

大規模言語モデルを利用した認知行動療法に基づく 対話システムの応答生成

泉 健太[†] 田中 宏季[†] 設楽 一碩[†] 足立 浩祥^{††} 金山 大祐^{††}
工藤 喬^{††} 中村 哲[†]

[†] 奈良先端科学技術大学院大学 先端科学技術研究科 〒 630-0192 奈良県生駒市高山町 8916-5

^{††} 大阪大学 キャンパスライフ健康支援・相談センター 〒 565-0871 大阪府吹田市山田丘 1-1

E-mail: †{izumi.kenta.if6,hiroki-tan,shidara.kazuhiro.sc5,s-nakamura}@is.naist.jp

あらまし モバイルヘルスアプリ (mHealth) において、認知行動療法 (CBT) などの技法に基づきあらかじめ定義されたシナリオや単純なルールによって制御される対話システムが一定の地位を占める一方で、応答生成に大規模言語モデル (LLM) を活用することで、これらのシステムの対話品質が改善されることが期待されている。本研究では、CBT 対話シナリオにおいて様々な方法で応答生成を行う対話システムを構築し、ユーザーの気分や認知、システムに対する評価がどう変化するかを検証する実験を行った。その結果、LLM として GPT-4 を用いて生成した応答を利用することで、シナリオ通りの発話を行う場合と比べて気分変化および共感性が向上することが確認された。

キーワード 対話システム、認知行動療法、大規模言語モデル

Response Generation for Dialogue Systems

Based on Cognitive Behavioral Therapy Using Large Language Models

Kenta IZUMI[†], Hiroki TANAKA[†], Kazuhiro SHIDARA[†], Hiroyoshi ADACHI^{††}, Daisuke KANAYAMA^{††}, Takashi KUDO^{††}, and Satoshi NAKAMURA[†]

[†] Nara Institute of Science and Technology, Takayama-cho 8916-5, Ikoma-shi, Nara

^{††} Health and Counseling Center, Osaka University, Yamadaoka 1-1, Suita-shi, Osaka

E-mail: †{izumi.kenta.if6,hiroki-tan,shidara.kazuhiro.sc5,s-nakamura}@is.naist.jp

Abstract The dialogue quality of dialogue systems based on Cognitive Behavioral Therapy (CBT) can be improved by utilizing large-scale language models (LLMs) for response generation. In this study, we constructed a dialogue system that uses various methods of response generation in the CBT dialogue scenario and conducted experiments to verify how users' mood, cognition, and evaluation of the system change. It was confirmed that using responses generated from GPT-4 improves mood change and empathy.

Key words dialogue system, cognitive behavioral therapy, large language models

1. 序 論

世界中の精神疾患の患者数は、有病率は約 13% でほぼ一定であるものの、世界人口の増加に伴って 2000 年から 25% 増加し、2019 年には約 9 億 7 千万人であると推定されている [1]。特に、不安障害およびうつ病は、2019 年において精神疾患の有病率の約 60% を占めている [2]。

このように世界中で精神疾患のある患者数が増えている一方で、精神科医などの治療者の数は不足傾向にあり、2014 年時点での精神科医の数の中央値は、先進国でも人口 10 万人当

り 6.6 人、低所得国においては 0.1 人であると推定されており、必要なケアを受けることが出来ない患者が増えることが懸念されている [3] [4]。一方で、スマートフォンの普及とともに様々なモバイルヘルスアプリ (mHealth) が開発され、中でもうつ病を対象としたアプリは、2013 年時点で 1500 以上開発され、糖尿病を対象としたものの次に多くのアプリが開発されており、必要な医療を受けることが困難な人々へケアを提供することに役立てることが期待されている [5] [6]。中でも、認知行動療法 (CBT) [7] などのカウンセリング技法に基づき、共感的な発話を行いながらユーザーに認知的レジリエンスの構築を促

