Evaluating LLMs to Measure Psychiatric Functioning

Hannah Park hp2501@nyu.edu

George Zhou gz2214@nyu.edu

Jin Choi jkc9890@nyu.edu

Stephen Spivack ss7726@nyu.edu

Abstract

The rapidly growing interest in the Large Language Model (LLM) in various domains brings the question of how the model will perform in the field of psychiatry. This paper presents an approach to employing LLMs for psychiatric diagnoses, leveraging clinically validated case vignettes. We aimed to develop an LLM that can diagnose disorders to patients and assess its diagnostic accuracy. We compared different types of open-source LLM and analyzed their diagnostic accuracy through different prompt engineering strategies. The study utilized three data sets on Anxiety, Stress, and Mood disorders. Although the LLMs had input token limitations, we sought few approaches such as using high-capacity LLMs, summarizing guidelines, or using the expertise of clinicians to only use essential features of guidelines. The study's findings reveal that the model's performance parallels the patterns of choices made by human clinicians, with notable accuracy in diagnosis but high false positive rates for "no diagnosis" scenarios. Future work includes translating and reformatting multilingual data and exploring potential biases. This research signifies a step towards AI applications in medical diagnosis, emphasizing the improvement of clinical decision-making processes.

1 Introduction

In recent years, the field of artificial intelligence and natural language processing has witnessed remarkable advancements, particularly with the advent of Large Language Models (LLMs). These LLMs, such as Open AI's GPT-3, Meta's Llama2, Mistral, have demonstrated exceptional power in understanding and generating human language. With their capabilities, LLMs rapidly exert influences across various domains, including media, education and healthcare. In healthcare, large language models (LLMs) have exhibited their potential to contribute in clinical settings, with instances such as a large dialog LLM like ChatGPT successfully passing part of the US medical licensing exam [1]. Hence, the application of LLMs to enhance diagnostic precision and support clinical decision-making has emerged as a vibrant area of research in the field of medicine [2].

The global demand for mental health services has been on a steady rise, yet there exists a considerable shortage of mental health specialists to meet this need. This shortfall is particularly pronounced in humanitarian emergencies, low-income countries, and other resource-constrained regions [3]. To address this treatment gap, there exists a compelling clinical need for innovative approaches to psychiatric diagnosis and assessment.

In response to this demand, our project seeks to extend the potential of LLMs to the field of psychiatry. Our goal is to develop a comprehensive large language model pipeline that facilitates psychiatric diagnosis when given clinically validated case vignettes. While we will assess the diagnostic accuracy of the LLMs, we also aim to uncover the influential factors that enable the model to mimic the common diagnostic patterns made by mental health and primary care professionals. To fulfill these objectives, we performed comparative analysis of various LLMs to select the most suitable model for development of our diagnostic pipeline. After determining the best performing model, we have designed the pipeline to systematically input data and gather diagnostic responses. In order to enhance the diagnostic accuracy, we have conducted experiments using different prompting strategies and attention weight analysis.

1.1 Related Work

The field of psychiatric diagnosis has been significantly shaped by a number of pivotal studies. Kogan et al [4] conducted an international field study evaluating the diagnostic guidelines for ICD-11 mood disorders. At the same time, Smith et al [5] was the first to use machine learning in psychiatric classification, showing the potential of automated systems to enhance diagnostic accuracy through the analysis of linguistic information and patterns learned from patient interviews, case studies and so forth. Furthermore, Jones et al [6] applied natural language processing (NLP) to extract insights from online mental health forums. Broadly, their work uncovered sentiments related to psychiatric symptoms. Complementary to this would be Wang et al [7], who used AI for clinical decision making in a psychiatric setting, showcasing the augmented efforts of human expertise and machine intelligence.

These foundational studies set the stage for the current project, which aims to advance psychiatric diagnosis by leveraging large language models (LLMs) to develop a comprehensive, yet accessible, psychiatric pipeline. Our aim is to bridge the gap in psychiatric evaluation accessibility for low-income areas. We aspire to achieve this by developing a mobile application designed to assist in the diagnostic process, offering a valuable resource where traditional access to psychiatric evaluation may be limited.

