of a famous author.

Prompt Engineering is the art of designing a prompt that programs an LLM to act in a specific manner [20] since some prompt are more easily interpreted by the model and hence gives better results [16].

E. A prompt pattern catalog

White et al. [20] introduced a framework consisting of multiple patterns which prompts can follow depending on the purpose of the prompt. They document each pattern by giving it a name and classification, describing the intent and context, the motivation, and the structure and key ideas. They also include a summary of the pros and cons of each pattern. These are designed to operate in the same way as software patterns so that recurring problems that occur when prompting can be systematically solved and eliminated [20].

The prompt patterns that we used in this paper are described in the following sections.

1) The Persona Pattern: This pattern is used to make the LLM design answers that correspond with those that a person with a specific perspective would give. The motivation to using this pattern might be that the user either does not know, or can not describe in detail how the LLM should behave, but can give an example of someone who behaves similarly.

2) The Template Pattern: This pattern can be applied if the user wishes the response of the chatbot follows a specific template. This can, for example, be useful if the answer of the chatbot will be processed by a computer, or if the user wants the chatbot to generate text in specific formats, such as JSON.

F. The Godspeed Questionnaire

When developing and testing AI or robots, the issue regarding user testing is one that the developer must decide on early. There are multiple ways of testing, one being an observation of the user during the demo, either through direct observation or through recording the test. The other is to let the user fill out a survey after the demo, answering a predetermined set of questions. There are positive and negative aspects to both of these. The main positive aspect of direct observation is that the developer can identify the user's behavior and reaction to the AI while the AI is tested. Measurements such as heart rate and skin conductivity have previously been used to measure the user's arousal to the AI in real-time. The major flaw of this kind of testing is that the measurements give no indication of whether the user is for example frustrated or satisfied. In order to get a better understanding of the user's experience a survey might be used. The idea is to let the user first test the AI and let the user answer a set of questions afterward. This might seem good but a major flaw of this type of study is that users might be biased in their response since the users need to reflect on their experience first after the experience is over [21].

In order to establish a baseline for good questionnaires with regard to Human-Robot Interaction, a series of questionnaires called Godspeed was published. The idea of the Godspeed questionnaire is to let the user answer questions based on a scale from one to five with regards to the user's impression of the robot [21]. For example:

Please rate your impression of the robot on these scales
Machine-like 1 2 3 4 5 Human-like

Worth noting is that, if for example two bots are similar, the interpretation of the questionnaire answers might be tough on

a small test audience. Therefore the number of participants in the study might need to be increased. Furthermore, the prior experience the user's got with robots might influence the answers to certain questions. And that the more humans as a whole get used to conversing with robots, the more likely they might be to recognize flaws in robots and therefore give answers weighted more towards the scale of machine-like [21].

G. The Session Rating Scale

The Session Rating Scale (SRS) is an instrument in psychotherapy and counseling used to assess the therapeutic alliance, meaning the collaborative relationship between a therapist and a client, from the perspective of the client. This framework aims to enhance the efficiency of therapy by evaluating the therapeutic interventions based on feedback from the client, thereby centering the therapy process on the client's unique experiences and perceptions [22].

The SRS consists of four items that the client fills out at the end of each therapy session. The assessment serves to measure the client's view of the session and provide insight for the therapist into the client's perspective. The four domains that the SRS consists of are [22]:

1) Relationship: The client's feelings about the therapeutic relationship, asking if they felt heard, understood, and respected during the session.
2) Goals and Topics: If the therapist worked on and talked about the issues and goals that the client wanted to focus on.
3) Approach or Method: The client's feeling of how the therapeutic approach used fitted their own preferences.
4) Overall: If the overall session felt right for the client and met their needs.

Each item is rated on a scale (which in our case was 5-point), which provides a quantitative measure of the therapeutic alliance. The scores help make it clearer where breaches and misalignment in the therapeutic relationship can be found and thus measures how good of a fit the therapeutic session was in the eyes of the client [22].

Note that the SRS originally is designed to be used by psychologists and therapists to adapt their approaches toward counseling. In this study, we instead use it as a way of evaluating the quality of the therapeutic alliance after a session.

H. The System Theory's Approach to Assessing Organizations and Markets

Participants within a field can be construed as components of a system, working in harmony to generate output. This output can be the result of either a specific or an unspecified input stream. Inputs within a work system can simultaneously be interpreted as physical resources or hardware, and as data funneled through the different technologies inherent to the system [13].

Within system theory, a system is constructed out of the following parts, ***Customers*** - all who receive the systems products or services without performing any work that favor the system itself, ***Products/Services*** - everything that the system produces to its customers, ***Processes and Activities*** - what occurs within the system to produce the product/service, ***Participants*** - all who work within the system, ***Information*** - all information created or used within the system, and ***Technologies*** - tools and automated agents the system requires in order to function [13].

In Section IV we describe how psychological therapy can be described as a work system, and what the components are.

III. Literature Review

Despite the fast advancements in Large Language Models and text generation, it is important to have a balanced perspective. Therefore this section will aim to explore the current limitations and challenges with the reliability of these AI models, before moving on to the potential and considerations of these models in psychological therapy, and their implications from a Human-Computer Interaction (HCI) perspective. We also describe a few HCI principles from previous research about chatbots.