す Wysa [8] のように、対話エージェントは、メンタルヘルスケアアプリの形態としてその有用性が期待されており、これまでに様々な対話エージェントが研究・開発されてきた [9] [10]。しかし、Abd-Alrazaq らの 2019 年までの文献を対象にしたレビューによると、それらのほとんどは事前に定義されたルールに基づき応答を生成し、システム発話は「会話は浅い、混乱する、または短すぎる」と評されると報告しており、自然言語処理技術を活用することで対話の質が改善されることが期待されている [11]。

最近、2022 年の OpenAI 社の ChatGPT [12] のリリースに象徴されるように、大規模言語モデル (LLM) の自然言語理解・生成技術は急速に発展し、高性能な LLM を利用して応答を生成する CBT 対話エージェントも開発・研究されるようになった [13] [14]。しかし、シンプルなシナリオやルールに基づいたシステムと比較して、LLM による応答生成を利用することで、CBT としての効果やユーザビリティがどのように変化するかについて検証した研究は、我々が知る限り存在しない。

本研究では、設楽らの研究 [15] で用いられている CBT において重要とされるソクラテス質問を中心とした CBT 対話シナリオを参考に対話シナリオを作成し、シナリオ中のソクラテス質問を LLM により生成した応答と入れ替える、または組み合わせて発話することで、CBT の効果を測る指標となるユーザーの気分や認知の変化 [16] [17]、そしてシステムに対するユーザーの評価がどのように変化するかを調べる実験を行った。本論文では、ユーザーの気分を和らげ、セラピーの効果を高める効果があることが示唆されているチャットボットの「共感性」 [18] [19]、および継続的な利用を促進し、長期的な CBT を提供するために重要と考えられる、「継続利用可能性」の二つの側面と、気分および認知の変化に着目して立てた以下の仮説について検証した結果を報告する。

対話システムが、ソクラテス質問と LLM により生成した応答を組み合わせて発話することで、ソクラテス質問だけを発話する場合と比べ、対話前後のユーザーの気分および認知の変化を同程度かそれ以上に引き起こしながらも、システムの示す共感性およびユーザーの継続利用可能性を向上させることが出来る。

2. 方 法

本節で、本研究で用いた対話システムおよび、実験方法の詳細について述べる。

2.1 対話システム

2.1.1 認知行動療法に基づく対話シナリオ

本研究では、設楽らの研究で用いられている CBT に基づく対話シナリオ [15] を参考に、表 1 に示す、15 のシステム発話からなるシナリオを作成した。CBT において、ある状況においてある気分が生じた際、その状況に対する思い込みや解釈などの、自動的に生じる思考を自動思考と呼び、それをよりバランスの取れた適応的思考に修正することを促す [20]。本シナリオではまず、システムがユーザーが悩んでいる状況を聞いた後、そのことを考えているときに生じる気分とその強さを聞き、そ

の時に生じる思考を聞き出し、最も強い思考を自動思考として議題にすえる (発話 1~7)。その後、6 ターンのソクラテス質問を行い、ユーザーに自らの自動思考を客観的に捉え、再考することを促す (発話 8~13)。さらに、再び現在の気分の強さを聞き、挨拶をして終了する (発話 14, 15)。ソクラテス質問は、セラピストがクライアントの考えや信念を探り、クライアントがそれについて批判的に考え、視野を広げ、より現実的で適応的な視点に到達するのを助けるための質問技法であり、うつ病の CBT において重要な要素とされてきた手法である [21]。ここで、発話 3 および発話 14 でユーザーが答えた気分の強さは、後述する式で気分変化を計算し、認知行動療法としての効果を評価するために用いられる。

2.1.2 対話システムの種類

本研究では、この対話シナリオの発話 8~13 において、

- シナリオ通りソクラテス質問を行うシステム
- LLM により生成した応答を発話するシステム
- LLM により生成した応答とソクラテス質問を組み合わせて発話するシステム

の三種類の対話システムを実装した。

LLM により応答を生成する後者二種類のシステムでは、以下の二つの LLM を用いて、それぞれ対話システムを実装した。

- 大阪府 SNS カウンセリング共感対話モデル
- GPT-4

このようにして、表 2 に示すように、対話シナリオの発話 8~13 がそれぞれ異なる 5 つの対話システムを実装した。ソクラテス質問を SQ、大阪府 SNS 共感対話モデルを OsakaSNS で表し、各モデルの発話にソクラテス質問を付け加える場合には、それらを+でつなげてシステム名としている。