2 Problem Definition and Algorithm

2.1 Task

To evaluate the accuracy of our diverse LLMs on clinical case vignettes, we employ a multiclass classification machine learning framework. The number of classifications may vary depending on whether we are testing mood, stress, or anxiety. In each trial, the input to the LLM model is specified through a prompt template, encompassing a generic “welcome” prefix, guidelines, case vignette, and initializing prompt. Crucially, the guidelines, which include various diseases within the specific class under examination, serve as the list of choices from which the model makes its selection. Consequently, the output corresponds to one of the diseases from the list of potential choices. This procedure is systematically repeated for all clinical case vignettes across all disease classes.

2.2 Algorithm

Overview of LLMs LLMs are a revolutionary class of machine learning algorithms that have transformed the field of NLP. These models undergo pre-training on extensive textual datasets, equipped with attention mechanisms and deep architectures that endow them with remarkable contextual awareness. One prominent example of LLMs is the Transformer architecture, employed in widely-known models like ChatGPT. This architecture incorporates distinctive features, including self-attention mechanisms that enable the model to assess the significance of various words, parallelization for enhanced computational efficiency, and positional encoding to convey information about the position of each token in a sequence. This characteristic makes LLMs the models of choice for text generation, a pivotal aspect in our current project. Our methodology revolves around harnessing this inherent contextual awareness to effectively capture nuanced patterns embedded within the clinical case vignettes we analyze.

Implementation of transformed-based LLMs

1. Tokenization: Input texting tokenized into smaller units that each correspond to an embedding in a higher-dimensional space
2. Positional encoding: Added to token embeddings to convey information about the positional sequence of each token
3. Input embedding: Token embeddings serve as initial input representation

4. Encoder stack: Input passed through multi-head self-attention mechanism and a fully-connected feedforward neural network (e.g., multilayer perceptron) to introduce non-linearity
5. Layer normalization and residual connections: All layers normalized with an added residual connection
6. Encoder output: Output represents input sequence with incorporated contextual information and corresponding dependencies between tokens
7. Decode stack: Processes output of encoder and generates auto-regressed output sequence
8. Softmax output layer: Produces probabilities for next token in sequence
9. Sampling or greedy decoding: Next token is chosen based on the predicted probabilities
10. Output and iteration: Selected token is part of input for next iteration; this process is repeated until a certain criterion is met

3 Experimental Evaluation

3.1 Data

Data utilized in our project consists of three distinct datasets, each focusing on different mental health disorders, specifically Anxiety, Stress and Mood disorders. The data collection was conducted through Qualtrics surveys, a web based software that allows users to create surveys. Each survey was administered to approximately 900 clinicians, and it was made available in multiple languages including English, Spanish, Chinese, French, German, Russian and Japanese. These surveys encompassed four crucial components to ensure comprehensive assessment of psychiatric diagnosis. These components included guidelines of mental disorder diagnosis (based on ICD-11), clinical vignettes that simulate real-world scenarios, clinician responses to vignettes and ground truth data for validation.

The clinical vignettes employed in our study were sourced from the World Health Organization's (WHO) validation studies [8], thereby ensuring their reliability and relevance to real-world clinical scenarios. Vignettes were evaluated by 16,000 mental health and primary care professionals across 158 countries to ensure their reliability and relevance to real-world clinical scenarios.

3.2 Methods

Our research approach sought to solve the issue of excessive token sizes of our guidelines, in addition to the vignettes and prompts, surpassing the input constraints of typical LLMs. We selected specialized LLMs, specifically Llama2 7B, Long-Llama, Llama2-32k, and Mistral 7B, known for their ability to process considerably larger data sets. This choice was critical for accommodating the detailed nature of psychiatric vignettes. Llama2 7B could handle relatively small token sizes which is up to 4096. Long-llama and Llama2-32k can both handle up to 32k token sizes, while Mistral 7B can take input token sizes up to 8000.

In parallel, we condensed the ICD-11 guidelines using summarization tools from other LLMs (Llama or ChatGPT), ensuring the preservation of diagnostic information while effectively reducing the input size for the models. The incorporation of 4-bit quantization also played a pivotal role, providing a balance between model size and computational effectiveness. Lastly, the expertise of mental health professionals was integral in isolating key elements from the guidelines, ensuring our model's focus remained on the most relevant while reducing the input limit of LLMs.

