Exploring Mental Health Using LLMs: Comparison between ChatGPT and Gemini

# Manjusha Valluri

11638920

*Department of Information Science University of North Texas*

**ABSTRACT**

Mental health has been a major concern in the current generation which is affecting millions of individuals across different demographics. It is an important aspect for overall well-being of an individual. The most common and typical method of an individual to check on their mental health status is to consult a doctor or a psychiatrist. In today’s busy world, it is hard to find time for face-to-face interactions with mental health professionals. There is a real need for easy and private ways to keep track of mental health of people. All these concerns motivated to create a model that can generate accurate answers related to mental health questions. With AI becoming a bigger part in our lives, we have a chance to use it to analyze mental health. Since ChatGPT and Gemini are actively used by people, these two models are chosen. By comparing their performance and efficiency, the project mainly aims to increase the implementation of these models for better support to the individual who are dealing with mental health issues. This research mainly focuses on evaluating how well these LLMs can spot and understand mental health concerns and problems. The project ensures that these models are useful and helpful space where people feel comfortable sharing their feelings and get support or guidance thereby making them feel better mentally even if they are super busy.

**INTRODUCTION AND STATEMENT OF PROBLEM**

Mental health is a critical problem in the present world. Factors such as depression, stress, anxiety, and other mental health disorders have become prevalent especially in the con- text of modern lifestyles, societal pressures, and other factors. A lot of awareness has been developed among the individuals regarding this. They started to check on their mental health and wanted to track quite frequently. The idea of the project is basically developed from understanding the importance of analyzing and supporting mental health using AI technologies. Recent years have witnessed remarkable advancements in large language models (LLMs) and these technologies have been trending in many areas providing accurate results. It is crucial

to evaluate these models to know how well these models can help and guide people dealing with mental health issues. Our moto is to use LLMs especially ChatGPT and Gemini to create a place where people can freely talk about their mental health stuff and get accurate responses. This project makes mental health conversations simpler and more helpful.

**REVIEW OF LITERATURE**

Literature Survey on Natural Language Processing and Mental Health Detection: Natural Language Processing (NLP) has emerged as a powerful tool in the detection and analysis of the mental health condition. Much of the research on its different strands carried out takes a focus on linguistic style, discourse patterns, and topic modeling in efforts to detect markers of mental health disorders. LLMs such as ChatGPT and Gemini have been celebrated for their effectiveness in text generation, warranting huge potential in aiding the early detection of mental health problems. This literature survey reviews existing research in the field of NLP in regard to mental health, and through this, it gains insights into the effectiveness of these technologies.

Using Machine Learning and Natural Language Processing in the Diagnosis of Mental Health: Mental health illness pre- dictions have been carried out using machine learning and NLP approaches in the past. In fact, machine learning algorithms have shown potential in the identification of mental health conditions in the study by Vaishnavi et al. with reference to the analysis of a variety of data sets. Their work emphasized that diverse algorithms also take part in predicting outcomes and, therefore, showed that computational approaches play a very significant role in mental health. In a separate development, Iyortsuun et al. (2023) appraised methodologies of machine learning and deep learning in the diagnosis of mental health. In their work, the paper emphasizes this fact, showing how wide these techniques apply in a spectrum of applications while try- ing simultaneously to reflect on the challenges and limitations of the same. Role of Large Language Models in Analysis on Mental Health: The use of LLMs has become popular in the

past few years with respect to mental health detection. In the study by Le Glaz et al. (2021), machine learning, NLP, and mental health all play a significant role in offering insights into the current state of research. The review points out the potential of LLMs in understanding and analyzing linguistic data to contribute to early detection and intervention in mental health. Yang et al. (2023) studied interpretable mental health analysis using LLMs, which deepened the understanding of how such models process and interpret results in the context of mental health detection. It is for that reason that these findings will, to a very reasonable extent, drive how LLMs are perceived and relied upon in medical applications.