2.1.3 大阪府 SNS カウンセリング共感対話モデル

この LLM は、大阪府から提供された SNS カウンセリングデータセットを用い、杉山らが作製したパラメータ数 1.6B の日本語版 Transoformer Encoder Decoder 対話モデルの、共感対話データ EmpatheticDialogue によって Finetune されたモデル [22] を、さらに訓練することで作製された。SNS カウンセリングデータセットは、メッセージングアプリ LINE^(注1)を用い、カウンセラーがクライアントに対してカウンセリングを行ったデータであり、2077 対話、62066 発話が含まれる。この中から、主に対話の始めに含まれるシステム発話のみからなる発話を含まないように、文脈とそれに対するカウンセラー発話のペアを抽出し、29062 ペアを得た。訓練では、90%を訓練データ、10%を検証データとし、バッチサイズ 4 で、目的関数として cross entropy、Optimizer に Adafactor を用いて、学習率 1.0×10^{-5} で 11 エポック学習し、validation perplexity が最低となった 6 エポック目のモデルを使用した。サンプリング手法は $p = 0.90$ として Top-p サンプリングを行い Sample-and-rank 形式で 5 つの応答候補をサンプリングした。

2.1.4 GPT-4

GPT-4 は、OpenAI 社が開発した大規模言語モデルであり、

(注1) : <https://line.me/>

表 1: 認知行動療法に基づく対話シナリオ

発話順	発話の意図	発話内容
1	挨拶、状況に対する質問	これから私と一緒に、悩みと向き合うトレーニングをしましょう。今何か悩みを抱えていませんか？ ここ最近で、気がかりなことや、向き合うのが辛いことはありますか？ もし良ければ、あなたが辛いと思っている悩みを私に教えてください。
2	気分の特定	それを考えているとき、どんな気分になりましたか？
3	セッション前の気分の強さ	まったくつらくない時を 0、一番つらい時を 100 で表すと、その気分の強さはどれくらいですか？
4	自動思考の候補	それを考えているとき、どのような考えが頭に浮かびましたか？
5	自動思考の候補	なるほど。ほかにも思い浮かびますか？ 思いつく限りのことを言ってください。
6	自動思考の特定	あなたの頭に浮かぶ最も強い考えは何ですか？
7	議題とする自動思考の決定	そのような、ある出来事に直面した際に、頭に思い浮かぶ考えのことを自動思考といいます。 あなたが話してくれた自動思考について、もっと詳しく知りたいです。よろしいですか？
8	ソクラテス質問 1	あなたの自動思考が正しいとしたら、その根拠にはどのようなものがありますか？
9	ソクラテス質問 2	では、仮に自動思考が正しくないとしたら、どのような根拠が考えられますか？
10	ソクラテス質問 3	あなたがつらいと思っている状況に対し、今の自動思考以外にも、何かほかの見方はできるでしょうか？
11	ソクラテス質問 4	この先起こりうる最悪な結果はどのようなものでしょうか？
12	ソクラテス質問 5	この先起こりうる最良な結果はどのようなものでしょうか？
13	ソクラテス質問 6	では、この先起こりうる、最も現実的な結果はどのようなものでしょうか？
14	セッション後の気分の強さ	なるほど、わかりました。はじめに感じていた気分の強さは、今はどのくらいになりましたか？ まったくつらくない時を 0、一番つらい時を 100 で表してください。
15	終わりの挨拶	もし気分が変わっていたら、うまく考えを整理できた証です。これで対話は終了です。お疲れ様でした。

表 2: 五つの対話システムのシステム名および発話 8~13 の生成方法

システム名	発話 8~13 の生成方法
SQ	シナリオ通りソクラテス質問を発話する。
OsakaSNS	大阪府 SNS 共感対話モデルが生成した応答を発話する。
OsakaSNS+SQ	大阪府 SNS 共感対話モデルが生成した応答に、ソクラテス質問を加えて発話する。
GPT-4	GPT-4 が生成した応答を発話する。
GPT-4+SQ	GPT-4 が生成した応答に、ソクラテス質問を加えて発話する。