Based on expert recommendations, we successfully extracted critical components from the ICD-11 guidelines, resulting in creation of three distinct guideline sets intended for model input. The first set included only the essential information. In the second set, we introduced additional background information along with the essential elements. In the third set, we combined essential features, additional background information, and a segment about boundaries with normality to help the model differentiate between disorders (Figure 1). Our purpose in developing these varied sets of guidelines was to assess the model's diagnostic accuracy with different levels of information input. This was particularly important due to the constraints of the limited token size we faced. Smaller input sizes offered us greater flexibility in model selection.

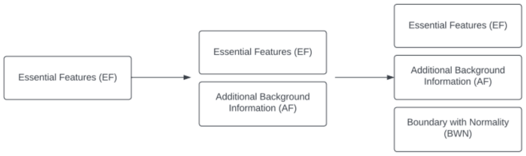
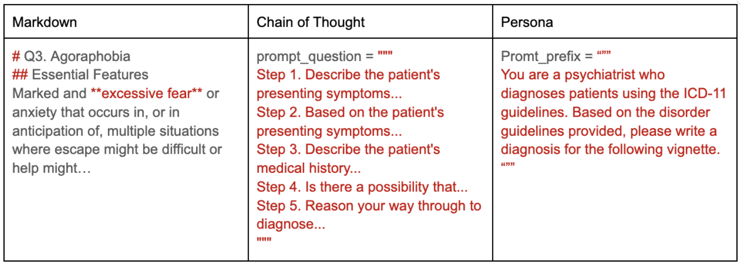

Essential Features (EF)
Essential Features (EF)
Additional Background Information (AF)
Essential Features (EF)
Additional Background Information (AF)
Boundary with Normality (BWN)

Figure 1: Levels of Guidelines



Markdown	Chain of Thought	Persona
# Q3. Agoraphobia ## Essential Features Marked and **excessive fear** or anxiety that occurs in, or in anticipation of, multiple situations where escape might be difficult or help might…	prompt_question = """ Step 1. Describe the patient's presenting symptoms... Step 2. Based on the patient's presenting symptoms... Step 3. Describe the patient's medical history... Step 4. Is there a possibility that... Step 5. Reason your way through to diagnose... """	Promt_prefix = “”” You are a psychiatrist who diagnoses patients using the ICD-11 guidelines. Based on the disorder guidelines provided, please write a diagnosis for the following vignette. “””

Figure 2: Examples of Prompting Strategies

3.2.1 Refining Prompting Techniques

Our project emphasized the importance of crafting effective prompts to draw accurate responses from LLMs. This process involved experimenting with diverse prompting strategies. Figure 1 depicts examples of prompting techniques used during experimentation:

- **Structured Markdown:** We employed a formatted approach to present case vignettes and guidelines, enhancing the model's interpretative capabilities. The idea is that by using markdown formatting, we can make LLMs to recognize the hierarchy within text, facilitating the recognition of text structure by the model.
- **Chain of Thought:** This strategy guided the model through a logical thought process similar to a psychiatrist's diagnostic approach, mimicking the cognitive steps in clinical reasoning. In many research findings, chain of thought.
- **Persona:** By embedding a professional persona within prompts, we aimed to align the model's responses with the expertise of psychiatric practitioners.
- **Guideline Order Randomization:** Guideline Order Randomization: To mitigate bias in the model's diagnostic process, we varied the presentation order of guidelines, ensuring a more unbiased approach in response generation. In Anxiety Disorder Guideline, the Generalized Anxiety Disorder (GAD) was the first disorder mentioned. Due to this, and also the fact that GAD is the most common, general diagnosis for an anxiety disorder, the model would consistently output GAD. Randomizing the guidelines improved the performance and not referring the diagnosis to GAD.

3.2.2 Output Evaluation Techniques

The assessment of the model's output involved the two followings:

Attention Weight Exploration: By examining the model's attention distribution across different parts of the prompt, we gained insights into its processing priorities.

Diagnostic Accuracy Comparison: The definitive measure of the model's effectiveness was its accuracy in psychiatric diagnoses. We compared its output against established ground truth labels, derived from professional assessments and ICD-11 standards, to quantitatively evaluate its diagnostic precision.

