Evaluating the Efficacy of Utilizing Large Language Models in Mental Disorder Diagnoses

Eric Gong

General Education Course 1179: Psychotherapy

Introduction:

Large Language Models (LLMs) – a form of Generative Artificial Intelligence trained on vast corpuses of textual data to enable not only the understanding of natural language input, but also its creation as output – have become center-stage in the minds of the populace since the release of ChatGPT in 2022 [1]. Be it households, workplaces, or research labs studying the cutting edge, LLMs have shown the remarkable ability to engage in a wide range of topics and specialties, albeit with varying degrees of success [2]. An LLM's adaptability to various differing tasks can be accredited to the vast breadth of topics covered in the training data, its ability to mimic multi-step reasoning to break larger problems into smaller logical steps, and most importantly, its ability to mimic learning, particularly in-context learning, where the LLM can understand how to complete new tasks through instructions and examples provided by the input prompt [1]. These traits generate strong interest among scientists and researchers who wish to explore the potential for LLMs to augment, or even entirely replace, tasks currently performed by humans. Be it generating satisfactory code with far greater speeds than software engineers, or even outperforming the average clinician in neurology examinations, various studies have demonstrated that LLMs hold promise in exceeding human capabilities in specific tasks [3,4].

The field of psychotherapy is, surprisingly, no stranger to this new obsession for replacing natural intelligence with the artificial. As a matter of fact, the first inklings of integrating technology into psychotherapy dates back over half a century prior, to ELIZA, a computer program developed at MIT [5]. While the original paper only intended to explore the interactions between humans and machines – with a computer therapist being the most convenient role, given that only in a patient-therapist conversation is it possible for a conversation to take place while assuming no knowledge of the outside world – ELIZA portrayed a far more convincing illusion of cognition than its designer Weizenbaum had intended, soon bringing ELIZA great fame, despite it being no more than a simple grammar-parsing program capable of deterministically re-phrasing user input to a fixed output [6]. Similar technology has been commercialized in the form of Woebot, a chatbot service that draws from pre-written responses, with the promise of helping users "develop coping skills for symptoms of anxiety and depression" and "encourage healthy lifestyle choices," in addition to being available 24-7, and (almost) free of cost, unlike traditional therapists [7].

Given the scalability of technology in comparison to human therapists, as well as the reduced cost, there is great interest in evaluating whether the move from fixed deterministic outputs to generative outputs via LLMs may hold promise for the field of psychotherapy. Indeed, researchers have already demonstrated that LLMs hold some promise for specific tasks within therapeutic settings. For instance, when provided with clinical vignettes of patients with obsessive-compulsive disorder (OCD), LLMs were able to correctly identify OCD as the primary diagnosis in almost all vignettes, outperforming the performance of medical healthcare professionals [8]. On the other hand, unfortunately, LLMs fall short of human therapists at the task of identifying and challenging a patient's cognitive distortions in the context of CBT [9]. In particular, researchers observed that while LLMs were able to offer advice which "technically [improved] the [patient's] original thought" the LLMs appeared to lack an understanding of, and thus ability to address, the cognitive distortions underlying the patient's thoughts [9].

But in addition to having worse performance than therapists, there are numerous factors which make the endeavor of utilizing LLMs to completely replace human therapists severely problematic. For instance, LLMs have been seen to encourage suicidal ideation, making them a severe danger to patients seeking therapy [10]. More generally, however, LLM's have been known to exhibit biases and enforce stereotypes, mirroring, unfortunately, the sentiments of the data that the LLM is trained on [11]. The propagation of these biases and stereotypes into any therapy the LLM would result in a complete failure to provide – in the words of Iwamasa – culturally competent therapy that takes into account the cultural, political, and environmental background that is unique to each patient [12]. Indeed, given that culturally competent therapy remains a challenge, even among trained therapists, there is little hope that there would be data available by which to train or fine-tune an LLM to be able to generate output capable of mimicking the response of what a culturally competent therapist may say. Thus, unsurprisingly, the American Psychiatric Association advises physicians to "remain skeptical of AI output when used in clinical practice," if at all [13].

