Investigating the Efficacy of the Natural Language Processing AI: ChatGPT in Emotion Recognition and Psychological Intervention

Bryce Petofi Towne

Sections of a Stage 1 Registered Report

Abstract

This registered report aims to compare the emotion recognition accuracy and effectiveness of psychological interventions provided by ChatGPT, an artificial intelligence (AI) language model, and human mental health professionals. The study employs a mixed-methods approach, incorporating quantitative and qualitative methodologies. Participants will be assessed on emotion recognition tasks, and a randomized controlled trial (RCT) will be conducted to compare the effectiveness of psychological interventions provided by ChatGPT and human professionals. Additionally, semi-structured interviews will be conducted to explore participants' experiences with ChatGPT and human-guided interventions. This comprehensive study design aims to provide valuable insights into the potential of AI in the field of mental health and to identify areas where improvements can be made to optimize AI-guided psychological interventions.

Key words: emotion recognition, natural language processing, mental health, psychological interventions, ChatGPT, human mental health professionals.

Introduction

The advancement of artificial intelligence (AI) and natural language processing (NLP) technologies has revolutionized various fields, including mental health care. A growing body of literature indicates that AI-powered tools may have the potential to accurately recognize emotions (Calvo et al., 2017) and provide support during psychological interventions (Fitzpatrick et al., 17; Vaidyam et al., 2019). This study aims to build upon existing research by specifically examining the effectiveness of ChatGPT, a state-of-the-art NLP AI, in emotion recognition and psychological intervention in comparison with human mental health professionals.

The choice of ChatGPT as the primary study object among other NLP AI models is based on its exceptional performance in generating human-like conversations. ChatGPT is grounded in the GPT (Generative Pre-trained Transformer) architecture, which has demonstrated superiority over previous NLP models in a variety of natural language understanding and generation tasks (Radford et al., 2018; Brown et al., 2020). Consequently, ChatGPT is expected to provide a more advanced and realistic conversational experience compared to other NLP AI models, making it a suitable candidate for exploring the effectiveness of AI-powered mental health interventions.

Additionally, ChatGPT has been pretrained on a diverse range of data, enabling it to generate coherent and contextually relevant responses (Radford et al., 2018). This ability to understand context and provide appropriate responses is essential for emotion recognition and psychological intervention, as it facilitates more accurate identification of emotional cues and tailoring of interventions to individual clients' needs (Gross et al., 2011). As a result, ChatGPT's advanced capabilities make it an appropriate model for investigating the potential of AI in mental health care settings.

There are also some potential advantages of using ChatGPT as a mental health intervention tool in comparison to human health professionals, which include several key factors: ChatGPT's objectivity as an AI language model allows for impartial assessments and interventions, potentially leading to improved mental health outcomes by avoiding personal emotions or biases that may affect human professionals (Miner et al., 2016). Additionally, individuals may find it easier to disclose sensitive or stigmatized information to ChatGPT due to the perceived non-judgmental and safe nature of AI, enabling the identification of critical factors impacting their mental health and allowing for more effective, tailored interventions (Lucas et al., 2014). Furthermore, ChatGPT's access to a vast database of information empowers it to provide a comprehensive range of evidence-based solutions for emotional distress (Vaidyam et al., 2019). By assessing the effectiveness of ChatGPT-guided interventions compared to those led by human health professionals, this study may contribute to the development of more accessible, cost-effective, and efficient mental health care solutions.

Recent research on ChatGPT-4, the latest version, a large language model with 1 trillion parameters, demonstrates its capabilities in various tasks across multiple domains, including natural language understanding, computer vision, logic reasoning, and common sense reasoning. The paper asserts that GPT-4 exhibits characteristics of artificial general intelligence (AGI), which refers to the ability to learn and execute any intellectual task that humans can perform (Bubeck, 2023).