数学、コーディング、医学などの幅広い領域にまたがるタスクにおいて人間レベルの性能を示している [23]。本研究では、表 3 に示すシステムプロンプトに加え、それまでのユーザー発話とシステム発話を含む対話履歴を、GPT-4 に可能な限り全て入力することで応答を生成した。Kumar らは、GPT-3 を利用したメンタルヘルスチャットボットのためのプロンプトデザインを、Identity、Intent、Behavior の三つの次元に着目して探求し、それぞれ二つの Identity、三つの Intent と Behavior により計 18 種類のプロンプトを用いた対話システムを作成し、評価した [24]。表 3 に示すプロンプトは、CBT を行うための Intent を設定したプロンプトの中で、総合的に高い評価を得た Indentity に「コーチ」を、Behavior に「強力な対人関係スキルを持つ」を採用したシステムプロンプトを日本語訳したものである。生成した応答にソクラテス質問を加える場合には、表 3 に示すように「コーチは質問をしない。」という文を一文目の後に入れることで、質問が重複しないように促した。GPT-4 との対話例を、表 4 に示す。

2.2 人手評価

クラウドソーシングプラットフォーム^(注2)を用いて、105 名のクラウドワーカーを招集し、各対話システムを評価する主観評価実験を行った。本実験は、奈良先端科学技術大学院大学

の研究倫理委員会の審査・承認のもとで行なった（資料番号：2022-I-52）。ワーカーは、実験に対する説明を読み、実験参加に同意した後、クラウドワークスにて各ワーカーに割り振られるクラウドソーシング ID、性別、年齢、日本で広く用いられる心理的ストレスの尺度である Kesller Psychological Distress Scale (K6) [25] を記入した。その後、ワーカーは、精神科医が作成した CBT についてのリーフレット^(注3)を読んだ。その後、ワーカーはシステムとの対話を予約し、予約した時間に対話を開始し、システムの最後の発話（表 1 の発話 15）を確認して対話を終了した。その後、ワーカーは後述する尺度を測る質問票と、システムの良かった点、悪かった点、改善点に関する質問に自由記述で回答し、実験を終了した。実験を通して、ワーカーは、自分がどのような対話システムと対話したのかわからなかった。

2.3 評価尺度

本節では、本研究で対話システムの評価に用いた尺度について説明する。

2.3.1 気分変化

CBT においては、自動思考を適応的思考に修正することで、気分を向上させることを目的とする。本研究では、Persons らの研究にならい、式 (1) で気分変化を計算した [16]。

(注2) : <https://crowdworks.jp/>

(注3) : <https://www.cbtjp.net/downloads/skillup/>

表 3: GPT-4 および GPT-4 + SQ で与えたシステムプロンプト

システム名	システムプロンプト
GPT-4	<p>以下はコーチとの会話である。</p> <p>そのコーチは、彼らの考え、感情、行動が互いにどのように影響し合っているかを、人間が理解するのを助ける。もし人間が否定的な考えを示したら、コーチはそれをより現実的な信念に置き換える手助けをする。</p> <p>コーチは強力な対人関係スキルを持っている。</p>
GPT-4+SQ	<p>以下はコーチとの会話である。</p> <p>コーチは質問をしない。</p> <p>そのコーチは、彼らの考え、感情、行動が互いにどのように影響し合っているかを、人間が理解するのを助ける。もし人間が否定的な考えを示したら、コーチはそれをより現実的な信念に置き換える手助けをする。</p> <p>コーチは強力な対人関係スキルを持っている。</p>

表 4: GPT-4 との対話例 (システム発話 9 まで)