A. Prompt engineering for chatbots within health care

Prompt engineering is quite a young subject of research and specifically within health care, where there are only roughly 300 submitted papers at the time of writing, according to the writers of [23].

In a research paper by Kumar et. al. [24], different prompts to GPT-3 were analyzed based on user conversations. The users spoke about mental well-being for roughly 5 minutes and were then tasked with answering a series of questions. In order to engineer the chatbot, the developers set up three modifiers, *identity*, *intent*, and *behavior*. They found that when changing these three pre-defined behaviors, the chatbot expressed itself very differently and focused on different parts of the conversation.

In a study by Wang et. al., two approaches to prompting are introduced, manual and automated prompting [23]. Manual prompting is described as manually creating prompts with the assistance of an expert within the specific field whereas automated prompting is when an LLM itself generated the prompts. Wang et. al. describes the use cases for these types and concludes that manual prompting generates responses with higher accuracy whilst automated prompts are easier and cheaper to generate. Something that has shown promising results when focusing on specific fields is what is called *zero-shot prompting* within manual prompting. *Zero-shot prompting* means that the LLM is given a single detailed prompt without any context upon which subsequent responses are generated [23].

B. Chatbots for psychological therapy

The introduction of chatbots in the field of psychology and therapy brings anticipation of scalable solutions that can serve many different patients concurrently independent of time and location [2]. Bendig et al. [2] concludes that most research in the field in question primarily revolved around pilot studies concentrating on stress, depression, and anxiety metric. The

findings of these studies underscored the positive impact of chatbot interaction on the mental health of the participants, but even though the results seem promising it is clear that there needs to be more research within the area.

Yet, Bendig et al. [2] brings light to the question of what role chatbots should have in therapy. Should it be employed as a therapeutic tool, or simply as an aid to lower the barriers to seeking mental health support? Cameron et al. suggests a different angle, proposing that chatbots have higher usability when focused on providing information rather than curing patients [25]. This opens up ways for further research into the roles and potential of chatbots in psychological treatment.

Furthermore, Cameron et al. affirm that chatbots present a safe and comfortable space for patients to communicate their concerns, meaning that chatbots have the potential to create therapeutic environments [25]. Nevertheless, it is crucial to acknowledge the need for more rigorous research in this area.

In their paper, Andersson et. al. [26] researches if online therapy can be used within the pschological field Cognitive Behavioral Therapy (CBT). Their study concluded that Internet based CBT (ICBT) could be a useful supplement to standard CBT. They argue that ICBT could make treatment more available and cost effective. They do however point out that the effectiveness of internet based treatment is not sufficiently tested, making it possible that ICBT is a worse alternative from a treatment perspective. Another fact that Andersson et. al. is that with the increased availability of ICBT, blended formats of the treatment might be developed. Until now, studies have been conducted on either face-to-face treatment och entirely digital treatment. By combining these, a more successfully format might be discovered. But this will require further studies from a psychological point of view [26].

C. Usability of chatbots from a Human-Computer-Interaction perspective

From a user's point of view, chatbots need to be more useful than traditional ways to get information, like websites or apps [27]. This could be done by providing something that traditional manuals and guides can not do, for example giving the user tips and guidelines that the user can try and then discuss with the chatbot. This trial-and-error approach, as observed by Jain [27], seems to resonate with the users more than only seeking assistance from external sources. He therefore suggests that chatbots should make minimal use of links to other websites in order to maintain user retention and achieve this iterative approach. Jain also points out that it is important that the chatbot has its own distinctive personality since users tend to accept the chatbot more if it has one.

To maintain user retention the chatbot should make use of a minimal amount of links to external services as well as provide the users with a fair amount of tips and guidelines since users favor a trial-and-error approach rather than searching the internet for manuals or guides.

Another study by Følstad and Skjuve [28] found that a major reason for user frustration with chatbots is centered around two main issues: firstly, the chatbot failing to comprehend the user prompt, and secondly, the chatbot lacking the necessary functionality to fulfill user requests. These challenges highlight the trade-offs between versatility and specificity that needs to be addressed in chatbot design. The Human-Computer Interaction perspective calls for a fine balance to be struck in chatbot design to ensure user-friendly and efficient interactions.

IV. Method

The study we present in this paper investigates the effectiveness of two chatbots, built on the GPT-3.5 and GPT-4 language models, in simulating psychological conversations. To critically evaluate the performance of these chatbots, our method involved a four-fold approach.

First, we created two chatbots. The first one, named Mike, was created with a basic system prompt to mimic the base version of GPT-3.5. The second one, named Laura, was created with a more extensive prompt, as well as the ability to send resources to the user, meaning some form of media, e.g. a YouTube video or a link to a website.

Secondly, we developed a survey in Google Forms, designed based on the Godspeed Questionnaire [21] for evaluating the computer science aspect, and the Session Rating Scale [22] for evaluating the simulation of a psychologist's session.

Thirdly, we created a test environment, a user interface that hosted these chatbots, developed with a Node.js back-end, Express.js, and a Vue.js front-end. This platform enabled our licensed psychologist participants to engage with the chatbots in a manner that mimics a real-world setting, which offered us valuable insights into the chatbots' performance. This platform can also be used as a platform for future studies of chatbots.

Finally, we let licensed psychologists try Mike and Laura and answer the form created on the Godspeed Questionnaire [21] and Session Rating Scale [22] to evaluate their performance.