By focusing on ChatGPT, this study hopes to add to the growing body of literature examining the potential applications of this advanced AI model, furthering our understanding of its strengths and limitations.

Progress in AI and NLP has led to the development of sophisticated conversational agents like ChatGPT, capable of engaging in human-like interactions (Radford et al., 2018). These agents have been increasingly utilized in mental health care, with promising results in applications such as mood tracking (Torous et al., 2015) and mental health triage (Miner et al., 2016). However, research directly comparing the effectiveness of AI to human mental health professionals in terms of emotion recognition and psychological intervention remains limited.

Emotion recognition is critical for psychological interventions, as it enables therapists to comprehend clients' emotional states and respond accordingly (Gross et al., 2011). Although previous studies have demonstrated that AI can recognize emotions with a high degree of accuracy (Calvo et al., 2017), the

generalizability of these findings to more advanced NLP AI models like ChatGPT requires further investigation. This study poses the following research question: To what extent can ChatGPT accurately recognize emotions compared to human mental health professionals? Based on the literature, the author proposes the following hypotheses:

Hypothesis 1 (H1): ChatGPT will demonstrate comparable accuracy in emotion recognition to human mental health professionals.

Hypothesis 0 (H0): There will be no significant difference in emotion recognition accuracy between ChatGPT and human mental health professionals.

In the context of psychological interventions, AI has shown promise in delivering cognitive-behavioral therapy (CBT) (Fitzpatrick et al., 2017), mindfulness-based interventions (Ly et al., 2015), and dialectical behavior therapy (DBT) (Rizvi et al., 2016). Despite these encouraging results, questions remain about the performance of AI compared to human mental health professionals, particularly in terms of empathy, understanding, and overall effectiveness (Ebert et al., 2015; Hollis et al., 2017). This leads to the second research question: How does the effectiveness of psychological interventions provided by ChatGPT compare to those provided by human mental health professionals? Accordingly, the author proposes the following hypotheses:

Hypothesis 2 (H2): ChatGPT-guided psychological interventions will be as effective as human-guided psychological interventions in reducing emotional distress.

Hypothesis 0 (H0): There will be no significant difference in the effectiveness of psychological interventions between ChatGPT and human mental health professionals.

To address these research questions and hypotheses, a mixed-methods approach will be employed, offering a comprehensive evaluation of ChatGPT's potential in mental health care (see Table 1 for a summary of the study design).

It is important to note that the evidence presented in this study will be correlational in nature, and causal inferences should be made with caution. Furthermore, any text that may appear to be recycled from the authors' own work or others' previous publications will be properly acknowledged and cited, in accordance with the publisher's plagiarism policy.

This research aims to contribute to the growing body of literature on AI and NLP applications in mental health care by providing a rigorous comparison between ChatGPT and human mental health professionals. By examining emotion recognition and the effectiveness of psychological interventions, this study seeks to provide valuable insights into the potential benefits and limitations of using advanced AI models, such as ChatGPT, in mental health settings. The results could inform the development of future AI-powered mental health tools and help establish best practices for their integration into clinical practice, and it may have the potential to serve as an early intervention for mental illnesses.

Methods

Ethics Information

This protocol describes research with human participants and complies with all relevant ethical regulations. The Institution: Hephaestus Education Technology Ltd. has approved the study protocol. Informed consent will be obtained from all human participants, and they will receive monetary compensation for their involvement in the study.

Pilot Data

Pilot data is not available.

Design

The study will employ a between-subjects design. Participants will be randomly assigned to one of two intervention groups: the ChatGPT-guided intervention group or the human mental health professional-guided intervention group. The randomization procedure will be performed using a computer-generated random sequence.

Data collection and analysis will not be performed blind to the conditions of the experiments. However, the investigators responsible for the outcome assessments will be blinded to group allocation.

