LLMs for Precision Psychiatry: A Preliminary Evaluation of Large Language Models for Automated Extraction of OPCRIT Symptom Rating

Sutanay Choudhury[1](PhD), PhD, Erika Carter[2] (MD), Alejandro Michel Zuniga[1], Colby Ham[1], Ben McMahon[3,], Sayera Dhaubhadel[3,], Silviu Bacanu[4], Hamed Abbaszadegan[2](MD), Scott Grogan[2] (MD), Ayman Fanous[2] (MD), Khushbu Agarwal[1](MS)
[1]Pacific Northwest National Laboratory, Richland, WA; [2]The University of Arizona College of Medicine, Phoenix, Arizona; [3]Los Alamos National Laboratory, Los Alamos, NM; [4]Virginia Commonwealth University, Richmond, Virginia

Introduction The Operational Criteria Checklist for Psychotic Illness (OPCRIT) provides a structured approach to rating the presence of signs and symptoms of psychosis, mania, and depression from clinical notes [1]. Automated extraction of such clinical features and their integration with genomics profiles or electronic healthcare records can accelerate personalized diagnosis and treatment selection [2]. However, the reliance on extensive manual review of clinical notes has presented a challenge for automated extraction of psychiatric clinical features in practice. The emergence of large language models (LLMs) and their strong performance on computational tasks involving natural language understanding and inference [3] provides a new motivation to pursue automated extraction of OPCRIT symptoms from clinical notes. In this presentation, we demonstrate the use of LLMs to extract OPCRIT features in a zero-shot inference setting. Both GPT-3.5 and GPT-4 have been evaluated on AP psychology questions datasets. However, extracting OPCRIT symptoms requires stronger reading comprehension and natural language inference performance over input texts that are larger by orders of magnitude (as compared to AP psychology questions). Therefore, our study is a more practical evaluation of any LLM's readiness for deployment in a clinical setting.

OPCRIT Criteria	Question
Bizarre behavior	Behavior that is strange or bizarre and incomprehensible to others. Patient seems to be responding to auditory hallucinations or thought interference.
Excessive activity	Patient is markedly over-active with hyper-activity lasting one or two weeks. This includes motor, social, and sexual activity.
Restricted affect	Patient's emotional responses are restricted in range and at the interview there is an impression of bland indifference and lack of contact.
Irritative mood	Patient's mood is predominantly irritable and lasts for multiple weeks.
Dysphoria	Patient has a persistently low or depressed mood, irritable and sad mood or pervasive loss of interest.
Loss of pleasure	Patient pervasive inability to enjoy any activity. Patient's inability to enjoy any activity include marked loss of interest or loss of libido.
Persecutory delusions	Patient has persecutory delusions with persecutory ideation.
Grandiose delusions	Patient has grossly exaggerated sense of own importance, has exceptional abilities or believes that he is rich or famous, titled or related to Royalty. Also included are delusions of identification with God, angels, the Messiah etc.
Non-affective hallucination	Patient has hallucinations in which the content has no apparent relationship to elation or depression.

Table 1. Illustration of nine OPCRIT questions selected for current study. Note that accurate extraction of these features may require the inference of complex psychiatric concepts, and subsequent logical composition of the criteria.

Methods We present a dataset of 24 clinical notes extracted from publicly available DSM5 Cases, corresponding to the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders [4]. The notes are drawn from four categories of patients with 6 notes/category: 1) anxiety 2) bipolar 3) depression and 4) schizophrenia. We selected 9 representative OPCRIT questions (Table 1) and had a subject matter expert label the notes for presence of respective symptoms. Next, we implemented the extraction of selected OPCRIT questions via three state-of-the-art models (GPT 3.5, GPT-4 and Llama2) in a zero-shot inference setting. The models were instructed to return a boolean True/False response for each OPCRIT feature given the clinical notes and provide explanatory sentences for true answers. We experimented with different question decomposition strategies for prompt design and computed the precision as the fraction of questions where the LLM-answer matches psychiatric expert.