3.3 Results

Figure 3 shows the model performance based on prompt types that are Zero-Shot-Prompting for each vignettes types, and Figure 4 shows results based on Few-Shot Prompting. The results in

Prompt Type	**Vignettes**		
	Anxiety	**Stress**	**Mood**
Llama 2+EF	0.182 (0.455)	0.09 (0.545)	0.238 (0.476)
Llama 2+EF+AF	NA	0.2 (0.5)	0.2 (0.476)
Llama 2+EF+AF+BWN	NA	0.273 (0.636)	0.409 (0.545)
Mistral+EF	0.364 (0.727)	0.273 (0.455)	0.238 (0.429)
Mistral+EF+AF	0.273 (0.545)	0.273 (0.636)	0.19 (0.429)
Mistral+EF+A+BWN	0.182 (0.545)	0.455 (0.636)	0.238 (0.524)

Table 1: Model Performance with Zero-Shot Prompting

Prompt Type	**Vignettes**		
	Anxiety	**Stress**	**Mood**
Mistral+EF+AF+BWN+CoT	0.455 (0.636)	0.273 (0.636)	0.286 (0.619)
Mistral+EF+AF+BWN+CoT+Persona	0.545 (0.8)	0.545 (0.727)	0.333 (0.714)

Table 2: Model Performance with Few-Shot Prompting

parentheses are the syndrome accuracy, which clinicians assessed by examining whether the model's choice of disorder fell within the same treatment group as the ground truth disorder. Generally, the Mistral model outperformed the Llama2 model except for Mood. Typically, increasing the amount of information provided improved the performance of both Mistral and Llama2 models for Stress vignettes. However for anxiety, the model performed worse except when added additional prompting strategies such as chain of thought and persona. For syndrome accuracy, usually adding more input improves the model's diagnostic accuracy.

Interestingly, when the models made incorrect diagnosis, they often selected diagnoses that were emblematic of each disorder type. For example, for anxiety, it was GAD, stress was PTSD, and mood was single depressive disorder. Moreover, when the correct diagnosis was no diagnosis, the model still offered a diagnosis, or a false positive. There were no false negative results.

We also conducted attention weight analysis to assess whether the model exhibits higher attention to specific portions of the prompts. If it did, we aimed to adjust the placements of important information. Figure 5 shows the attention weights heat map for an example prompt based on the Mistral model. Each row represents individual tokens within the prompt while columns represent the parameters in the Mistral Model. The weights are based on averaging 33 hidden states. The visualization demonstrates that the Mistral model seems to "pay attention" to the texts in a uniform manner.

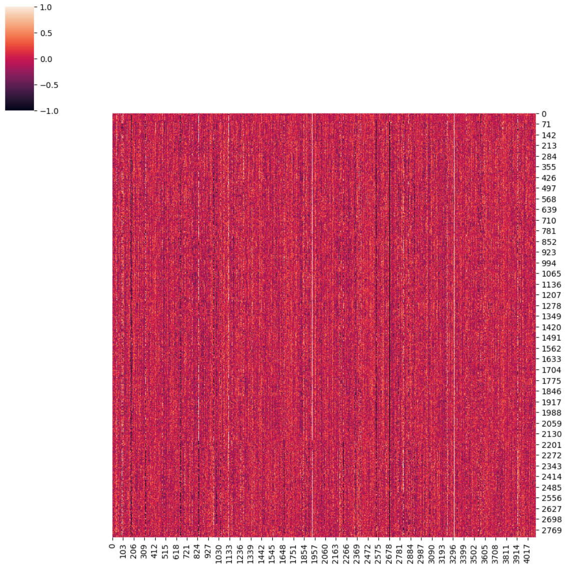


Figure 3: Attention Weights and Tokens Heat Map for Mistral

4 Discussion

As mentioned in the results section, when the models get the wrong diagnosis, they tend to pick diagnoses that are emblematic of each disorder type. This means the performance of our models follow the pattern of mistakes and success human clinicians make.

For different disorders types, different techniques had differing effects of making the model make the correct diagnosis. For example, there was not any noticeable improvement when adding more information for the mood disorder results. This may be due to the immense token size and the structure of the prompts. Perhaps different disorders require different prompt types for the models to be effective.

The performance of all prompt types were below 50 percent. However, the purpose of this analysis was not necessarily to prompt engineer the models to make the correct diagnosis, but rather is a step towards a more nuanced approach of leveraging AI for medical diagnosis.