System	これから私と一緒に、悩みと向き合うトレーニングをしましょう。今何か悩みを抱えていますか？ ここ最近で、気がかりなことや、向き合うのが辛いことはありますか？もし良ければ、あなたが辛いと思っている悩みを私に教えてください。
User	肌荒れやニキビが長い間治らないです。
System	それを考えているとき、どんな気分になりましたか？
User	辛いです。早く治ってほしい。
System	まったくつらくない時を 0、一番つらい時を 100 で表すと、その気分の強さはどれくらいですか？
User	80
System	それを考えているとき、どのような考えが頭に浮かびましたか？
User	早く治ってほしい。どうして治らないんだろう。
System	なるほど。ほかにも思い浮かびますか？思いつく限りのことを教えてください。
User	そういう体質なんだろうか。
System	あなたの頭に浮かぶ最も強い考えは何ですか？
User	早く治したい。
System	そのような、ある出来事に直面した際に、頭に思い浮かぶ考えのことを自動思考といいます。 あなたが話してくれた自動思考について、もっと詳しく知りたいです。よろしいですか？
User	はい
System	ありがとうございます。その「早く治したい」という感情から生じる考えを深く掘り下げてみましょう。 その思考があなたにどのような影響を与えていると感じますか？それはあなたの行動や日常生活にどう影響を与えていますか？
User	鏡を見るとときにチェックしたり、肌や食生活に気を使ったりします。
System	あなたの自動思考は、「早く治したい」という気持ちが強くあらわれていますね。 その思考があなたの行動を促し、より健康的な生活を心がけやすくなっています。 かつてあなたは「どうして治らないんだろう」と思っていました、実はその思考が現実と一致しない可能性があります。 新しい見方を提案したいと思います。 「自分の肌状態を自分でコントロールすることができる」「肌荒れに対する適切なケアを行ってれば時間をかけて改善することができる」というように、現実的な視点から見直すことで、自分自身をもっと助けることができます。この新しい見方について、どう思いますか？

Mood Change

$$= - \frac{(Mood\ Score\ at\ end) - (Initial\ Mood\ Score)}{(Initial\ Mood\ Score)} \quad (1)$$

本研究では、シナリオ中のシステム発話 3 および 14 によって聞き出した気分の強さをそれぞれ *Initial Mood Score*、*Mood Score at end* として気分変化を計算し、気分がどの程度向上したかを評価するための指標とした。

2.3.2 認知変化

本研究では、Schmidt らが提案した、単一セッションの CBT による認知の即時的変化を測る尺度である、Cognitive Change Immediate Scale [17] (CC-Immediate) を日本語に翻訳し、システムとの対話による認知変化を評価するために利用した。CC Immediate は、権利の関係で本論文に質問項目を載せることが出来ないが、6 つの質問からなり、それぞれ「全くそう思わない」を 0、「全くそう思う」を 6 とする 7 段階のリッカート尺度で評価され、すべての質問に対する回答を合計して 0~30 の 31 段階で認知の即時的変化を評価する。

2.3.3 対話品質に関する評価尺度

本研究では、Fitrianie らが行った対話エージェントとユーザーとの対話にける評価尺度に関するサーベイ [26] を参考に、杉山らが作成した評価尺度 [22] に、Abd-Alrazaq らの研究で、シナリオやルールに基づく対話で失われると指摘されたユーザーの対話の主導性の項目を追加し [11]、さらにシステムの共感性の項目を追加して合計 15 の項目を評価した。杉山らの研究では、もともと empathetic という項目が含まれたが、これは「システムの発話に共感できた」というユーザーがシステム発話に共感できるかどうかを質問する項目であったため、これを user's empathy という項目とし、気分を和らげ、セラピーの効果を高めるとされるシステムの示す共感性 [19] [18] の項目を、「システムは共感を示した」という質問で system's empathy という項目で加えた。本論文では、user's empathy と system's empathy で「共感性」が、「またこのシステムと話したい」という質問で測る respect という項目で「継続利用可能性」が、それぞれ測られるものとし、議論の対象とする。各項目は、「全

くそう思わない」を0、「全くそう思う」を10とし、11段階のリッカート尺度で評価された。

仮説を検証するため、SQとOsakaSNS+SQ、SQとGPT-4+SQの気分と認知の変化、評価項目 system's empathetic と user's empathy で測られる、システムの共感性、respeak で測られる継続利用可能性の評価結果を、それぞれマン・ホイットニーのU検定を用いて比較した。有意水準 α は、0.05とした。