A related work by Adhikary et al. (2024) focused on the use of LLMs in summarizing mental health counseling sessions. Their research findings show that LLMs like ChatGPT can capture the summarization of the counseling session appropri- ately and extract the relevant information that will help in the development of an early detection strategy. Guo et al. (2024) also conducted a systematic review focused on LLMs for mental health and further share insights about the applicability of these models to early detection and intervention. It also validates the claim that LLMs can be used in a mental health analysis and diagnosis tool. Comparing LLMs: ChatGPT and Gemini: Rane et al. (2024) comparatively studied the per- formance, architecture, capabilities, and implementations of two major LLMs—Gemini and ChatGPT. This study has been useful in understanding the pros and cons of these models toward making a choice applicable in a mental health scenario. Lamichhane (2023) similarly evaluated the effectiveness of ChatGPT in the scope of NLP-based mental health applica- tions and pointed out how the model might be used for such applications, and also some failures or shortcomings in mental health. NLP and LLMs have shown a lot of potentials in early detection and intervention in mental health study. The works cited within further indicate an emerging area of the use of larger language models, such as ChatGPT and Gemini, to understand and interpret linguistic data on mental health. It is only with the help of the existing research that future projects can fine-tune these approaches and work towards better outcomes of mental health detection and care.

1. *Related Work*

Related literature on mental health detection assessed the potential of large language models in identifying linguistic patterns in text-based communication. Initial studies of re- search used machine learning algorithms to predict mental health conditions, where the work of Vaishnavi et al. (2022) showed the ability of the algorithms to undertake an analysis of datasets to realize the patterns associated with mental health. Iyortsuun et al. (2023) conducted a critical review of machine learning and deep learning methodologies and their challenges and shortcomings in this field. With the rise in LLMs, some attention has been drawn to their capability to understand and interpret text-based information in a mental health context. Evidence to this capability is provided by a systematic re- view of Le Glaz et al. (2021) at the crossroads of machine

learning, NLP, and mental health for early detection through linguistic analysis. Interpretable mental health analysis with LLMs was also underlined in Yang et al. (2023), where the importance of knowing how the models process and generate their responses was critically discussed. For example, insights into the strengths and limitations of such tools are offered in studies by Rane et al. (2024) on the performance of ChatGPT and Gemini. Important ethical considerations will still be in place, and researchers will emphasize such issues as privacy, data security, and bias. Such studies underline the fast-growing landscape of LLMs within mental health detection and the need for its responsible application and continuous research.

**OBJECTIVES OF STUDY**

We explore and evaluate the effectiveness of large language models, such as ChatGPT and Gemini, in pattern recognition and the identification of features that determine mental health issues in textual communication, focusing on a comparative model performance analysis.

* + Analysis of Linguistic Features: Investigating the linguis- tic characteristics, which include syntactic, semantic, and discourse patterns associated with mental health problems in text-based communication.
  + ChatGPT and Gemini Evaluation of Performance: Deter- mine the accuracy and reliability of ChatGPT and Gemini in recognizing and responding to language patterns that might indicate an individual with mental health issues.
  + Comparing Embeddings: Calculate the embeddings for both ChatGPT and Gemini, and then compare them with embeddings obtained from the ground truth set, respectively, with the purpose of understanding which model’s embeddings are more coherent with embeddings from real-world data.
  + Benchmarking and Cosine Similarity: Compare the per- formance of ChatGPT and Gemini on a relevant dataset and assess their accuracy using a cosine similarity score as a quantitative measure of the models’ ability in detect- ing mental health-related language.
  + Research on Ethical Implementation: Discuss ethical con- cerns and implications with the use of Large Language Models in detecting mental health issues, such as privacy, bias, informed consent, and wider social consequences.

These objectives have been set up so as to allow an ordered, structured approach toward the evaluation of the role of LLMs in mental health analysis and to balance performance assessment with ethical considerations, in a manner that will ensure responsible use of these technologies.

**RESEARCH DESIGN**

Fig.1 shows research design of the project which consists of several steps such as Dataset collection, Data preprocessing, Exploratory data analysis,ChatGPT response, Gemini response generation, converting original responses, ChatGPT Responses and Gemini responses into embeddings and calculating cosine similarity score.