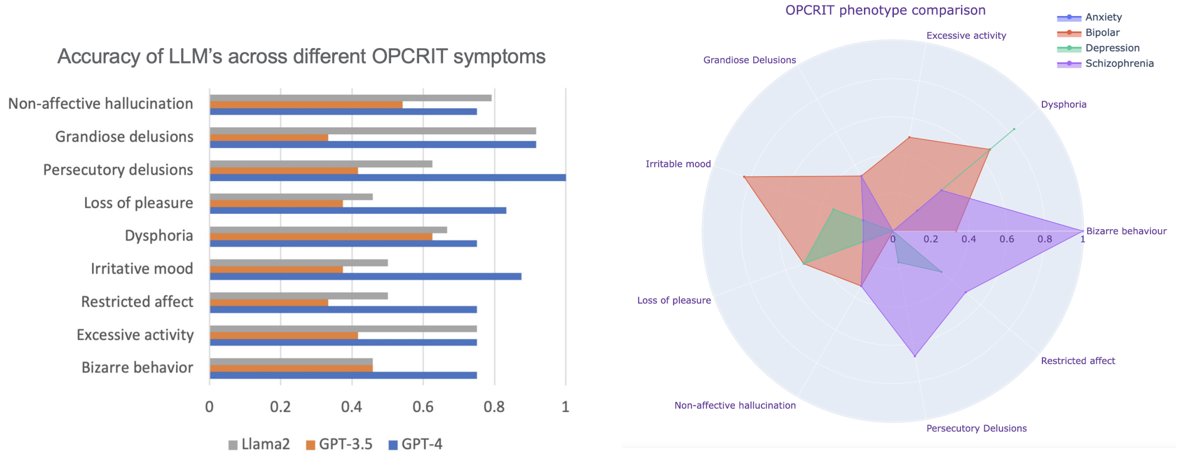

Figure 1. (left) GPT-4 reports performance with highest precision. LLama2 was comparable or outperformed GPT-4 in three tasks, and demonstrated superior performance over GPT-3.5 in 8 out of 9 tasks. (right) The figure on the right provides a performance breakdown of GPT-4 in terms of the different disorders and closely mirrors the rating patterns of the human expert.

Results Figure 1 reports the performance for each LLM compared to ground truth ratings provided by human expert. For Llama2, we experimented with all possible model variants and found best performance with the model with 70-billion parameters, smaller size models (Llama-13b chat, and GPT-2) showed inconsistent and ill-formatted results and are not presented here. Figure 2 presents the qualitative analysis of explanations generated by LLMs.

- {'case type': 'Schizophrenia', 'Bizarre behavior match': True, 'Bizarre behavior explanation': ['Although Mr. Baker had much academic success as a teenager, his behavior had become increasingly odd during the past year.', 'His sister said that she had recurrently seen him mumbling quietly to himself …..}
- {'case type': 'Bipolar', 'Excessive activity match': True, 'Excessive activity explanation': "The patient's behavior as described in the clinical notes matches the criteria. He was markedly over-active or hyperactive as evidenced by his refusal to sit and running through the ER, his leg bouncing rapidly up and down, and his incessant walking that led to blisters on his feet….}
- {'case type': 'Bipolar', Grandiose Delusions match': True, 'Grandiose Delusions explanation': ['He referred to himself as the "New Jesus" and declined to offer another name.', 'Despite being restrained, the patient remained giddily agitated, talking about receiving messages from God.', 'When asked when he had last slept, he said he no longer needed sleep, that he had "been touched by Heaven."}

Figure 2. Qualitative Analysis: Explanatory sentences generated by GPT-4 for different patients and OPCRIT features.

Discussion: Extraction of OPCRIT symptoms is a complex natural language understanding and inference task that challenges leading LLMs. Psychiatric symptoms are often described using ambiguous, complex, and varied language. For example, "auditory hallucinations" may be noted as "hearing voices" or implied from delusions. Our work demonstrates strong performance of GPT-4 both in terms of precision as well as explanatory capabilities. Clinical notes in a deployment setting are however likely to be noisier and could approach the maximum context size supported by these LLMs. Hence, further fine-tuning of such models will be required to attain similar results in a real deployment setting. Our work demonstrates the potential of LLMs in performing accurate psychiatric phenotyping at scale. This stands to unlock new possibilities for detecting subthreshold features of illness, subtyping illness presentations and identifying patients who differentially respond to treatment, which can be used in modes of prevention.

References:

1. McGuffin, P., et al, 1991. A polydiagnostic application of operational criteria in studies of psychotic illness: development and reliability of the OPCRIT system. Archives of general psychiatry, 48(8), pp.764-770
2. Kimbrel, N.A.et al., 2022. A genome-wide association study of suicide attempts in the million veterans program identifies evidence of pan-ancestry and ancestry-specific risk loci. Molecular psychiatry, 27(4)
3. Touvron, H. et al, 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288
4. Guha, M., 2014. Diagnostic and statistical manual of mental disorders: DSM-5. Reference Review