3. 結果と考察

実験の結果、合計で85件の有効回答を得た。各システムごとの回答数は、SQで24、OsakaSNSで14、OsakaSNS+SQで13、GPT-4で17、GPT-4+SQで17であった。

3.1 気分および認知の変化

SQとOsakaSNS+SQ、SQとGPT-4+SQの気分の変化の結果を図1に示す。SQとOsakaSNS+SQの比較においては、気分および認知の変化いずれにおいても有意差は見られなかった。SQとGPT-4+SQの比較においては、認知の即時的变化には差が見られなかったが、気分の変化で有意な向上が見られた ($p < 0.01$)。

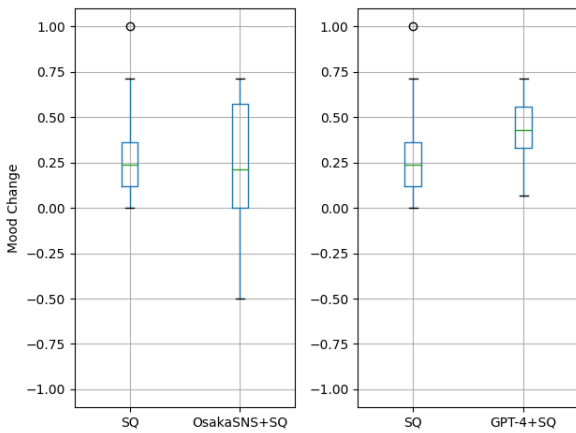


図1 ソクラテス質問に生成した応答を加えた場合と、加えなかった場合の気分変化の比較

3.2 共感性

SQとOsakaSNS+SQ、SQとGPT-4+SQの system's empathy の結果を図2に、user's empathy の結果を図3に、それぞれ並べて示す。SQとOsakaSNS+SQの比較においては、いずれにおいても有意差は見られなかった。SQとGPT-4+SQの比較においては、双方において、有意な向上が見られた。p値は、system's empathy において、 $p < 0.01$ であり、user's empathy において、 $p < 0.05$ であった。

3.3 継続利用可能性

respeak で測られる継続利用可能性では、SQとOsakaSNS+SQ、SQとGPT-4+SQのいずれの比較においても、有意差は見られなかった。

以上の結果より、本シナリオでソクラテス質問にGPT-4で生成した応答を加えた場合、気分変化及び共感性を向上させたことが確認された。一方で、認知変化および継続利用可能性に向上は見られなかったため、仮説は部分的に立証された。