As such, the most foreseeable application of LLMs in the field of psychotherapy may lie in their ability to increase the efficiency of trained therapists, assisting with routine tasks, or offering suggestions that the therapist could consider, but easily reject should the suggestion prove to be biased, incorrect, or harmful. In doing so, LLMs would increase the bandwidth of therapists, allowing them to treat more patients, with greater efficiency, without patients suffering the harms that may result if an LLM were directly exposed to a patient for the purpose of therapy, with no trained human therapist present. In fact, LLMs are already being put to work, analyzing patient-therapist transcripts to identify potential factors which distinguish therapists who are more successful from those that are not, in hopes that this information can increase the efficiency of therapists, allowing them to see more patients in similar amounts of time [14]. In addition, similar to the study conducted on LLMs identifying patients with OCD from vignettes, it may be possible for LLMs to assist therapists in the process of diagnosing the mental disorders a patient is struggling with, as a first step for creating an informed treatment plan for the patient. Thus, in this study, we aim to examine the efficacy of LLMs in generating accurate diagnoses for various differing mental disorders, which, if shown to be effaceable, has the potential to reduce mental labor on the part of the therapist: instead of needing to actively diagnose mental disorders based on the patients' behaviors, the therapist would need only to confirm whether the LLM-reported disorder diagnoses are valid.

Methodology:

To ensure that the LLMs have a reliable framework upon which to judge whether a patient ought to be diagnosed with a particular disorder, and to ensure that the LLM's definitions of mental disorders is in agreeance with that of therapist, we shall make use of the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), which not only provides a list of mental disorders as defined by the American Psychiatric Association, but also specific list of diagnostic criteria for each defined mental disorder by which to judge whether or not a patient should be diagnosed with the respective disorder [15]. As acknowledged by the DSM-5, "it is not sufficient to simply check off the symptoms in the diagnostic criteria to make a mental disorder diagnosis. A thorough evaluation of these criteria may assure more reliable assessment… the relative severity and salience of an individual's signs and symptoms and their contribution to a diagnosis will ultimately require clinical judgment" [15]. As such, an LLM that is capable of issuing an accurate diagnosis as per the DSM-5's criteria will hold promise in supporting and supplementing the judgement of a therapist who would be responsible for considering the nuances of the patient's conditions, but notably, this capability would not demonstrate the LLM's ability to fully replace a therapist in the context of issue the diagnosis.

So as to test the efficacy of an LLM in this role, we will make use of the "Casebook for DSM-5," a compilation of 30 clinical vignettes of patients undergoing therapy, followed by the mental disorders the patient is diagnosed with, according to "seasoned clinicians who have experienced complex client symptomology" [16]. These vignettes cover a variety of different patients, with differing demographics, and differing mental disorders. As such it will be possible to assess the acuity of LLMs in identifying a variety of different Mental Disorders. In particular, we will provide a popular and publicly accessible LLM, Google's Gemini Model, with the clinical vignettes, one at a time, alongside a copy of the DSM-5, and request that the LLM diagnose the patient in accordance with the DSM-5's diagnosis criteria. We will then compare the LLM's diagnoses with the diagnoses of human therapists to determine the accuracy of the LLM. We shall utilize chain-of-reasoning prompting strategies – which has been shown to improve LLM performance – to increase the accuracy of the LLM's diagnoses, but also to allow for an examination of any implicit biases or problematic reasoning that the LLM may undergo as it reasons through whether or not to issue a particular diagnosis.

