Evaluating LLMs to Measure Psychiatric Functioning

NYU Center for Data Science

Group Members: Hannah Park, George Zhou, Jin Choi, Stephen Spivack
Mentors: Matteo Malgaroli, Kyunghyun Cho, Geoffrey M. Reed

RESEARCH DESCRIPTION

Project Goals

- Assess the accuracy of Large Language Models (LLMs) to infer ICD-11 psychiatric diagnoses from clinically validated case vignettes.
- Assess the diagnostic accuracy of LLMs in evaluating psychiatric diagnosis across multiple languages
- Study agreement between LLMs responses and diagnostic assessments made by a diverse group of mental health professionals

Research Approach

- Design LLMs pipeline to systematically feed data and collect response.
- Compare model performance of different LLMs.
- Improve diagnostic accuracy by experimenting different prompting strategies and through attention weight analysis.

Related Work:

- Clinical Vignettes validation studies performed by World Health Organization (WHO). It was evaluated by 16,000 mental health and primary care professionals across 158 countries

DATASET AND METRICS

Data

- Data from three studies funded by the World Health Organization on Anxiety, Stress, and Mood disorders.
- Data collection through Qualtrics surveys encompassed four essential components:
  - Guidelines of Mental Disorder Diagnosis (ICD-11)
  - Clinical vignettes
  - Clinicians' responses to vignettes
  - Ground truth for validation
- Surveys were provided in multiple languages: English, Spanish, Chinese French, German, Russian, and Japanese.

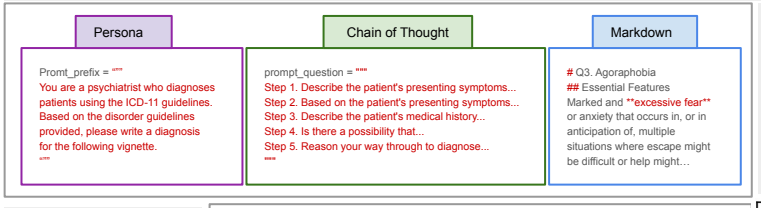
Evaluation Metrics

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

METHODS

Model Capacity

The main challenge of the project was the large token size which exceeded the LLM input limit. Several methods were implemented to tackle this issue, which were to (1) use LLMs that take much larger inputs, such as Long-Llama, Llama2-32k, or Mistral, (2) summarize the ICD-11 guidelines, (3) incorporate 4-bit quantization, and (4) work with experts and collaborators to find which element are most important in the guideline.



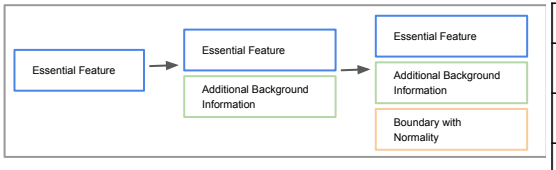
Prompt Engineering

[1] Identifying the best prompting strategy involved several steps:

1. Markdown
2. Chain of Thought
3. Persona
4. Randomizing guideline orders

Levels of Guidelines

To conserve token limits, progressively evaluate the performance of guidelines based on their level of importance in the element.



	Elements	Token Size
Anxiety	EF	1724
	EF + AF	3492
	EF + AF + BWN	4164
Stress	EF	1948
	EF + AF	2423
	EF + AF + BWN	2803
Mood	EF	742
	EF + AF	859
	EF + AF + BWN	1483

RESULTS

Generally, the Mistral model performs better than the Llama2 model except for Mood. Typically feeding in more information improves the model results for Mistral and Llama2 for Stress vignettes. However, for anxiety the models performs worse except when adding chain of thought and persona.

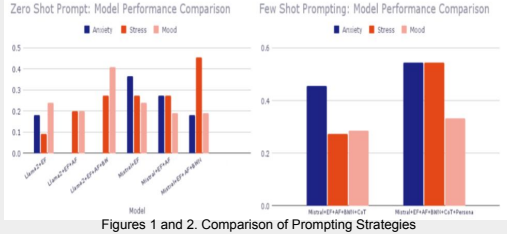
Interestingly, when the models get the wrong diagnosis, they tend to pick diagnosis that are **emblematic** of each disorder type. For anxiety it is GAD, stress is PTSD, and mood is single depressive disorder. When the correct diagnosis is no diagnosis, the model will usually give a false positive. There **were no false negative results**.

Prompt Type	Vignettes		
	Anxiety	Stress	Mood
Llama 2+EF	0.182	0.09	0.238
Llama 2+EF+AF	NA	0.182	0.2
Llama 2+EF+AF+BWN	NA	0.273	0.409
Mistral+EF	0.364	0.272	0.19
Mistral+EF+AF	0.273	0.273	0.19
Mistral+EF+A+BWN	0.182	0.455	0.19

Table 1: Model Performance with Zero-Shot Prompting

Prompt Type	Vignettes		
	Anxiety	Stress	Mood
Mistral+EF+A+BWN+CoT	0.455	0.273	0.19
Mistral+EF+A+BWN+CoT+Persona	0.545	0.545	0.333

Table 2: Model Performance with Few-Shot Prompting



Figures 1 and 2. Comparison of Prompting Strategies

CONCLUSION

The project aimed to identify the most effective ways to improve clinical-decision making rather than simply achieving correct diagnoses. Our models' behavior closely mirrored the typical clinician responses including misdiagnoses typically seen in clinical practice. Additional information didn't necessarily improve diagnostics, indicating that tailored prompting strategies for different disorders might optimize accuracy. This research is a step towards a more nuanced approach of leveraging AI for medical diagnosis.

FUTURE PLANS

(1) **Translate** the data currently in different languages to English and format it according to our existing prompt formats for the purpose of linguistic comparisons. (2) Examine the **inter-rater agreement between LLMs and clinicians survey responses** to explore similarities in mistakes and response patterns. (3) Lastly, **study potential bias** that can affect diagnostic abilities of the model such as varied socio-demographic characteristics amongst patients.

References: [1] Perez, Ethan, et al. "True Few-Shot Learning with Language Models: 35th Conference on Neural Information Processing Systems, NeurIPS 2021." *Advances in Neural Information Processing Systems 34 - 35th Conference on Neural Information Processing Systems, NeurIPS 2021*, edited by Marc'Aurelio Ranzato et al., 2021, pp. 11054–70. *NYU Scholars*, http://www.scopus.com/inward/record.url?scp=85131828022&partnerID=8YFLogxK