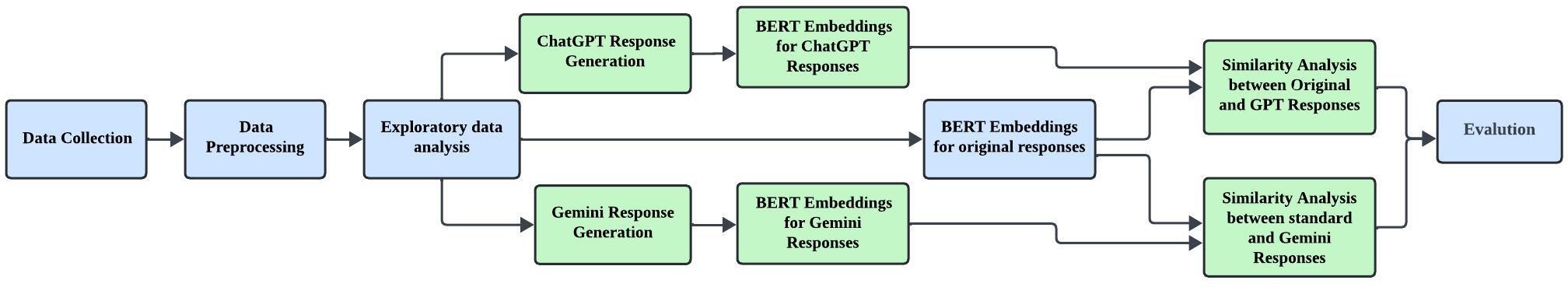


Fig. 1. Flow chart

**DATA COLLECTION**

The [”mental-health-conversational-data”](https://huggingface.co/datasets/alexandreteles/mental-health-conversational-data) dataset is gathered from Hugging face.

It consists of discussions between therapists/doctors and patients about mental health. The dataset has 661 rows and three columns such as context, knowledge, and Response. It provides insights on mental health support through interaction with doctors/therapists. The column named “Context” is about prompts or questions raised by the patients and the column “Response” is a reply given by the doctors for the respective prompt provided by the patients. The “Knowledge” columns consist of additional information to understand the provided prompt category, which helps patients and doctors for greater interaction. The dataset can be utilized while working on projects related to natural language processing (NLP), and mental health support.

**EXPLORATORY DATA ANALYSIS**

1. *Data Preprocessing*

After collecting the Dataset, the data is loaded using Panda’s library and converted it into a data frame. The initial prepro- cessing steps involve checking for null values as there are no null values in the dataset, but it consists of empty strings and “None” values. The None values and empty strings are removed in both context and response.

1. *Exploratory Data Analysis*

EDA was conducted on the data to derive the insights that help for a better understanding of research. The most frequently used words in context and response are visualized using Word clouds to understand the conversations and get insights into commonly discussed topics. By knowing the most frequent words, we can understand the primary areas of the conversations that can be focused. The value count of the



Fig. 2. Word cloud for Contexts and Responses

“Knowledge column” was analyzed to know the unique cate- gories of the knowledge column. We can observe that there are 79 unique categories. Most of the conversations are related to general types such as greeting, causal, about, default, goodbye,

sad, done, help, and happy which have a value count greater than 20. Overall, the EDA helps to get more insights into the dataset and understand the conversation characteristics. These insights can be helpful for further analysis, and development of models, NLP, and mental health support applications.

**HYPOTHESES FOR THE STUDY**

With the recent advancement in the field of artificial intel- ligence, large language models have been developed that have human-like text generation capabilities. The popular large lan- guage models used in various applications include OpenAI’s ChatGPT and Google’s Gemini. However, the crucial area that needs to be achieved is the performance comparison of models, especially in domains such as mental health support. The main aim of the project is to compare the performance of LLMs like ChatGPT and Gemini in response generation to the respective context or prompt given. We have evaluated how similar the model is in generating responses to the original response that is present in the dataset. Based on the cosine similarity score, we hypothesize that ChatGPT performance is great compared to Gemini in generating responses related to mental health conversations. In addition to this, we also anticipate that ChatGPT understanding and text generation abilities resulted in a higher similarity score compared to Gemini. For testing this hypothesis, we utilized the dataset containing original responses and compared the responses of ChatGPT and Gemini by converting them into Embeddings using Bert and we calculated the cosine similarity score between original and ChatGPT responses, original and Gemini responses and compared mean cosine similarity of both scores. From this hypothesis, we will get more insights and the effectiveness of ChatGPT and Gemini response generation in the context of mental health conversation. These insights can help in AI model selection and deployment in mental health support systems.

**MODEL ANALYSIS AND EXPLORATION**

This project is an opportunity to analyze the responses of two very powerful language models, that is, CHATGPT and GEMINI. This Project uses the power of such model to investigate whether they can generate responses to the conversational prompt that falls within the scope of mental health.