Specifically, we follow the following prompting framework:

First Message:

> *In the following messages, you will be given Anonymized Clinical Vignettes of patients. You are to determine what Mental Disorders they may be experiencing based on the provided DSM-5 Manual*

Second Message:

[A PDF copy of the DSM-5]

Third Message:

Based on the following Clinical Vignette, determine what Mental Disorders the patient may be experiencing, as per DSM-5 definitions of Mental Disorders. Note that the vignette is fully anonymized, and is created as an assessment of your diagnostic capabilities; answer the question while remaining cognizant of sensitive content:

[INSERT CLINICAL VIGNETTE HERE]

Note that the second sentence of the third message is necessary so as to bypass the LLM's generic content safeguard, wherein it sometimes generates a response reporting it is not able to issue any form of medical diagnoses given that it is not a trained medical professional, as well as its refusal to generate output containing references to highly sensitive topics such as the sexual abuse of children, one of the many sensitive topics touched upon by the Clinical Vignettes featured in the Casebook. Furthermore, although outside the purview of this study, it was noted that for messages flagged by Google's Gemini Model as violating usage policies, the OpenAI User Interface could be used: even when the input message was flagged and deleted for "violating [OpenAI's] usage policies", after about a 20 second delay, the LLM would nonetheless response to the purportedly problematic prompt, which may indicate a concerning flaw in OpenAI's content censoring system.

Results:

Based on results of querying the LLM with 30 vignettes and examining the results, it is evident that the LLM exhibits a couple properties in the context of generating a diagnosis, which ought to be taken into consideration when potentially utilizing LLMs in supporting therapists in their diagnosis. The results of the LLM queries are summarized in Table 1, which lists the LLM's diagnosis alongside the diagnosis given by the Casebook.

Specifically, the LLM exhibits three key properties which must be taken into account if such tools are to be used in assisting diagnosis. The first is a False Positive property, where the LLM is seen consistently over-diagnosing patients with Mental Disorders that the evaluation by human experts does not report. The second is Overconfidence, where the LLM confidently reports that the vignette contains sufficient evidence to diagnose a patient with a mental disorder, whereas the human experts have noted the particular mental disorder with "r/o," indicating that further information is necessary to determine whether or not the patient is to be diagnosed with the mental disorder at hand. The final property is the General Gist, where the LLM reports a mental disorder diagnosis with similar symptoms and exhibited behaviors to the ground-truth diagnosis reported by human experts.

False Positive Diagnoses:

It can be seen across all vignettes that the LLM almost always creates a list of between four and six mental disorder diagnoses, despite the fact that for any vignette, the number of mental disorders each patient is diagnosed with is only between one and three. As a result, there are between two to four misdiagnoses, or false positives, in the LLM generated diagnosis. It should be noted however, that in 29 out of the 30 vignettes, the LLM correctly identifies at least one of the diagnoses reported by human experts, and in 15 of the 16 vignettes where human experts diagnosed the patient with more than one mental disorder, the LLM was able to correctly identify at least two of the mental disorders the patient experienced. Finally, in 12 of the 30 vignettes, the LLM was able to generate a list of diagnoses where every mental disorders identified by human experts could be found in the LLM's predictions.

Overconfident Diagnoses:

Out of the 30 vignettes, seven of them contain "r/o" indications for mental disorders by human experts, indicating that the diagnoses of the mental disorder at hand would require further information about the patient to ascertain. In six of these seven vignettes, the LLM's response featured these uncertain diagnoses as a certainty. As such, it can be seen that in instances where the evidence is inconclusive, but leans towards the potential presence of a particular mental disorder, the LLM leaps to conclusions and diagnoses the patient with the mental disorder in question.

General Gist Diagnoses:

In seven of the LLM responses out of the 30 vignettes, the LLM listed mental disorder diagnoses that are similar to a mental disorder identified by human experts. In particular, similar mental disorders are defined in this case for if the LLM reports a more generalized umbrella of the expert-determined mental disorder, a more specific sub-category of the mental disorder, or a mental disorder with similar symptoms and resulting behaviors. Some examples of similar mental disorders include the LLM reporting that a patient should be diagnosed with Level 1 on the Autism Spectrum, when human experts instead have diagnosed the patient with Social (Pragmatic) Communication Disorder. As per the DSM-5, Social Communication Disorder "manifested by deficits in understanding and following social rules of both verbal and nonverbal communication in naturalistic contexts," thus subtly differing from Level 1 on the Autism Spectrum, which, in addition to being described as "Difficulty initiating social interactions, and clear examples of atypical or unsuccessful responses to social overtures of others" is distinguished from Social Communication Disorder in that patients "May appear to have decreased interest in social interactions." Similarly, there are instances when the LLM reports that a patient may be experiencing Major Depressive Disorder, when instead human experts identify the patient as having Persistent Depressive Disorder, or another similar but distinct Mental Disorder for example. In all cases, it appears that the LLM is unable to fully recognize, or extract from the vignette, the subtleties that differentiate the similar yet nonetheless differing Mental Disorders.