Sampling Plan

A power analysis was conducted to determine the required sample size for this study, using the lowest available or meaningful estimate of the effect size found in the existing literature. Based on a power of 0.95 and an alpha level of 0.05, a total sample size of 180 participants (90 per group) was determined to be necessary to detect a significant difference between the ChatGPT and human mental health professional-guided intervention groups.

Inclusion criteria for participants will be adults aged 18 and above experiencing mild to moderate emotional distress. Exclusion criteria include individuals with severe emotional distress or a history of severe psychiatric disorders. The maximum feasible sample size is set at 180 participants, and data collection will cease once this number is reached. Should any participants withdraw from the study or be excluded due to technical errors, additional participants will be recruited to maintain the required sample size.

Data Inclusion/Exclusion Criteria:

the author will clearly outline all criteria for data inclusion and exclusion, including sample characteristics, technical error exclusion, and conditions under which data would be replaced. These details will be summarized in Table 1.

Analysis Plan

Data pre-processing steps will include screening for data entry errors, assessing normality, and handling missing data using multiple imputation techniques. The primary analysis will involve mixed-effects models to compare changes in emotional distress scores between the ChatGPT-guided intervention group and the human-guided intervention group, controlling for baseline distress levels and potential covariates (e.g., age, gender). Appropriate correction for multiple comparisons will be applied.

Data Availability

Upon acceptance of the Stage 2 manuscript, raw data and materials will be shared publicly. The data will be deposited in a recognized data repository (e.g., Open Science Framework) and made accessible under a Creative Commons Attribution License.

Code Availability

All code associated with this study will be made publicly available upon acceptance of the Stage 2 manuscript. The code will be shared on a recognized code-sharing platform (e.g., GitHub) and made accessible under an appropriate open-source license. Links to the live versions of the code for data

simulation, power analyses, and pilot data analysis will be provided. The code will be made available for peer-review but can be placed under public embargo until Stage 2 acceptance.

Limitations and Future Directions

Despite the rigor and careful planning of this study, there are several limitations that warrant consideration and provide directions for future research.

1. Generalizability: The study's findings may have limited generalizability due to the specific population included, which focuses on individuals experiencing mild to moderate emotional distress. Future research should investigate the effectiveness of ChatGPT and other AI-driven interventions in diverse populations, including those with more severe mental health concerns or specific disorders.
2. Lack of long-term follow-up: This study focuses on the immediate effects of the ChatGPT and human mental health professional-guided interventions on emotional distress. The long-term effects of these interventions, including potential relapse rates, remain unknown. Future studies should incorporate long-term follow-up assessments to better understand the sustainability of any observed treatment gains.
3. Limited range of intervention types: This study focuses on the effectiveness of ChatGPT in delivering psychological interventions in general, without specifying a particular therapeutic approach (e.g., CBT, DBT, or psychodynamic therapy). Further research should investigate the efficacy of ChatGPT in delivering specific evidence-based therapies and compare these findings to human mental health professionals specialized in those approaches.
4. Emotion recognition assessment: The study assesses the emotion recognition capabilities of ChatGPT and human mental health professionals using a limited set of emotion recognition tasks. Future research should employ a wider range of tasks, including those assessing the ability to recognize nuanced or complex emotions, as well as the recognition of emotions from nonverbal cues (e.g., facial expressions, vocal tone).
5. Therapeutic alliance: The current study does not address the potential impact of the therapeutic alliance on treatment outcomes. The therapeutic alliance is a crucial factor in the success of psychological interventions, and it is unclear whether AI-driven interventions like ChatGPT can establish a similar alliance with clients as human mental health professionals. Future research should examine the role of the therapeutic alliance in ChatGPT-guided interventions and its impact on treatment outcomes.
6. Ethical considerations: While this study adheres to strict ethical guidelines, the broader ethical implications of using AI-driven interventions in mental health care warrant further exploration. Issues related to data privacy, informed consent, and the potential for AI bias should be considered in future research and guidelines for AI integration in mental health care.