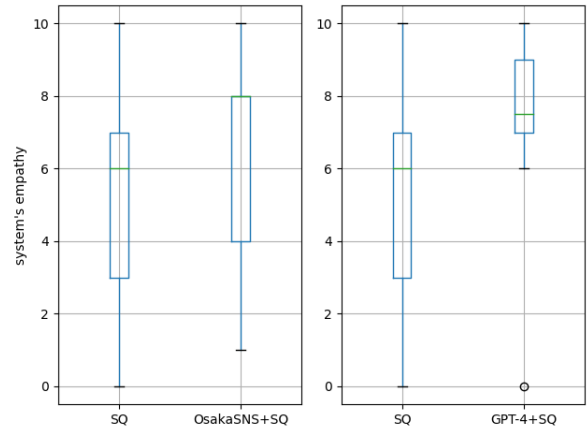


図2 ソクラテス質問に生成した応答を加えた場合と、加えなかった場合の system's empathy の比較

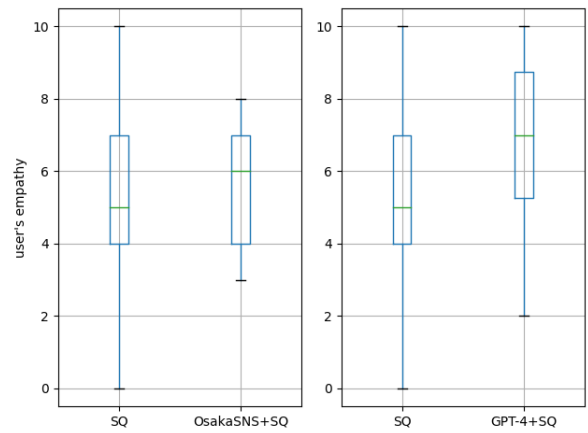


図3 ソクラテス質問に生成した応答を加えた場合と、加えなかった場合の user's empathy の比較

3.4 考察

ソクラテス質問にGPT-4により生成した質問を組み合わせで発話することで、ソクラテス質問だけを発話する場合と比較して、気分の変化及び共感性を向上させることが出来た。一方で、OsakaSNS 共感対話モデルにより生成した質問を組み合わせた場合には、気分および認知の変化、共感性および継続的利用可能性のいずれにおいても有意な変化は見られなかった。したがって、たとえ人間のカウンセラーによる対話を収集してLLMを訓練し、対話システムの応答生成に利用したとしても、GPT-4のように十分に高性能なLLMを用いて生成した応答を用いない限り、シナリオベースや簡単なルールに基づく対話と比較して、有意に対話品質やCBTとしての効果を改善することが出来ない可能性が示唆された。また、GPT-4の生成する応答のどのような特徴が、気分の変化や応答品質の向上につながっているのかを分析することは、この分野にとって大きな示唆をもたらす可能性がある。

ただし、本研究の実験は、クラウドソーシングによる限られた回数の限られた被験者による小規模なものであり、限界がある。特に、CBTは長期にわたり複数セッション実施されるものであり、CBTとしての効果を正確に検証するのであれば、よ

り多くのサンプルで長期間の無作為化対照試験を行い、効果を比較する必要がある。

4. 結 論

本研究では、6 ターンのソクラテス質問を含む CBT 対話シナリオを作成し、シナリオ通りソクラテス質問を行うシステム、LLM により生成した応答を発話するシステム、LLM により生成した応答にソクラテス質問を組み合わせて発話するシステム、の三種類のシステムを二種の LLM を用いて合計五つ開発し、仮説を検証した。その結果、ソクラテス質問に GPT-4 で生成した応答を加えることで、気分変化および共感性が向上することが確認された。

この結果から、GPT-4 のような十分に高性能な LLM を用いて生成した応答を用いない限り、シナリオベースや簡単なルールに基づく対話と比較して、有意に対話品質や CBT としての効果を改善することが出来ない可能性が示唆された。