Conclusion:

It is clear that as a stand-alone tool, LLM's lacks the ability to provide a patient with an accurate diagnosis. The success seen with the research study that assessed an LLM's ability to diagnose OCD likely yielded successful results due primarily to the fact that LLMs tend to be overconfident and given many false positives. Notably, that study only gave the LLM vignettes of patients diagnosed with OCD, and completely lacked negative controls. As a result, it is unsurprising that the LLM would have appeared to have essentially 100% accuracy; the inaccuracy of the LLM lies in its inability to realize that a patient does not have a certain mental disorder, even if they exhibit some of the symptoms for the disorder, or similar symptoms to the disorder. Indeed, this particular inaccuracy was left completely unevaluated by the aforementioned study.

However, this finding does not necessarily make the use case of LLMs within psychotherapy a complete impossibility. In particular, given that the LLM overcompensates in the diagnosis, issuing a significant number of false positives and overconfident predictions, it may prove useful in aiding therapists in identifying Mental Disorders that may otherwise be accidentally overlooked. The bulk of the mental labor of analyzing the patient's behavior can be offset to the LLM, while the therapist can examine the list of mental disorders generated by the LLM, so as to determine if each one fully correct, or should be marked as needing to be ruled out through further evidence, or if there is a similar mental disorder that more appropriately describes the patient's circumstances. Thus, based on this study, LLM currently show promise in serving a role similar to an inexperienced but book-smart and eager assistant, whose suggestions – while not always correct, and should be taken with caution – may prove invaluable in the instances where a tired but experienced therapist may accidentally overlook a key piece of information regarding the patient's behavior, and needs a simple hint in the right direction.

Sources:

1. Minaee. S *et al*., (2024) Large Language Models: A Survey. *Arxiv.* https://arxiv.org/pdf/2402.06196
2. Raiaan M. A. K. *et al*., (2024) A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, vol. 12, pp. 26839-26874. https://ieeexplore.ieee.org/document/10433480
3. Haque. A, (2024) LLMs: A Game-Changer for Software Engineers?. *Arxiv.* https://arxiv.org/pdf/2411.00932
4. Schubert, M. C. Wick, W. Venkataramani, Varun. (2023) Performance of Large Language Models on a Neurology Board–Style Examination. *JAMA Network.* https://pmc.ncbi.nlm.nih.gov/articles/PMC10704278/
5. Weizenbaum, J. (1966) ELIZA—a computer program for the study of natural language communication between man and machine Download ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9, pp. 36–45 https://canvas.harvard.edu/courses/140019/pages/week-11
6. Shrager, J. (2024) ELIZA Reinterpreted: The world's first chatbot was not intended as a chatbot at all. *Arxiv.* https://arxiv.org/html/2406.17650v1
7. Woebot Health. https://woebothealth.com/
8. Kim, J *et al.,* (2024). Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ digital medicine*, *7*(1), 193. https://pmc.ncbi.nlm.nih.gov/articles/PMC11271579/

9. Hodson, N., & Williamson, S. (2024). Can Large Language Models Replace Therapists? Evaluating Performance at Simple Cognitive Behavioral Therapy Tasks. *JMIR AI*, *3*, e52500. https://pmc.ncbi.nlm.nih.gov/articles/PMC11322688/

10. Miller, G. Lennette, B. (2024) Breaking Down the Lawsuit Against Character.AI Over Teen's Suicide. *Tech Policy Press*. https://www.techpolicy.press/breaking-down-the-lawsuit-against-characterai-over-teens-suicide/