Our participant group was limited to licensed psychologists for testing the two chatbots, thereby ensuring an ethical approach to our research. The detailed specifics of our method, the rationale behind our decisions, and the exact implementation of the chatbots and the test environment are elaborated further in the sections below.

A. Survey and evaluation

In order to gather the data needed to answer our research questions we gathered a series of questions within both computer science (CS) and psychology. The CS questions were used to get an indication of the human likeness and the overall interaction experience with the chatbots and these were formulated according to the Godspeed Questionnaire format. The survey also allowed users to comment on their ratings.

The psychological questions served the purpose to align our "sessions" with those a human psychologist would have with a patient. These were formulated in accordance with the SRS described above. A full list of questions can be found in the appendix.

The survey was structured in an order such that the participant would answer some questions personal questions (age, sex, education), try one of the chatbots, and answer questions about their experience, then continue with the same format

for the next chatbot. Finally, the subjects had to compare the chatbots and their experiences, as well as decide if, and in that case which, chatbot they would prefer to use as a personal psychologist.

Upon first reaching the platform developed to host the chatbots, the user was asked to create an account. Upon creation, the participant would get a random order of which the bots were shown on the platform. In this way, we could avoid bias towards one of the bots based on previous conversations.

As said, we chose to limit our subject group to licensed psychologists. This is due to the need for ethics when evaluating the bots on real patients. The participants were found thanks to Birger Moëll, who had access to channels with psychologists.

B. Implementation of the chatbots

As stated, we created two different chatbots which were given names: Mike and Laura. Both of them were also given individual profile pictures to make the users distinguish between them easier than if they had been called for example Chatbot A and B. This info was clearly shown to the user in the chat page (as can be seen in Figure 1 showing the chat for Laura) This was also done to give the chatbots a more distinct personality since this makes it easier for the user to accept it [27]. The fact that we gave them individual names and ”faces” also meant that we gave them all a gender. This could potentially disrupt our study since gender bias might serve as an underlying factor when comparing the two bots.

We opted for OpenAI’s LLM GPT-3.5 as the primary LLM for our chatbots, despite the release of GPT-4 by OpenAI. This decision was influenced by two key factors. Firstly, the GPT-4 API is considerably slower, which means users would need to wait for a longer time to get the responses from the chatbot. This delay could deter potential participants and hence impact our study. Secondly, the OpenAI API exhibits some instability, occasionally returning only errors due to overload. As GPT-4 takes longer to generate responses, it is more prone to return errors when called. Consequently, we decided to utilize the GPT-3.5 model, as it demonstrates greater speed and stability. However, it’s important to note that the GPT-3.5 API can also be overloaded at times, which may have influenced our results.

The individual chatbots are described below.

1) Mike: Mike was developed using the GPT-3.5 language model. It was instructed with the starting prompt:

”You are an AI Psychologist. Give short answers like you are having a verbal conversation.”

The prompt is designed to be as basic as possible, only instructing the model to act like an AI Psychologist. It however also includes that the bot should give short answers that would be similar to a verbal conversation since it otherwise gave essay-like answers which we deemed to be too long. This is mostly because the time to generate this long response was quite extensive, and it did not in our opinion make for lifelike conversations.

This chatbot is used to benchmark how well-suited GPT-3.5 is for psychology with only a minor modification.

2) Laura: Laura, also built using the GPT-3.5 language model, was instructed with the starting prompt:

”You are a world-class psychologist who is incredibly compassionate and understanding. Give answers that confirm the user’s feelings and acknowledge their problems. Then try to help the user with their problems. Try to mirror the user’s feelings and make them feel like you are taking them and their problems seriously.”

In addition to acknowledging the user’s feelings, this version of the chatbot aimed to establish a supportive and empathetic therapeutic environment. This prompt was designed with the results of Moëll’s pilot study in mind, where one of the results was that his bot, which utilized the same principles as ours, was lacking empathy and being too focused on direct actions the user could take. It also uses an established therapy method of affirming the patient and acknowledging his or her problems and worries, while mirroring their behavior. It was constructed in consultation with Moëll, who is as previously stated a licensed psychologist. This way, we could instruct the chatbot to approach the patient in a way that is similar to how a psychologist would approach one.

The prompt also follows the *persona pattern* described by White et al. in their prompt pattern catalog [20], in order to further facilitate the chatbot’s behavior as being from the perspective of a psychologist. Note that this is not an ideal framework for our study since it does not focus the chatbot on behaving in a way that is desired when acting as a psychologist, but makes the chatbot act like it has the perspective of one. We hope that the instructions developed with Moëll compensate for this.

The chatbot could also determine if it would be appropriate to provide additional resources, such as a YouTube video, website link, app recommendation, or an AI-generated exercise, and in that case, attach that resource to the message. This was done by instructing another agent based on GPT-4 to classify if it was appropriate to append any of these resources to the end of the last chatbot message, and in that case generate a search term for the YouTube video, website or app, or generate an exercise for the user. If a search term was generated, a simple search on YouTube or Google was made using their API, and the top result was chosen.

We were able to use GPT-4 for this since the answers that this agent would provide are much shorter, which means the fact that GPT-4 is slower is less detrimental. The prompt that was given to this agent is quite long, and can hence be seen in the Appendix. It follows the template pattern in order to ensure that the answer is in a format that can be parsed by the code that appends the resource.