5 Conclusion

This research sought to discover capabilities of LLMs in the field of psychiatric diagnosis. We focused on enhancing the accuracy of diagnosing Anxiety, Stress, and Mood disorders, leveraging a range of open-source LLMs and experimenting with different prompt engineering strategies. Our findings indicate that LLMs can indeed replicate the diagnostic patterns commonly observed in human clinicians. An intriguing aspect of our results was the high rate of false positives in scenarios requiring a "no diagnosis" conclusion, highlighting the need for further refinement in the diagnostic approach.

The study also revealed that simply adding more information to the model does not guarantee improvement of the diagnostic accuracy of the LLMs. This observation suggests that the effectiveness of LLMs in psychiatric diagnostics may not solely depend on the quantity of data but also on how this data is processed, suggesting that tailoring specific prompting strategies to different disorders may be the key to optimizing diagnostic accuracy. Moreover, our research also underlines some of the future work in several key areas. This includes translating and formatting multilingual data to facilitate broader linguistic comparisons, fine-tuning LLMs with additional mental health data, and a more thorough examination of the agreement between LLMs and clinicians to explore similarities and differences in their diagnostic approaches.

In conclusion, our project marks a significant step towards integrating LLM in the field of medical diagnosis. By closely mimicking the response patterns of clinicians, including their errors, our LLMs offer a promising foundation for future developments.

5.1 Student Contribution

All team members had an equal contribution in generating LLM pipelines, assessing model performance, and writing the report. In terms of project design, all members collaborated to build the LLM pipelines through regular meetings. When it came to programming and evaluating models, we each worked on different models and prompting strategies, successfully completing our respective tasks. For the report, each member was responsible for specific sections: Hannah handled the project objectives and data, Jin worked on the methods section, George focused on the results section, and Stephen took care of the discussion and conclusion.

References

[1] Gilson, Aidan, et al. “How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment.” JMIR Medical Education, vol. 9, Feb. 2023, p. e45312. DOI.org (Crossref), https://doi.org/10.2196/45312.

[2] Karabacak, Mert, and Konstantinos Margetis. “Embracing Large Language Models for Medical Applications: Opportunities and Challenges.” Cureus, May 2023. DOI.org (Crossref), https://doi.org/10.7759/cureus.39305.

[3] Van Heerden, Alastair C., et al. “Global Mental Health Services and the Impact of Artificial Intelligence–Powered Large Language Models.” JAMA Psychiatry, vol. 80, no. 7, July 2023, p. 662. DOI.org (Crossref), https://doi.org/10.1001/jamapsychiatry.2023.1253.

[4] Kogan, C. S., Stein, D. J., Maj, M., First, M. B., Emmelkamp, P. M. G., Reed, G. M., ... Evans, S. C. (2016). The International Classification of Diseases (ICD)-11 for Mood and Anxiety Disorders: The Work of the Mood Disorders Expert Advisory Group. World Psychiatry, 15(2), 117–118.

[5] Smith et al: Utilizing Machine Learning for Psychiatric Classification Smith, K., Hames, A., Hagan, T. (2018). Machine Learning in Mental Health: A Scoping Review of Methods and Applications. Psychological Medicine, 48(9), 1291–1308. https://doi.org/10.1017/S0033291717002408

[6] Jones et al: Natural Language Processing in Mental Health Forums Jones, R. B., Hoare, P., Elton, R. (2018). Natural Language Processing in Mental Health Applications Using Hyperdimensional Computing: A Meta-Analysis. JMIR Mental Health, 5(2), e11158. https://doi.org/10.2196/11158

[7] Wang et al: Integration of AI in Clinical Decision Support Systems Wang, Z., Shah, A. D. (2019). Understanding the Adoption of Clinical Decision Support Systems in Behavioral Health: Application of the Unified Theory of Acceptance and Use of Technology in Mental Health. JMIR Mental Health, 6(4), e14041. https://doi.org/10.2196/14041

[8] Evans, S. C., Roberts, M. C., Keeley, J. W., Blossom, J. B., Amaro, C. M., Garcia, A. M., Stough, C. O., Canter, K. S., Robles, R., Reed, G. M. (2015). Vignette methodologies for studying clinicians’ decision-making: Validity, utility, and application in ICD-11 field studies. International Journal of Clinical and Health Psychology, 15(2), 160–170. https://doi.org/10.1016/j.ijchp.2014.12.001