Natural Language Explanations for Suicide Risk Classification Using Large Language Models

William Stern, Seng Jhing Goh, Nasheen Nur, Patrick J Aragon, Thomas Mercer, Siddhartha Bhattacharyya, Chiradeep Sen, and Van Minh Nguyen
Contact: sternwill970@gmail.com, nurn@fit.edu

FLORIDA TECH

INTRODUCTION

Suicide and mental illness pose significant global challenges, compounded by barriers to seeking help and the overwhelming volume of online support requests. Automated methods can assess risk but lack interpretability. Natural language generation shows promise in explaining reasoning for suicide risk in natural language. This study evaluates their effectiveness compared to expert-written responses, aiming to improve clinical support.

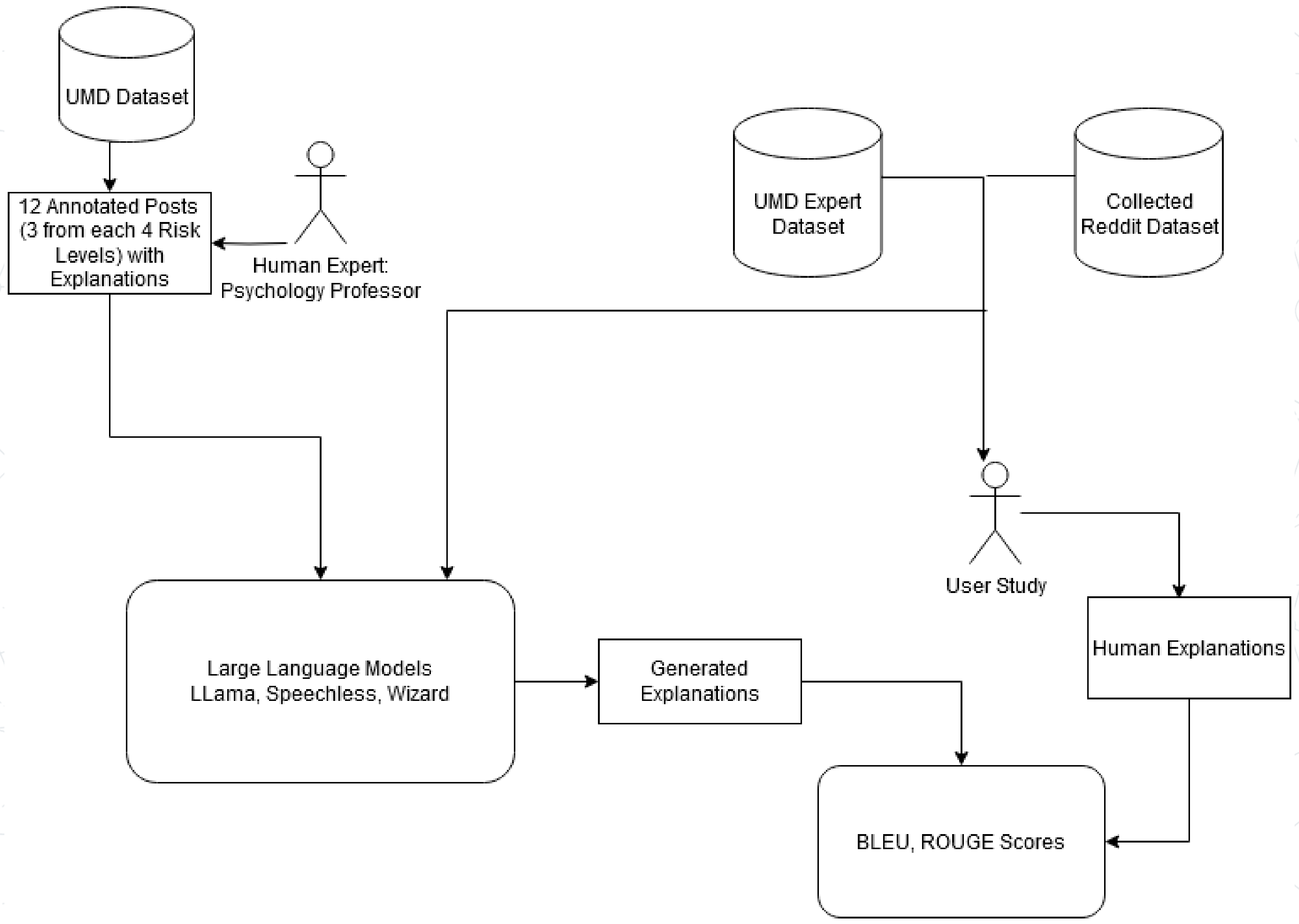
OBJECTIVES

1. I have been feeling very sad lately. I have had some thoughts of suicide. I feel very alone. ☹
2. I have been feeling not great, but I worked with my therapist and feel much better. ☺
3. Nothing matters anymore. I am going to kill myself when I get off work. ☹

Can we also utilize Large Language models (LLM) to explain why each post corresponds to a risk level?

LLaMA Chat 70B: This post represents **medium risk**. The poster mentions a history of **recent suicidal thinking**, feelings of **loneliness/social isolation, decreased interest** in activities they used to enjoy (school), as well as **hopelessness** about their academic performance, which could increase the future lethality of any potential attempt at self-harm.