C. Creating the test environment as a user interface

As stated above, the tech stack for the web application was chosen to be a Node.js back-end that implements Express.js which provides the user with a Vue.js front-end. Both Node.js and Vue.js are popular frameworks with regard to server-side as well as client-side web development. In order to easily and quickly set up a mobile-first front-end, the front-end

framework Bootstrap was used. This allowed us to focus on the overarching design of the client while minimizing problems regarding responsiveness and mobile-first development.

The client, as seen in Figure 1, is designed with the likes of ChatGPT and Bing Chat in mind. This means that the user has the possibility of creating an account on our website and as a result accessing the chatbots. When signing in the user will be redirected to a page describing the survey. From there, the user can choose to chat with the different chatbots in any order they want. The user has the possibility to change the chatbot which it is conversing with at any point, thus saving the current conversation and making it accessible at any point.

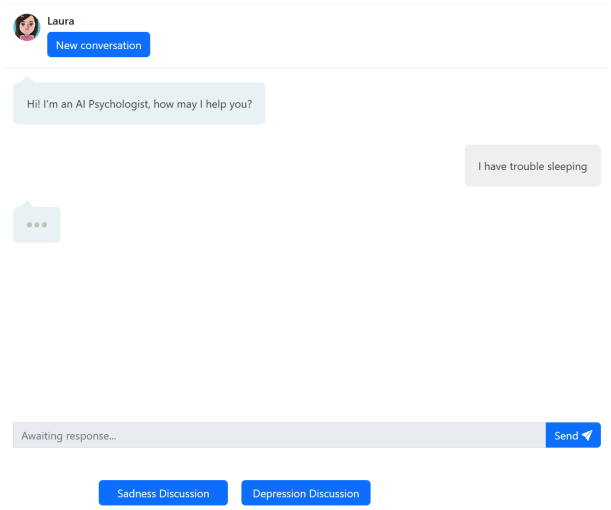

Figure 1: Chat interface for Chatbot: Laura

Furthermore, the client-side also links the user to the survey form, created in Google Forms. Google Forms was chosen for its efficiency and ease of use. Its integration with other applications and services Google developed was also fundamental to the choice of a survey application. In reality, the users should be redirected to the website from the form and not go from the website to the form, but the link was included anyways.

By using already established frameworks for the development of the application, we could make sure that the accessibility of the application meets the standards previous studies have shown it requires to meet. By following best-practice guidelines, other applications, for example, text-to-speech tools, can easily identify the necessary texts and read them out loud to the user, in order to increase accessibility.

The main advantage of creating our own application is that we have full control of what is saved about the user and the conversation, which we would not have had if we used for example Facebook Messenger or Discord. The application allows the user to easily switch between conversations and chatbots and gives the user the freedom of choice that is described as a necessity within previous studies with chatbots. Furthermore, with us in full control over the application, we could design it so that the user can have different chat interfaces for different chatbots. If we want one chatbot to be able to send media such as video, audio, or PDFs, we could implement that on that specific chatbot, which we did.

D. Ethics

In order to ensure that the study was conducted with research ethics and according to all regulations, we only used licensed psychologists as test subjects. This way, we did not have to conduct a large-scale ethical review since that would have been too extensive for this study.

E. Depicting Psychological Therapy as a Work System

When examining psychological therapy through the lens of work systems theory, we define the involved system as follows:

- **Customers:** Patients actively engaged in psychological therapy.
- **Services:** Provision of mental health services or treatment of mental illnesses.
- **Processes and Activities:** Concrete meetings between a patient and a certified psychologist.
- **Participants:** Certified and non-certified psychologists involved in patient treatment, along with other personnel such as medical secretaries or IT support roles.
- **Information:** Dialogues with psychologists, patient-related data, etc.
- **Technologies:** Resources used by psychologists, including automated agents.

V. Results

We have divided our results into two parts, one part for the quantitative results of the study and one for the qualitative results.

Out of the users that registered an account, 52% got Mike as their first bot and 48% got Laura as their first bot. There were 38 users who signed up and tried either of our chatbots. 15 people tried only one of the bots, out of these, 67% tried only Laura and 33% tried only Mike. Out of the remaining 23 people who tried both of the chatbots, 7 people responded to our survey.

A. Quantitative questions

When computing the average values of each of the quantitative questions for Mike and Laura (as seen in Figure 2), we see that Mike achieved a higher value on all questions except for the one regarding *Conversation elegancy*, where the chatbots got the same average value, and the one regarding how the user *felt heard* in the conversation, where Laura had a higher average than Mike. The exact values on the 1 - 5 rating scale can be viewed in the Appendix.

B. Qualitative questions

This section will be divided into aggregated responses for each chatbot respectively, as well as comparisons between them.