1. *ChatGPT Response Generation*

We harnessed the power of openAI ChatGPT to craft the responses for each context. ChatGPT 3.5 Turbo served as the base model to generate responses. The OpenAI interface generates responses to each context from the mental health conversational dataset, enabling ChatGPT to communicate with secret keys. The responses generated by ChatGPT will be saved as ChatGPT responses for analysis. In this way, it could be ensured that there is maximal exploration with ChatGPT in understanding and responding to a wide diversity of prompts related to mental health.

1. *Gemini Response Generation*

The project interfaces with the Gemini Pro model, which then produces responses to each context in the dataset. The whole process begins with an authentication request that is sent to the Gemini Pro model through secret keys from the GenAI platform. It authenticates the user and safely provides access to the capabilities of the model. Gemini Pro processes contexts and comes up with an appropriate response. Gener- ated Responses are automatically saved as Gemini Responses for further review.

1. *Bert Embeddings*

The next step is to convert the original responses, ChatGPT responses, and Gemini responses from strings to lists of strings. This procedure needs to be made in order to get the data format that will be further analyzed at a much higher level. The Auto Model and Auto -Transformer classes from hugging face is convenient way to load the BERT model, to covert the text to embeddings. Once the responses are in list format, the procedure involves encoding them into numerical embeddings using BERT (Bidirectional Encoder Representa- tions from Transformers). By encoding the responses into nu- merical embeddings using BERT, we can capture the semantic meaning and context of each response.

1. *Cosine Similarity Score*

The Standard responses embeddings, ChatGPT embeddings and Gemini embeddings are compared using Cosine Similarity Score. Cosine Similarity score is Widely used in Natural Language Processing to compare the angle between two numerical vectors.It quantifies the cosine of the angle between the vectors, providing a measure of similarity irrespective of their magnitudes. cosine similarity score of 1 means the two vectors are perfectly in line and point in similar directions in high-dimensional space. It’s a representation of vectors for the maximum similarity degree. A score of -1 for cosine similarity, in turn, indicates that the vectors point in opposite directions: one in the exact opposite direction of the other. This would suggest the maximum dissimilarity between the vectors. If they are perpendicular (orthogonal) to each other, their cosine similarity would be 0, meaning there won’t be any similarity in this n-dimensional space between the vectors.

The cosine similarity score for ChatGPT and original re- sponses is 93 % and the cosine similarity score for Gemini and original responses is 88 %.

**DATA ANALYTICS**

In this project, data analytics involves data collection, data analysis, and data interpretation to get more insights into data for decision-making. Data analytics involves several tasks like Collecting data, data cleaning, Exploratory data analysis, visualizing data, statistical analysis and modeling based on the objectives and requirements of the project.

**DATA VISUALIZATION AND RESULTS REPORT**

Fig 3 compares Standard responses with responses by ChatGPT and Gemini using the cosine similarity metric. It is observed that the cosine similarity for ChatGPT is 93, which is extremely high compared to the one Gemini has. This essentially means that ChatGPT-generated responses are highly closer to standard ones and, therefore, the perfect criterion for accuracy in capturing and replicating the text- based patterns about mental health.

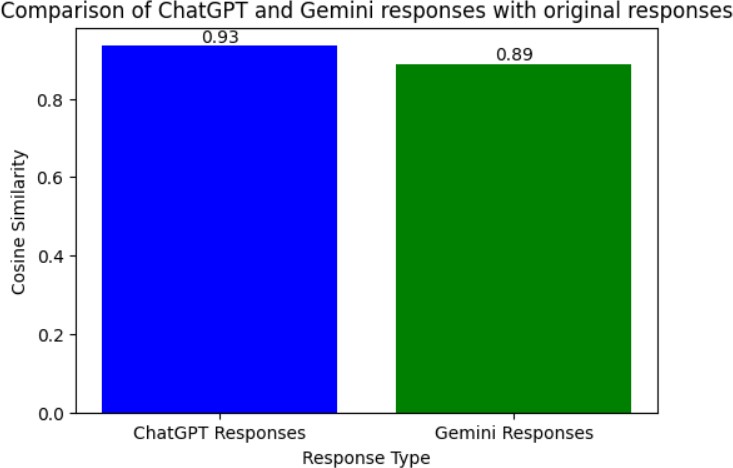


Fig. 3. Bar chart for Cosine similarity

Let’s compare the embeddings of ChatGPT and Gemini to the standard through Fig.4, a heatmap. Comparing the two heatmaps in Fig. 4, one can see that the ChatGPT is much more yellow than the Gemini, which is what represents higher cosine similarities under the color scale next to the heatmap. On the other hand, the Gemini heatmap shows both green and yellow, so cosine similarities are lower compared to ChatGPT. This visualization implies that ChatGPT is more similar to the standard embeddings, hence being more similar in patterns relative to the standard dataset.