METHODOLOGY


UMD Dataset
12 Annotated Posts (3 from each 4 Risk Levels) with Explanations
Human Expert: Psychology Professor
UMD Expert Dataset
Collected Reddit Dataset
User Study
Human Explanations
Large Language Models LLama, Speechless, Wizard
Generated Explanations
BLEU, ROUGE Scores

Instruction: Tell me the risk of suicide in the following social media post and explain the reasoning step by step with quotes. The risk can be no, low, medium, or high. Take the whole post into context.
Input: I am feeling very sad lately…
Response: This post is (no/low/medium/high) risk because…

RESULTS

	LLaMA 2 Chat 13B		WizardLM 13B		SpeechlessLM 13B		LLaMA 2 Chat 70B	
Avg. F1 Scores	BLEU	ROUGE-1	BLEU	ROUGE-1	BLEU	ROUGE-1	BLEU	ROUGE-1
No risk	0.194	0.211	0.185	0.234	0.204	0.198	**0.297**	**0.271**
Low risk	0.208	0.216	**0.248**	**0.236**	0.227	0.226	0.198	0.197
Medium risk	0.221	**0.250**	**0.254**	0.225	0.219	0.229	0.232	0.228
High risk	0.170	0.219	0.269	0.250	0.230	0.233	**0.280**	**0.271**
Entire dataset	0.230	0.442	**0.276**	0.482	0.272	0.472	0.265	**0.495**

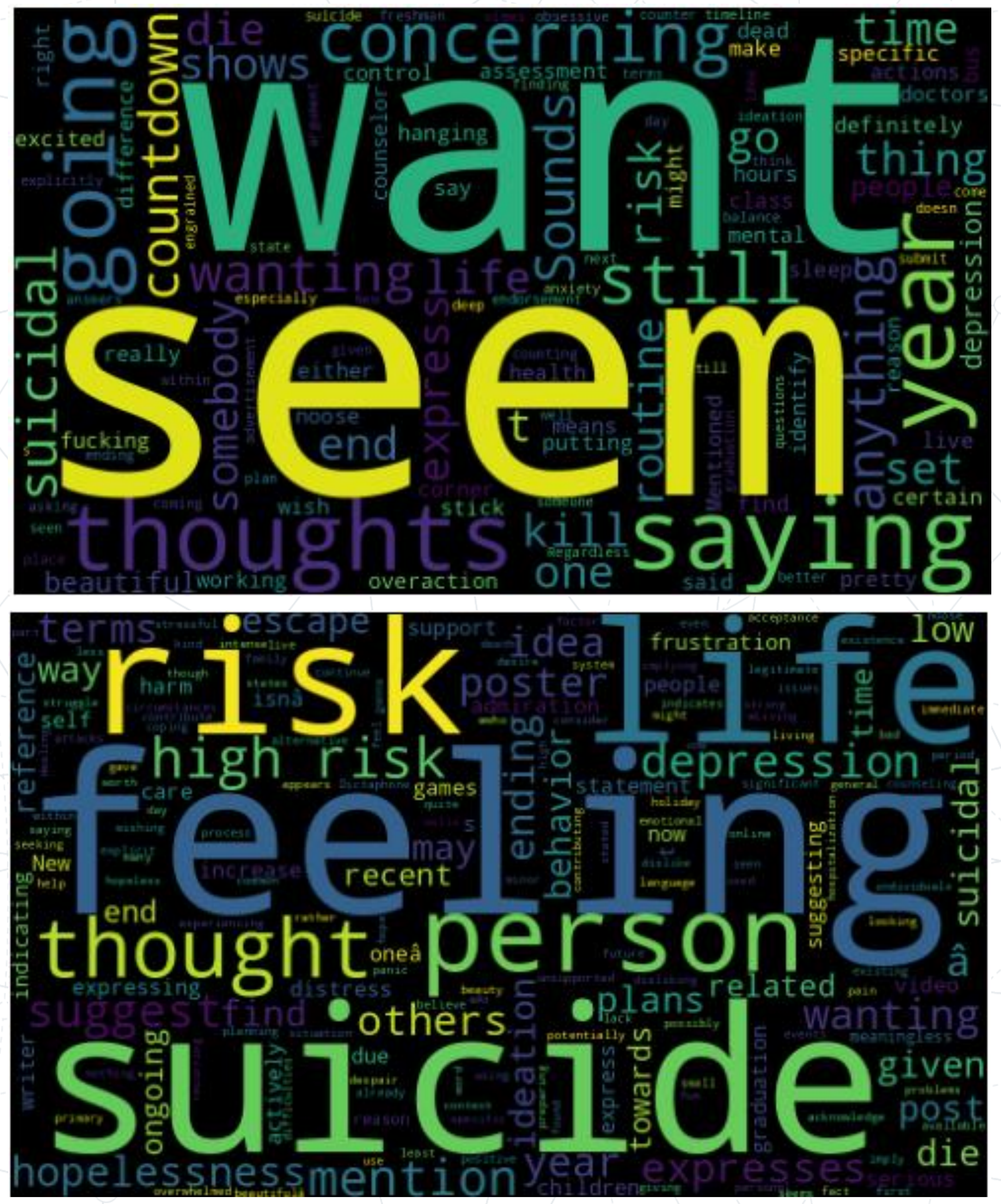
Generation evaluation metrics for UMD dataset

Avg. F1 Scores	BLEU	ROUGE-1
LLaMA 2 Chat 13B	0.222	0.391
WizardLM 13B	0.283	0.514
SpeechlessLM 13B	0.266	0.464
LLaMA 2 Chat 70B	**0.286**	**0.525**

Generation evaluation metrics for collected Reddit dataset

Model	F1	Precision	Recall
MentalBERT Base	0.39	0.42	**0.43**
WizardLM 13B	**0.41**	**0.43**	0.42
Speechless 13B	**0.41**	0.40	**0.43**
LLaMA Chat 13B	0.34	0.42	0.36
LLaMA Chat 70B	0.32	0.36	0.33

Classification accuracy of LLMs on the UMD dataset



Word-clouds generated on explanations originally annotated as medium-risk posts from UMD dataset: Human participants (top) vs. model-generated from WizardLM (bottom)

LLMs can provide suicide risk predictions and explanations