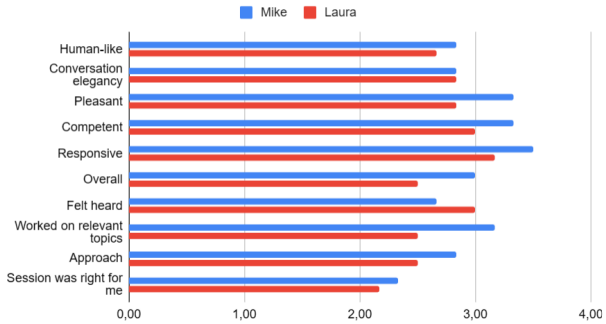

Figure 2: Quantitative results for Chatbots Mike and Laura

1) Mike: Overall the users were happy with Mike. They gave strengths such as "*He gave me some actionable tips for self-care and validated my feelings*" and "*Attentive and normalizing, did not provide too much information or tips too quickly.*". The weaknesses were given as for example "*The answers sometimes felt a bit "generalized", like they weren't aimed specifically towards me*" or "*It felt like I was googling stuff rather than in therapy.*". One user also felt that Mike asked too few questions but specifically pointed out that that does not have to be a bad thing in all cases.

With regards to using Mike in real therapy, all users felt that Mike was best suited for use in easier environments with the sole purpose of answering simple questions and leading the user to the right sort of therapy. For example, one user suggested that Mike could be used for initial screening within children or parental care, or in areas where psychologists tend to get many of the same questions repeatedly.

2) Laura: Two users had problems with Laura, stating that she either "buffered" or did not answer any of the questions given. Except for these two users, who probably experienced the already described problems with the OpenAI API overloading, the people who participated were quite divided in their opinions. For example, one user said that "*She was very soft and gentle and giving me relevant advice and exercises*" while another user said that "*Too quick to suggest diagnoses and solutions.*" or "*She first sent a very annoying video with an american guy talking about stress reduction techniques. and then she told me to seek a mental health professional. i didn't feel as though i got much help.*". Despite this, a theme in our results is that users felt that Laura was good at making the conversation feel natural and smooth, like a human, and that she handled the problems they describe with care and sympathy.

The main weaknesses of Laura were said to be that she was too fast when giving advice based on the information and that she quickly suggested seeking therapy elsewhere. Another user also felt that Laura was too impersonal and that humans are more complex than chatbots can handle. She pointed out that giving answers quickly or diagnosing the patient might not always be the solution to the problem. One user gave concerns that despite her being good at giving advice and helpful resources regarding suicidality, the patient might not be susceptible to the advice given over a screen.

The users were overall united in the fact that Laura should not be used for real-life therapy. The users complained about the lack of "*active interventions*" and the fact she lacked a certain intuitive feeling that psychologists need to have. Despite this, two users thought that with a little more work and fine-tuning, there might be a possibility for Laura to be used on real patients. The psychologists thought that if Laura were to be used on patients, it would be as a first stage or to answer questions. One user also suggested that Laura could be used to answer basic questions and generate routines for example treatment of sleep deprivation.

3) Comparison of Mike and Laura: When comparing Mike and Laura, the users were divided as to which one was the better of them. Some users did not find any particular differences between them while one thought that Laura was more human-like than Mike. One user felt that Laura was too quick to answer questions while Mike made more inquiries to the user, trying to guide them into finding where their problems came from themselves. 42.9% felt neither bot was suitable as a personal psychologist, while 28.6% preferred Laura and 28.6% preferred Mike. One user that felt that Laura would be best for personal psychology felt so because she was "*more gentle and humanlike*" whilst two users that would not like any of the bots did so because "*both were machine like and gave advice before asking questions*" and that if she were to seek therapy, she "*would want someone to work with me in a more in depth manner towards a goal that we have formulate*" and that both the chatbots were good at making her feel validated but they lacked the more in-depth approach she would expect from a psychologist.

VI. Discussion

A. Quantitative Assessment of the Chatbots

The numerical data from our study revealed a preference for Mike over Laura across most measured dimensions. Except for the categories of "Conversation Elegancy" and "Felt Heard", Mike outperformed Laura, indicating a more successful implementation of the counseling function.

Interestingly, despite the only substantive differences between Laura and Mike being the extensiveness of their prompts and Laura's additional resources, the psychologists in our study favored Mike. This could suggest a potential drawback of an extensive prompt or added resources, but further investigation is required to conclusively determine this.

It's important to note that other underlying factors, such as the difference in response time—Laura made two API calls compared to Mike's single one—could also have influenced the user experience. Moreover, the suitability of Laura's additional resources, generated by the GPT-4 model, in conjunction with her GPT-3.5 responses might have negatively impacted her perceived performance.

Although the exact reasons for Mike's preference remain inconclusive from our quantitative results, they nonetheless provide important considerations for the future design of psychology-focused chatbots based on GPT models.

B. Qualitative Assessment of the Chatbots

In our study, we analyzed the responses of two chatbots, Laura and Mike, both based on the LLM GPT-3.5. The

comparison was rooted in the differentiation of their responses to the same user inputs, a result of the specific prompts given to each.

Laura's responses were characterized by empathy and understanding, with her prompts oriented toward validation and consolation. This was reflected in user feedback, where participants noted feeling "heard" and understood by Laura. However, she fell short in providing viable solutions and maintaining productive conversations. Her propensity to prematurely suggest diagnoses and inability to accurately interpret users' descriptions of their problems highlighted these shortcomings.

Mike, in contrast, was more solution-oriented and reliable in his responses. Although less empathetic than Laura, Mike was able to provide basic interventions and validations that were appreciated by the users.

Our findings suggest that the design of a chatbot's prompts plays a pivotal role in determining its proficiency in specific areas. While a greater focus on empathy can yield a chatbot that is more comforting, this might come at the cost of its problem-solving abilities.