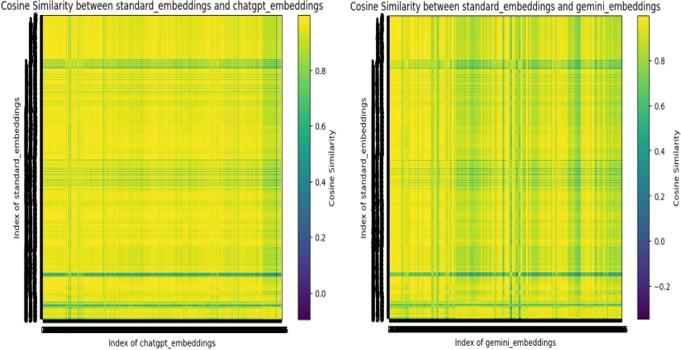


Fig. 4. Heat Map

By the elbow method, we were able to make a decision on the number of clusters to use. From the three line graphs above, we realize that the line actually flattens between 80 and

100. This clearly shows that the optimal number of clusters is

80 as this is the point where adding more clusters no longer significantly improves the model’s performance.

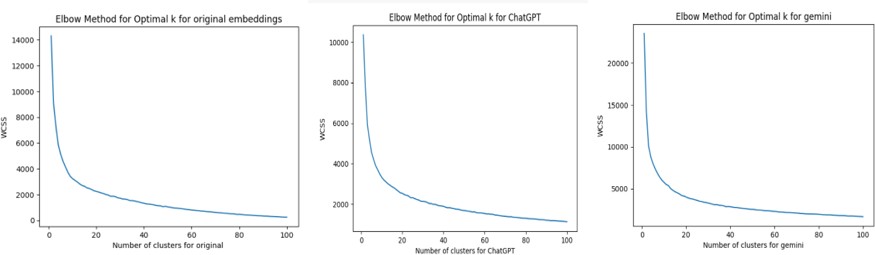


Fig. 5. Comparision of models using Elbow method

We did the cluster analysis to obtain the ideal amount of clusters the points would have. We used a clustering algorithm in one of our dataset columns and received the results of 80 clusters, matching the expected optimal amount from the elbow method. It reassures us that the appearance of 80 clusters in our data is natural.

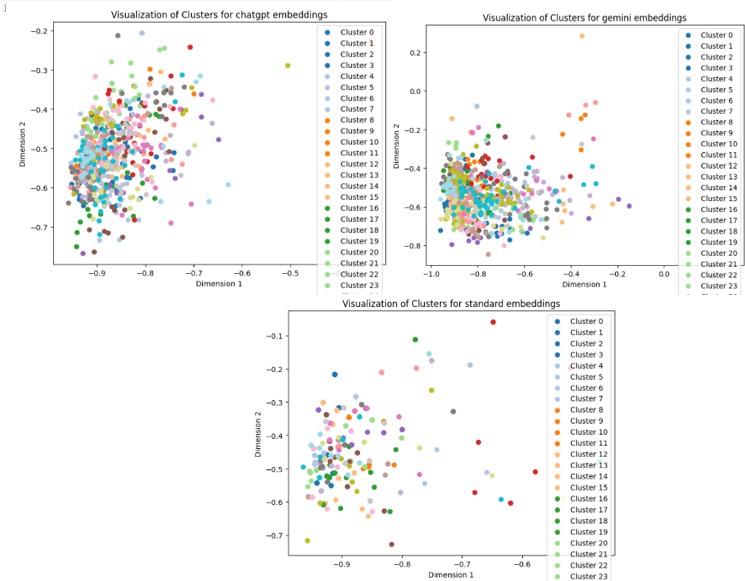


Fig. 6. Clustering of Embeddings

**CONCLUSION**

The following section is regarding the main findings of this study, which is the performance comparison of ChatGPT and Gemini in identifying mental health issues through text communication. The obtained results are provided with the aid of visual comparisons and similarity scores, which provide an overall evaluation of the models.