This study expects the identified limitations and future directions highlight areas for further investigation can contribute to the development of more effective, accessible, and ethically sound AI-driven mental health interventions.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
2. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

3. Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649-685.
4. Ebert, D. D., Zarski, A. C., Christensen, H., Stikkelbroek, Y., Cuijpers, P., Berking, M., & Riper, H. (2015). Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PloS one, 10(3), e0119895.
5. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR mental health, 4(2), e7785.
6. Gross, J. J., Sheppes, G., & Urry, H. L. (2011). Emotion generation and emotion regulation: A distinction we should make (carefully). Cognition and emotion, 25(5).
7. Hollis, C., Falconer, C. J., Martin, J. L., Whittington, C., Stockton, S., Glazebrook, C., & Davies, E. B. (2017). Annual Research Review: Digital health interventions for children and young people with mental health problems–a systematic and meta-review. Journal of Child Psychology and Psychiatry, 58(4), 474-503.
8. Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior, 37, 94-100. https://doi.org/10.1016/j.chb.2014.04.043
9. Ly, K. H., Topooco, N., Cederlund, H., Wallin, A., Bergström, J., Molander, O., ... & Andersson, G. (2015). Smartphone-supported versus full behavioural activation for depression: a randomised controlled trial. PloS one, 10(5), e0126559.
10. Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA internal medicine, 176(5), 619-625.
11. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
12. Rizvi, S. L., Hughes, C. D., & Thomas, M. C. (2016). The DBT Coach mobile application as an adjunct to treatment for suicidal and self-injuring individuals with borderline personality disorder: A preliminary evaluation and challenges to client utilization. Psychological services, 13(4), 380.
13. Torous, J., Staples, P., Shanahan, M., Lin, C., Peck, P., Keshavan, M., & Onnela, J. P. (2015). Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR mental health, 2(1), e3889.
14. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64(7), 456-464.

Acknowledgements

The author would like to thank ChatGPT for its contributions in proofreading and improving the language clarity and structure of this report, especially in translating Chinese into English.

Author Contributions

Bryce Petofi Towne is the sole contributor to this study. Bryce Petofi Towne conceived the study design, developed the research questions and hypotheses, and prepared the research protocol. Bryce

Petofi Towne was responsible for drafting the manuscript, revising it critically for important intellectual content, and giving final approval for the version to be published. Bryce Petofi Towne also agrees to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing Interests

The authors declare no competing interests, financial or non-financial, related to this study.

Table 1. Design Table

Question	Hypothesis	Sampling Plan (e.g., power analysis)	Analysis Plan	Interpretation given to different outcomes
Q1	H1: ChatGPT will demonstrate comparable accuracy in emotion recognition to human mental health professionals. H0: There will be no significant difference in emotion recognition accuracy between ChatGPT and human mental health professionals.	Power analysis based on the lowest available or meaningful effect size from the literature, aiming for a power of 0.95.	Independent samples t-test comparing the emotion recognition accuracy between ChatGPT and human mental health professionals.	If $p < 0.05$: Evidence supporting H1, suggesting that ChatGPT has comparable emotion recognition accuracy to human professionals. If $p \geq 0.05$: Inconclusive evidence; neither H1 nor H0 can be confirmed.
Q2	H2: ChatGPT-guided psychological interventions will be as effective as human-guided psychological interventions in reducing emotional distress. H0: There will be no significant difference in the effectiveness of psychological interventions between ChatGPT and human mental health professionals.	Power analysis based on the lowest available or meaningful effect size from the literature, aiming for a power of 0.95.	Mixed-design ANOVA with intervention type (ChatGPT vs. human professional) as the between-subjects factor and time (pre-intervention vs. post-intervention) as the within-subjects factor.	If $p < 0.05$ for the interaction effect: Evidence supporting H2, suggesting that ChatGPT-guided interventions are as effective as human-guided interventions in reducing emotional distress. If $p \geq 0.05$ for the interaction effect: Inconclusive evidence; neither H2 nor H0 can be confirmed.