C. Potential Improvements and Future Directions

Our study sheds light on the significant influence of prompts on the behavior of chatbots. The differences in performance and user feedback for Laura and Mike, both based on the same LLM GPT-3.5, demonstrate this impact. This offers promising opportunities for refining chatbot performance through the enhancement of prompt design.

Future research could investigate more sophisticated prompt engineering techniques, with a specific emphasis on psychological applications. By developing prompts that strike a balance between empathy and problem-solving, we could create chatbots that cater more effectively to users' needs.

It is also certain that LLMs will improve much in the coming years, which means that studies like this one will have to be repeated for these models and their specific attributes and characteristics. Our evaluation method, including the website, can of course be used for this, which will make future work less time-consuming.

However, it's worth noting that our study's participants were exclusively psychologists. Thus, to fully understand the efficacy of our prompt modifications, further research should involve different user groups, including actual patients. The divergence in feedback between psychologists and patients could yield further insights into prompt design and its influence on chatbot performance.

D. The future for LLM-based psychologists

The future of LLM-based psychologists looks promising but also challenging, as underscored by the results of this study. While our research indicates that the current LLMs are not yet ready to substitute human therapists, they do offer the potential for supplementary roles within the therapeutic landscape. This potential is, however, predicted by several factors that require further development and refinement.

This study shows that the current LLMs are not advanced enough to be used for psychological therapy. There are multiple reasons for this given by the psychologists that participated in our study. One of the reasons the psychologists felt this is that in many cases, a patient seeks the help of a professional psychologist because they need to speak to a human that listens to them and validates their feelings at the same time that psychological progress is being made. The chatbots that can be developed based on the current LLMs are still quite machine-like and thus the need for human interaction is not met. It is currently unclear whether this will ever be successfully met but as of right now, they are not.

On the other hand, chatbots like Mike and Laura, though imperfect in their current form, could serve to lessen the workload of human psychologists by answering frequently asked questions or offering basic guidance. Multiple subjects in our study described that they get a lot of the same basic questions which have mostly the same standard answers. The use of chatbots to answer these could free up more time for psychologists to focus on the more complex aspects of patient care, offering a form of triage in managing patient flows. This potential augmentation of human therapy also opens a path toward making psychological assistance more accessible.

LLMs, once sufficiently developed, could potentially serve as a gateway to professional treatment for individuals who might be hesitant or reluctant to seek help from a human therapist due to stigma, logistical issues, or other barriers. This hypothesis is supported by our findings showing that some psychologists perceived the bots as helpful in redirecting them to other resources. Previous studies also show that chatbots can be a useful tool to extend the reach of treatment and make it more readily available [2], as detailed in Section I.

The fact that some psychologists found the bot to be helpful at redirecting to other instances contradicts what Jain found in [27], namely that chatbot should minimize redirecting to attain a more iterative approach with the user, as discussed in Section III.

In addition to serving as an entry point, LLM-based psychologists like Mike and Laura could perform a vital role in acknowledging and affirming the user's emotions, as suggested by our results. As some of the psychologists who tested the bots said, the chatbots, particularly Laura, were predominantly focused on validating user emotions as opposed to providing interventions. While this reflects a limitation in the proficiency of the chatbots, it also indicates their potential in offering affirmation and support to users.

However, for this to be realized, the reliability, responsiveness, and competence of such LLMs need to be improved, as underscored by our research and existing literature.

E. System Theory Based Implications of Chatbot Implementation for Psychological Therapy

Utilizing the framework of the system theory, this section describes the potential impacts of integrating chatbots for tasks of simpler nature, as suggested by our research findings.

Within the 'Technologies' component of the system, the introduction of LLM Chatbots is projected to significantly amplify the volume of patient data collected during the initial phases of psychological therapy. The 'Services' rendered would remain unchanged, as the intent behind chatbot incorporation is to bolster the system's existing components

rather than modify the provided services. Furthermore, the interaction with the chatbot would represent a new 'process' within the system. This evolution of the system process, in turn, necessitates the inclusion of new support roles to assist the chatbots, adding to the 'Participants' in the system.

F. Validity and reliability of our research

There were only seven people who answered our survey and thus participated fully in our research. The fact that we only have the opinions of seven people to base our analysis on might mean that there is low validity in our data since the law of large numbers cannot be applied. This is also why we did not include any non-basic statistical analysis of our data.

We do however see that these people's opinions were quite alike regarding the two chatbots, which indicates a high reliability. The area where there was the most misalignment in opinions was on whether any of the bots should be used for therapy on patients. This question divided the participants where some thought that neither should be used, some thought Laura, and some thought Mike. If we had had more participants, there might have been a more clear answer to this question as well, but for now, we have to assume that opinions are split on the matter.

Another potential fault in our research is the fact that we only had psychologists try and evaluate the chatbots. Psychologists might not seek the same qualities in a psychological session as patients with no previous experience would. It might also differ what a psychologist seeks from a session with a chatbot, than what a patient does. The psychologists as a whole felt that Mike was better than Laura and that he seemed more competent within the field. There might be the case that patients rather would like Laura because she for example is more affirmative in her responses. Further research would have to be made in this area since we deliberately kept patients outside our scope for ethical reasons.