11. Miller, K. (2024) Models Are Reinforcing Outdated Stereotypes. *Stanford University Human-centered Artificial Intelligence.* https://hai.stanford.edu/news/covert-racism-ai-how-language-models-are-reinforcing-outdated-stereotypes

12. Iwamasa, G. Y. *et al.* (2019) Culturally responsive cognitive behavior therapy. *American Psychological Association.* https://canvas.harvard.edu/courses/140019/files/19700657?wrap=1

13. The Basics of Augmented Intelligence: Some Factors Psychiatrists Need to Know Now. *American Psychiatric Association.* (2023) https://www.psychiatry.org/News-room/APA-Blogs/The-Basics-of-Augmented-Intelligence

14. Jee, C. Heaven, W. D. (2021) The therapists using AI to make therapy better. *MIT Tech Review.* https://www.technologyreview.com/2021/12/06/1041345/ai-nlp-mental-health-better-therapists-psychology-cbt/

15. American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association

16. Ventura, E. (Ed.) (2017). *Casebook for DSM-5. Springer Publishing Company, LLC*

Supplemental Figures:

Table 1: LLM Predicted Diagnoses and Diagnoses given by the Casebook

Patient	Predicted Disorders	Actual Disorders and Noted Observations
Dylan	OCD, ARFID, ASD (Level 1), GAD, Provisional Tic Disorder	OCD, Tic Disorder, Social (Pragmatic) Communication Disorder
Carol	Alcohol Use Disorder (Severe), PTSD, MDD, BPD, Other Specified Trauma- and Stressor-Related Disorder	Alcohol Use Disorder, Severe, R/O PTSD, R/O Unspecified Depressive Disorder
Keith	RAD, PTSD, Persistent Depressive Disorder, Enuresis, Other Specified Trauma- and Stressor-Related Disorder	Reactive Attachment Disorder, Child Neglect, Child Psychological Abuse, R/O Child Sexual Abuse, Enuresis, Avoidant/Restrictive Food Intake Disorder
Carla	Somatic Symptom Disorder, GAD, Persistent Depressive Disorder, Avoidant Personality Disorder	Somatic Symptom Disorder, R/O Medical Conditions, R/O Anxiety, R/O Unspecified Depressive Disorder
Todd	GAD, Social Anxiety Disorder, Persistent Depressive Disorder, Adjustment Disorder with Mixed Anxiety and Depressed Mood	Generalized Anxiety Disorder, R/O Medical Condition, R/O Depression, R/O Other Anxiety Disorders
John	MDD (Recurrent, Moderate), BED, Medication-Induced Sexual Dysfunction, Adjustment Disorder with Mixed Anxiety and Depressed Mood	Adjustment Disorder With Depressed Mood, Citalopram-induced Sexual Dysfunction, BED, Overweight
Michael	Gender Dysphoria, MDD (Recurrent, Moderate to Severe), Adjustment Disorder with Depressed Mood, Possible PTSD	Gender Dysphoria, Persistent Depressive Disorder (Early Onset, Intermittent Major Depressive Episodes, Severe), Parent-Child Relational Problem
Jamie	Anorexia Nervosa (Restricting Type), ADHD (Predominantly Inattentive), ASD (Level 1), Developmental Coordination Disorder, Possible ARFID	ARFID, ADHD (Predominantly Hyperactive/Impulsive, Partial Remission), Asthma, Dyspraxia, Academic Problem, Sibling Relational Problem
Maria	Selective Mutism, Adjustment Disorder with Mixed Anxiety and Depressed Mood, Possible Social Anxiety Disorder	Selective Mutism, Parent-Child Relational Problem
Jessica	PTSD, NSSI, MDD (Recurrent, Moderate to Severe), Possible Adjustment Disorder with Mixed Anxiety and Depressed Mood	PTSD (Delayed Expression), Child Sexual Abuse, Self-Harm
Rhonda	PTSD, BPD, Substance Use Disorder, Possible Dissociative Symptoms, Possible Adjustment Disorder	PTSD, BPD, Child Sexual Abuse, Sibling Relational Problem, Legal Problems, Self-Harm, R/O Substance Abuse
Jeremy	Schizoaffective Disorder (Bipolar), OCD, Paranoid Personality Disorder, ADHD (Combined), MDD, Possible IED, Possible Identity Confusion/Dissociation, Possible Substance Use	Schizoaffective Disorder (Depressive), Limited Social Support, Frequent Unemployment
Dan	Delusional Disorder (Erotomanic), OCD, Possible Interpersonal/Marital Issues, Possible Avoidant Personality Traits, Possible Substance Use/Escape Mechanisms, Possible Mood Disorder	Delusional Disorder (Erotomanic), Unspecified Personality Disorder, Employment Problems
Tim	Delusional Disorder (Persecutory), Possible PPD, MDD, Possible Substance Use, Possible Adjustment Disorder with Mixed Anxiety and Depressed Mood	Delusional Disorder (Persecutory), Family Disruption, Employment Problems