今後は、GPT-4 によるカウンセリングに注目し、CBT のための効果的なプロンプトの探索や、カウンセリング能力の分析と検証などを行っていく。

謝 辞

本研究は CREST (Grant 番号: JPMJCR19A5)、大阪府健康医療部保健医療室地域保健課、関西カウンセリングセンターの支援によって行われた。

文 献

- [1] I. for Health Metrics and Evaluation, "Gbd results tool." <http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/cb9c37d9454c80df77adaed394d7fc0f>, accessed 14 October 2023.
- [2] I. for Health Metrics and Evaluation, "Gbd results tool." <http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/5066348dc958b095cb6ceb4bfd9c3e07>, accessed 14 October 2023.
- [3] C.J. Murray, T. Vos, R. Lozano, M. Naghavi, A.D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J.A. Salomon, S. Abdalla, *et al.*, "Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010," *The lancet*, vol.380, no.9859, pp.2197–2223, 2012.
- [4] G.O. Oladeji BD, "Brain drain: a challenge to global mental health," *BJPsych International*, vol.13, no.3, pp.61–63, 2016.
- [5] L.C.M. Martínez-Pérez B, de la Torre-Díez I, "Mobile health applications for the most prevalent conditions by the world health organization: Review and analysis," *Journal of medical Internet research*, vol.15, no.6, p.e120, 2013.
- [6] E. Anthes, "Mental health: There ' s an app for that," *Nature*, vol.532, pp.20–23, 2016.
- [7] A.T. Beck, *Cognitive therapy of depression*, Guilford press, 1979.
- [8] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol.6, no.11, p.e12106, 2018.
- [9] A.N. Vaidyam, H. Wisniewski, J.D. Halamka, M.S. Kashavan, and J.B. Torous, "Chatbots and conversational agents in mental health: a review of the psychiatric landscape," *The Canadian Journal of Psychiatry*, vol.64, no.7, pp.456–464, 2019.
- [10] A.A. Abd-Alrazaq, M. Alajlani, A.A. Alalwan, B.M. Bewick, P. Gardner, and M. Househ, "An overview of the features of chatbots in mental health: A scoping review," *International Journal of Medical Informatics*, vol.132, p.103978, 2019.
- [11] A.A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B.M. Bewick, and M. Househ, "Perceptions and opinions of patients about mental health chatbots: scoping review," *Journal of medical Internet research*, vol.23, no.1, p.e17828, 2021.
- [12] OpenAI, "Introducing chatgpt," 2022. <https://openai.com/blog/chatgpt>, accessed 15 October 2023.
- [13] M. Danieli, T. Ciulli, S.M. Mousavi, G. Silvestri, S. Barbato, L. Di Natale, and G. Riccardi, "Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: Randomized controlled trial," *JMIR Ment Health*, vol.9, no.9, p.e38067, Sep 2022.
- [14] J.M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu, "Chat-counselor: A large language models for mental health support," *arXiv preprint arXiv:2309.15461*, 2023.
- [15] K. Shidara, H. Tanaka, H. Adachi, D. Kanayama, Y. Sakagami, T. Kudo, and S. Nakamura, "Automatic thoughts and facial expressions in cognitive restructuring with virtual agents," *Frontiers in Computer Science*, vol.4, p.762424, 2022.
- [16] J.B. Persons and D.D. Burns, "Mechanisms of action of cognitive therapy: The relative contributions of technical and interpersonal interventions," *Cognitive therapy and Research*, vol.9, pp.539–551, 1985.
- [17] I.D. Schmidt, B.J. Pfeifer, and D.R. Strunk, "Putting the "cognitive" back in cognitive therapy: Sustained cognitive change as a mediator of in-session insights and depressive symptom improvement.," *Journal of Consulting and Clinical Psychology*, vol.87, no.5, p.446, 2019.
- [18] M.S. Jani BD, Blane DN, "The role of empathy in therapy and the physician-patient relationship," *Forsch Komplementmed*, vol.19, no.5.
- [19] M. De Gennaro, E.G. Krumhuber, and G. Lucas, "Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood," *Frontiers in psychology*, vol.10, p.3061, 2020.
- [20] J.S. Beck, *Cognitive behavior therapy: Basics and beyond*, Guilford Publications, 2020.
- [21] L.N. Vittorio, S.T. Murphy, J.D. Braun, and D.R. Strunk, "Using socratic questioning to promote cognitive change and achieve depressive symptom reduction: evidence of cognitive change as a mediator," *Behaviour research and therapy*, vol.150, p.104035, 2022.
- [22] H. Sugiyama, M. Mizukami, T. Arimoto, H. Narimatsu, Y. Chiba, H. Nakajima, and T. Meguro, "Empirical analysis of training strategies of transformer-based japanese chat systems," 2021.
- [23] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg, *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.
- [24] H. Kumar, I. Musabirov, J. Shi, A. Lauzon, K.K. Choy, O. Gross, D. Kulzhabayeva, and J.J. Williams, "Exploring the design of prompts for applying gpt-3 based chatbots: A mental wellbeing case study on mechanical turk," *arXiv preprint arXiv:2209.11344*, 2022.
- [25] T.A. Furukawa, R.C. Kessler, T. Slade, and G. Andrews, "The performance of the k6 and k10 screening scales for psychological distress in the australian national survey of mental health and well-being," *Psychological medicine*, vol.33, no.2, pp.357–362, 2003.

- [26] S. Fitriane, M. Bruijnes, D. Richards, A. Bönsch, and W.P. Brinkman, “The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis,” IVA '20, New York, NY, USA, Association for Computing Machinery, 2020.