VII. Conclusion

Based on our study, we can say that an extended version of LLMs such as GPT-3.5 and GPT-4 seem to perform worse than a base version of the same LLMs with regards to psychological therapy. In our view, this is because a prompt that is too extensive will constrain the chatbot into following rigid instructions, instead of answering according to its training. However, the fact that the more extensive chatbot performed worse in our case does not mean that more extensive prompt engineering makes for worse chatbots but instead shows that more extensive prompts can sometimes restrict the model. If the prompts are developed enough, maybe with iterative testing, it could be possible to receive responses more suited for the specific field. More research within prompt engineering will however be required to potentially reach the desirable goal.

In conclusion, while current LLM-based chatbots are not ready to replace human therapists, they show potential as supplementary tools for psychologists. They can lessen the workload of psychologists by answering common questions and providing basic guidance. Additionally, chatbots can serve as an entry point to psychological treatment, increasing accessibility. Although further improvements are required, LLM-based chatbots have the potential to enhance therapy by acknowledging and affirming user emotions, contributing to the overall support provided.

Furthermore, if chatbots were to be implemented as supplementary roles, the work system as a whole would not change as the chatbots only have a supporting role, but it would create new processes that have to be incorporated into the psychologists' work.

VIII. Acknowledgements

We want to express a deep appreciation to all those who made this study possible.

First and foremost, our utmost gratitude to Birger Moëll, Ph.D. Student, KTH, whose expertise, patience, and eagerness added so much to our research experience. Thank you for introducing us to the world of machine learning, for all the hours you put in, and for always encouraging us to ask more questions.

We also want to thank everyone else in the project group who helped us with everything from testing prototypes to proofreading this paper.

Finally, we thank our supervisors at KTH: Jonas Beskow, Professor, KTH, for his guidance, always nudging us in the right direction and making us see the greater context, and Mattias Wiggberg, Ph.D., KTH, for your enthusiasm in helping us make this study relevant in an IEM perspective.

References

[1] D. Vigo, G. Thornicroft, and R. Atun, "Estimating the true global burden of mental illness", English, *The Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, Feb. 2016, ISSN: 2215-0366. DOI: 10.1016/S2215-0366(15)00505-2.

[2] E. Bendig, B. Erb, L. Schulze-Thuesing, and H. Baumeister, "The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review", *Verhaltenstherapie*, pp. 1–13, 2019. DOI: 10.1159/000501812.

[3] G. Tyen, M. Brenchley, A. Caines, and P. Buttery, "Towards an open-domain chatbot for language practice", in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 234–249. DOI: 10.18653/v1/2022.bea-1.28.

[4] K. Roose, *How chatgpt kicked off an a.i. arms race*, Feb. 2023. [Online]. Available: https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html.

[5] *Conversation understanding: Open domain vs. closed domain*, Jun. 2022. [Online]. Available: https://symbl.ai/blog/conversation-understanding-open-domain-vs-closed-domain/.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", *Nature*, vol. 323, pp. 533–536, 1986.

[7] D. Jurafsky and J. H. Martin, "7.6 training neural nets", in *Speech and language processing*, 3rd ed. Pearson Education, 2014, pp. 150–158. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.

[8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[9] P. R. Iyer, A. M.R, and H. R., "Research perspectives and advancements in open domain chatbots", *International Journal of Engineering and Advanced Technology*, vol. 9, no. 4, pp. 1672–1678, 2020. DOI: 10.35940/ijeat.d8734.049420.

[10] [Online]. Available: https://platform.openai.com/docs/guides/chat/introduction.

[11] Nov. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health.

[12] Dec. 2014. [Online]. Available: https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/lag-2003460-om-etikprovning-av-forskning-som_sfs-2003-460.

[13] S. Alter, "Work system theory: Overview of core concepts, extensions, and challenges for the future", *Journal of the Association for Information Systems*, vol. 14, pp. 72–121, Feb. 2013. DOI: 10.17705/1jais.00323.

[14] Mar. 2023. [Online]. Available: https://openai.com/research/gpt-4.

[15] Y. Shen, L. Heacock, J. Elias, *et al.*, "Chatgpt and other large language models are double-edgednbsp; nbsp; nbsp; nbsp; nbsp; nbsp; nbsp; nbsp; nbsp; nbsp; swords", *Radiology*, vol. 307, no. 2, 2023. DOI: 10.1148/radiol.230163.

[16] C. Si, Z. Gan, Z. Yang, *et al.*, *Prompting gpt-3 to be reliable*, 2023. arXiv: 2210.09150 [cs.CL].

[17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, New York, NY, USA: Association for Computing Machinery, 2021, pp. 610–623, ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. [Online]. Available: https://doi.org/10.1145/3442188.3445922.

[18] K. Cho, B. van Merrienboer, Ç. Gülçehre, *et al.*, "Learning phrase representations using rnn encoder–decoder for statistical machine translation", in *Conference on Empirical Methods in Natural Language Processing*, 2014.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", *arXiv preprint arXiv:1409.0473*, 2014.

[20] J. White, Q. Fu, S. Hays, *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt", Feb. 2023. DOI: 10.48550/arXiv.2302.11382.

[21] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots", *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009. DOI: 10.1007/s12369-008-0001-3.