Inclusion criteria:

1. Participants aged 18 or older.
2. Participants experiencing mild to moderate emotional distress.

Exclusion criteria:

1. Participants with severe mental health disorders.
2. Participants currently receiving psychological treatment.
3. Data exclusion and replacement procedures:
4. Participants who do not complete the entire intervention will be excluded from the analysis.
5. Technical errors leading to incomplete or corrupted data will result in the exclusion of the respective participant's data.
6. Excluded participants' data will be replaced by new participants who meet the inclusion criteria and do not meet any exclusion criteria.

Supplementary information

Detailed Experimental Procedures

In this Supplementary Information section, the author provide a comprehensive description of the experimental procedures to ensure that other researchers can replicate this study with precision. The detailed methodology includes participant recruitment, intervention administration, data collection, and data analysis procedures, as well as an overview of the study plan.

Participant Recruitment

Participants will be recruited through online forums, social media platforms, and local mental health organizations. Advertisements will specify the inclusion and exclusion criteria, and potential participants will be directed to an online screening questionnaire. The questionnaire will assess eligibility based on age, current mental health status, and ongoing psychological treatment.

Human mental health professionals (n = 55) will be recruited through professional networks, clinical organizations, and online forums. Eligible professionals must hold a minimum of a master's degree in psychology, counseling, social work, or a related field and have at least two years of clinical experience. All participating professionals will provide documentation of their qualifications and licensure.

Intervention Administration

Eligible participants will be randomly assigned to either the ChatGPT-guided or human mental health professional-guided intervention group. Both interventions will be conducted through a secure online platform. The ChatGPT-guided intervention will utilize the latest version of the ChatGPT AI, while the human-guided intervention will be led by licensed mental health professionals. The interventions will follow standardized protocols based on evidence-based psychological techniques, and each session will last for 50 minutes. Participants will attend weekly sessions for a total of eight weeks.

Data Collection

Emotion recognition accuracy data will be collected using a standardized emotion recognition task. Participants will be assessed before and after the intervention period. Additionally, emotional distress levels will be measured using self-report questionnaires administered before and after the intervention.

Data Analysis

Data analysis will be performed using the specified statistical tests in the Design Table (Table 1). Power analysis and other necessary calculations will be conducted using appropriate statistical software.

Detailed Study Plan

To investigate the research questions and hypotheses, a mixed-methods study will be employed, comprising both quantitative and qualitative approaches. The study will be conducted in three stages, as outlined below.

- **Stage 1: Emotion Recognition Task**

A sample of mental health professionals (n = 55) and ChatGPT will be assessed on their ability to recognize emotions in a series of textual transcripts. These transcripts will be derived from therapy sessions, and any identifying information will be removed to maintain confidentiality. The transcripts will be coded for emotional content by an independent team of experts who will establish a "gold standard" for emotion recognition

Both the mental health professionals and ChatGPT will be required to identify the emotions present in the transcripts and assign a confidence rating to their responses. The accuracy of emotion recognition will be calculated by comparing their responses to the gold standard. Statistical analysis, such as t-tests or non-parametric alternatives, will be employed to determine whether there are significant differences in emotion recognition accuracy between ChatGPT and human professionals.

- **Stage 2: Psychological Intervention Experiment**

A randomized controlled trial (RCT) will be conducted to compare the effectiveness of psychological interventions provided by ChatGPT and human mental health professionals. Participants experiencing mild to moderate emotional distress (n = 180) will be randomly assigned to either the ChatGPT intervention group or the human intervention group. Both groups will receive an 8-week intervention program, comprising one session per week.

Participants' emotional distress levels will be assessed using validated self-report measures, such as the Patient Health Questionnaire-9 (PHQ-9) and the Generalized Anxiety Disorder-7 (GAD-7), at pre-intervention, post-intervention, and 3-month follow-up. Changes in emotional distress scores will be compared between the two groups using mixed-effects models or other appropriate statistical.