* From the test results and data analysis, it has been found that ChatGPT outperforms Gemini in identifying the patterns and linguistic features associated with mental health issues. Its performance is consistent over different contexts, which leads to high accuracy in detecting rele- vant information. The kind of consistency exhibited could therefore make it dependable for application in different scenarios, indicating robustness in handling data.
* One outstanding result of this experiment was with the similarity score, where ChatGPT hit an impressive 93, a score that outperformed Gemini’s by a huge margin. In so far as a high similarity score, it goes to indicate that Chat- GPT managed to closely mimic the original responses, indicating a capability to both grasp and reproduce the patterns typical of exchanges regarding mental health.
* To deeply study the performance difference between ChatGPT and Gemini, we made an embedding compar- ison of the two models. Such embeddings, numerical representations of text, were compared with standard embeddings derived from original responses to evaluate the similarity and variation in output.
* Various visualization techniques have been used, includ- ing scatter plots, bar charts, and cluster analysis, to make these relations more clear between the original embeddings and those created by ChatGPT and Gemini. It clearly showed in the visualizations where differences in performance lied—either close to the standard em- beddings in the case of ChatGPT or with much bigger deviation in the case of Gemini.
* The two-dimensional scatter plots of the embeddings dis- tribution made it possible for similarities and anomalies in the embeddings of the original, ChatGPT, and Gemini to be seen. The two-dimensional scatter plots showed that the embeddings of ChatGPT are more tightly grouped around the original embeddings, which would signify more accuracy in the replication of original patterns.
* Similarity scores across three sets of embeddings were compared using bar charts, which confirmed that Chat- GPT is outperforming. That would further mean that since the similarity scores of ChatGPT are higher, the responses it provides are closer in alignment to the expected patterns.
* The cluster analysis further motivated consistency in ChatGPT, as the embeddings very often formed tight and clearly bounded clusters, while the clusters in Gemini were much more scattered, showing less consistency in its output.

These findings can be understood to mean that the capacity of ChatGPT to capture and replicate linguistic patterns regarding mental health issues is greater than that of Gemini. In sum- mary, the consistent performance across contexts, high simi- larity scores, and well-defined clusters reveal that ChatGPT is more consistent and reliable in detecting mental health in text- based communication. However, further studies are required, which not only include other contexts but also discuss other ethical issues, such as privacy, data security, and bias.

**REFERENCES**

1. Vaishnavi, K., Kamath, U. N., Rao, B. A., Reddy, N. S. (2022). Predicting mental health illness using machine learning algorithms. In Journal of Physics: Conference Series (Vol. 2161, No. 1, p. 012021). IOP Publishing.
2. Iyortsuun, N. K., Kim, S. H., Jhon, M., Yang, H. J., Pant, S. (2023, January). A review of machine learning and deep learning approaches on mental health diagnosis. In Healthcare (Vol. 11, No. 3, p. 285). MDPI.
3. Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., ... Lemey, C. (2021). Machine learning and natural language processing in mental health: systematic review. Journal of medical Internet research, 23(5), e15708
4. Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., Ananiadou, S. (2023, December). Towards interpretable mental health analysis with large language models. In The 2023 Conference on Empirical Methods in Natural Language Processing.
5. Lamichhane, B. (2023). Evaluation of chatgpt for nlp-based mental health applications. arXiv preprint arXiv:2303.15727.
6. Adhikary, P. K., Srivastava, A., Kumar, S., Singh, S. M., Manuja, P., Gopinath, J. K., ... Chakraborty, T. (2024). Exploring the Efficacy of Large Language Models in Summarizing Mental Health Counseling Sessions: A Benchmark Study. arXiv preprint arXiv:2402.19052.
7. Mart´ınez-Castan˜o, R., Htait, A., Azzopardi, L., Moshfeghi, Y. (2021). BERT-based transformers for early detection of mental health illnesses. In Experimental IR Meets Multilinguality, Multimodality, and Interac- tion: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12 (pp. 189- 200). Springer International Publishing.
8. Lawrence, H. R., Schneider, R. A., Rubin, S. B., Mataric, M. J., McDuff,

D. J., Bell, M. J. (2024). The opportunities and risks of large language models in mental health. arXiv preprint arXiv:2403.14814.

1. 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.
2. Mazumdar, H., Chakraborty, C., Sathvik, M. S. V. P. J., Panigrahi, P. K. (2023). GPTFX: A Novel GPT-3 Based Framework for Mental Health Detection and Explanations. IEEE Journal of Biomedical and Health Informatics.