Mike	Alcohol Use Disorder (Severe), Alcohol-Induced Major Neurocognitive Disorder, Possible Encephalopathy, Possible MDD (Secondary), Possible Social Anxiety, Possible Schizophrenia/Schizoaffective	Alcohol-Induced Major Neurocognitive Disorder (Amnestic-Confabulatory, Persistent), Alcohol Use Disorder (Severe)
George	NPD, Sexual Sadism Disorder, Possible ASPD, Possible Substance Use (Alcohol), Possible Co-occurring Mood Disorder	Sexual Sadism, NPD, Employment Problems, Family Problems, Social Problems, Legal Problems, Self-Harm, R/O Substance Abuse
Jonathan	GAD, Social Anxiety Disorder, Possible OCD, Possible MDD, Possible Adjustment Disorder with Anxiety, Possible Perfectionistic Traits	Generalized Anxiety Disorder
Luz	Female Sexual Interest/Arousal Disorder, Possible PTSD, Possible Adjustment Disorder with Mixed Anxiety and Depressed Mood, Possible GAD, Relational/Family Issues	Female Sexual Interest/Arousal Disorder (Lifelong, Generalized, Moderate)
Nathan	Possible Gender Dysphoria, OCD, Possible Social Anxiety, Possible MDD, Possible Adjustment Disorder with Anxiety	OCD
Bryant	Voyeuristic Disorder, Paraphilic Disorder (Voyeuristic), Possible Sexual Dysfunction/Impairment, Possible Lack of Empathy/Understanding of Harm	Voyeuristic Disorder, Potential Problems with University/Legal System, Tension in Living Situation
Adrienne	Excoriation Disorder, Possible PTSD, Possible GAD, Possible BDD, Possible Adjustment Disorder with Mixed Anxiety and Depressed Mood	Excoriation Disorder, R/O GAD, Lack of Coping Skills, Peer Relationship Problems, Estrangement from Father
Jacob	ASD, ODD, Possible ADHD (Inattentive), Possible Social (Pragmatic) Communication Disorder, Possible Anxiety (Generalized/Social)	ASD (Requiring Support), No Intellectual Impairment, No Language Impairment
Jason	MDD, Sexual Dysfunction (ED - Psychological), Possible GAD, Possible Adjustment Disorder with Depressed Mood, Relationship/Communication Issues	Erectile Dysfunction, MDD (Recurrent, Moderate)
Bashir	PTSD, Conduct Disorder, Substance Use Disorder (Alcohol/Cannabis), Adjustment Disorder with Mixed Anxiety and Depressed Mood, Possible Antisocial Traits	MDD (Moderate), Nonsuicidal Self-Injury, Lack of Coping Skills, Legal Problems, Social Support Problems

Link 1: Hyperlink to the Raw LLM Response for Each of the 30 Clinical Vignettes

https://docs.google.com/document/d/19VWt8d6M2_ucXfSAupYhphXjephVKm9U/edit