- **Stage 3: Qualitative Evaluation**

A subset of participants (n = 20) from each intervention group will be selected using purposive sampling to ensure diversity in terms of age, gender, and severity of emotional distress. They will be invited to participate in semi-structured interviews to explore their experiences with the ChatGPT and human-guided interventions. The interviews will be audio-recorded, transcribed, and analyzed using thematic analysis to identify patterns and themes related to the perceived effectiveness, empathy, understanding, and overall satisfaction with the interventions.

This comprehensive and detailed experimental procedure, along with the three-stage study plan, ensures that the study is methodologically sound and can be replicated by other researchers. The inclusion of clear criteria for the recruitment of human mental health professionals and participants, as well as the use of standardized tasks and intervention protocols, strengthens the validity and reliability of the findings.

Detailed Information about the ChatGPT-guided Intervention:

The ChatGPT-guided intervention will be based on well-established, evidence-based techniques and approaches that are commonly used in human-guided interventions. These techniques and approaches will be tailored to suit the AI's capabilities and ensure a fair comparison with human-guided interventions.

Cognitive Behavioral Therapy (CBT) Techniques: ChatGPT will be trained to employ CBT techniques such as cognitive restructuring, problem-solving, and behavioral activation. The AI will help participants identify and challenge negative thought patterns, develop coping strategies, and engage in activities that promote well-being.

Mindfulness and Relaxation Techniques: ChatGPT will provide guided mindfulness exercises and relaxation techniques, such as deep breathing, progressive muscle relaxation, and visualization, to help participants manage stress and anxiety.

Psychoeducation: The AI will be equipped to provide information and resources about mental health issues, coping mechanisms, and self-help strategies to participants, promoting mental health literacy.

Goal Setting and Monitoring: ChatGPT will assist participants in setting realistic, achievable goals and tracking their progress throughout the intervention. This process will facilitate the development of self-efficacy and motivation for change.

To ensure that the ChatGPT-guided intervention is evidence-based and comparable to human-guided interventions, a team of mental health professionals and AI experts will collaborate to develop a standardized intervention protocol. The protocol will be based on current best practices in mental health care and will be adapted to suit the AI's capabilities**.**

Potential Biases, Limitations, and Strategies for Minimization:

Demand Characteristics: Participants might change their behavior or responses based on their perception of the study's purpose. To minimize this bias, the study will use a double-blind design, ensuring that both participants and researchers are unaware of the group assignments.

Experimenter Bias: Researchers might unintentionally influence the results through their expectations or interactions with participants. To address this, the study will use clearly defined protocols, and researchers will be trained to adhere to these protocols consistently.

Attrition Bias: Participants who drop out during the study might differ from those who complete it, leading to biased results. To minimize this bias, the study will use intention-to-treat analysis and monitor dropout rates throughout the intervention.

Social Desirability Bias: Participants might provide responses that they perceive as more socially acceptable, rather than accurate reflections of their experiences. To minimize this bias, anonymous self-report questionnaires will be used, and participants will be encouraged to provide honest feedback.

Incorporating Additional Objective Measures of Emotional Distress:

To strengthen the validity of the study's findings, the following objective measures can be incorporated alongside self-report questionnaires:

Physiological Measures: Assessments of heart rate variability, cortisol levels, and other biomarkers can provide objective indicators of stress and emotional distress.

Behavioral Observations: Participants' behavior during the intervention, such as their engagement, attentiveness, and nonverbal cues, can be observed and analyzed to provide additional insights into their emotional state.

Ecological Momentary Assessment (EMA): EMA involves repeated real-time assessments of participants' emotional states and experiences in their natural environment, reducing the impact of recall bias and providing a more accurate representation of their emotional distress.