[22] B. Duncan, S. Miller, J. Sparks, *et al.*, "The session rating scale: Preliminary psychometric properties of a "working" alliance measure", *Journal of Brief Therapy*, vol. 3, pp. 3–12, Sep. 2003.

[23] J. Wang, E. Shi, S. Yu, *et al.*, *Prompt engineering for healthcare: Methodologies and applications*, 2023. arXiv: 2304.14670 [cs.AI].

[24] H. Kumar, I. Musabirov, J. Shi, *et al.*, *Exploring the design of prompts for applying gpt-3 based chatbots: A mental wellbeing case study on mechanical turk*, 2022. arXiv: 2209.11344 [cs.HC].

[25] G. Cameron, D. Cameron, G. Megaw, *et al.*, "Towards a chatbot for digital counselling", in *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*, ser. HCI '17, Sunderland, UK: BCS Learning amp; Development Ltd., 2017. DOI: 10.14236/ewic/HCI2017.24. [Online]. Available: https://doi.org/10.14236/ewic/HCI2017.24.

[26] G. Andersson, P. Cuijpers, P. Carlbring, H. Riper, and E. Hedman, "Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: A systematic review and meta-analysis", *World Psychiatry*, vol. 13, no. 3, pp. 288–295, 2014. DOI: https://doi.org/10.1002/wps.20151.

[27] M. Jain, P. Kumar, R. Kota, and S. N. Patel, "Evaluating and informing the design of chatbots", *Proceedings of the 2018 Designing Interactive Systems Conference*, Jun. 2018. DOI: 10.1145/3196709.3196735.

[28] A. Følstad and M. Skjuve, "Chatbots for customer service", *Proceedings of the 1st International Conference on Conversational User Interfaces*, Aug. 2019. DOI: 10.1145/3342775.3342784.

About the Authors

Isak Nordgren (left) and **Gustaf Svensson** (right) are both students at the Royal Institute of Technology (KTH), Sweden, and are both enrolled in the Industrial Engineering and Management program. Both authors have made substantial contributions to all parts of the paper, where Isak has focused more on the machine learning aspects to create the models, while Gustaf has focused more on the test environment and web development aspects. The evaluation and analysis have been designed and carried out in collaboration. No part of the thesis can be attributed to only one author. The authors have complemented each other throughout the process, and given their own unique insight into the study. With a well-distributed workload, the authors have successfully completed this bachelor's thesis.

APPENDIX

A. Code on Github

https://github.com/leegrash/kbt-chat-app

B. Questions in the survey

Describe your interaction with [Bot name] shortly
Scale 1-5: Machinelike/Humanlike
Scale 1-5: The conversation moved rigidly/The conversation moved elegantly
Scale 1-5: Unpleasant/Pleasant
Scale 1-5: Incompetent/Competent
Scale 1-5: Unresponsive/Responsive
What do you think the main strength of the chatbot was?
What do you think the main weakness of the chatbot was?
How appropriate do you think it would be to use [Bot-name] for real therapy and treatment?
In what context do you think it would be most helpful to use [Bot-name]? Why?
Scale 1-5: Rate the overall quality of [Bot-name]
Other comments regarding your experience with [Bot-name]
Scale 1-5: I did not feel heard, understood, and respected. / I felt heard, understood, and respected.
Scale 1-5: We did not work on or talk about what I wanted to work on and talk about. / We worked on and talked about what I wanted to work on and talk about
Scale 1-5: The therapist's approach is not a good fit for me. / The therapist's approach is a good fit for me.
Scale 1-5: There was something missing in the session today. / Overall, today's session was right for me.

What did you experience as the largest difference between the chatbots?
Which one would you prefer to use as a personal psychologist? Why?
Is there anything that you felt was missing from one or both of the chatbots?
Other comments

C. Prompt given to the resource-agent described in Section IV

"I will send you a conversation between a user and an AI Psychologist. Your task is to determine if it would be appropriate to append a youtube link, a link to a website, a link to an app, or an AI Generated exercise to the last message by the AI Psychologist.

If you think it is appropriate to send a youtube video: send a query to search on youtube that is fitting to the last message. Send it in this format: "Youtube: {query}".

If you think it is appropriate to send a link to a website: send the name of the website that is fitting. Send it in the format: "Website: {website_name}".

If you think it is appropriate to send a link to an app: send the name of the app that is fitting. Send it in this format: "App: {app_name}".

If you think it is appropriate to send an AI generated exercise: Design an exercise. Send it in this format: "Exercise: {exercise_description}".

If you do not think it is appropriate to send anything: send "0".

Keep in mind that it is not appropriate to send many resources in sequence, and that too many resources may confuse the user.

This is the conversation:
###
[Conversation on "User: message, Chatbot: message"-format]
###"

D. Exact mean values of the quantitative part of the study

Human-like	2.83
Conversation elegancy	2.83
Pleasant	3.33
Competent	3.33
Responsive	3.5
Overall	3.0
Felt heard	2.67
Worked on relevant topics	3.17
Approach	2.83
Session was right for me	2.33

Table I: Quantitative results for Mike

Humanlike	2.67
Conversation elegancy	2.83
Pleasant	2.83
Competent	3.0
Responsive	3.17
Overall	2.5
Felt heard	3.0
Worked on relevant topics	2.5
Approach	2.5
Session was right for me	2.17

Table II: Quantitative results for Laura

TRITA-EECS-EX-2023:417