

Natural Language Explanations for Suicide Risk Classification Using Large Language Models

William Stern¹, Seng Jhing Goh¹, Nasheen Nur^{1*}, Patrick J Aragon², Thomas Mercer¹, Siddhartha Bhattacharyya¹, Chiradeep Sen³ and Van Minh Nguyen⁴

¹Electrical Engineering and Computer Science, Florida Institute of Technology, Melbourne, FL 32901 USA

²School of Psychology, Florida Institute of Technology, Melbourne, FL 32901 USA

³Mechanical and Civil Engineering, Florida Institute of Technology, Melbourne, FL 32901 USA

⁴Mathematical Sciences, Florida Institute of Technology, Melbourne, FL 32901 USA

Abstract

In recent years, automated machine learning has assisted mental health researchers in analyzing large amounts of data. Despite this, many individuals and organizations hesitate to use these methods, as they are black-box methods. Our work uses off-the-shelf generative large language models (LLM) to generate natural language explanations for suicide risk from users' Reddit posts. We benchmark various language models utilizing annotations and explanations by psychology experts. Generated explanations can accurately explain the risk of suicide using evidence from users' posts. We optimized this method for low-resource settings by leveraging pre-existing general instruction-tuned and quantized models. We conduct user studies with experts to compare the explanations and predictions generated by our approach with human expert perception and report some interesting findings for further improving our approach. We find that LLMs can effectively classify and respond with helpful reasoning of a suicidal risk diagnosis.

Keywords

mental health, large language models, social media, explainable AI

1. Introduction

Suicide and mental illness are highly prevalent throughout the world today. Unfortunately, many people do not seek the assistance of a mental health professional or are unable to do so. Often, this is due to the cost and availability of mental health treatment or the stigma surrounding it. People turn to online communities for various reasons, including seeking advice or venting to peers. At times, these social media posts can show warning signs of mental illness. Moreover, individuals receiving mental health treatment may explicitly post about their conditions on social media, sharing additional information. These posts could provide valuable information to mental health professionals, 24/7 emergency responders, friends, and family members of the individual. Ideally, if an individual creates a post suggesting they are experiencing symptoms of a mental health condition, these posts can be used to signal a need for help and ultimately get the individual closer to avenues for treatment.

One of the harshest consequences of mental illness is suicide. There is a high prevalence of people struggling with suicidal thoughts who seek help online. One of the

largest communities for this is the r/SuicideWatch subreddit on Reddit's social media website. This group has more than 441 thousand members and dozens of new posts every day¹. The sheer volume of new posts daily makes it impractical for moderators or mental health professionals to review all of them and offer advice or resources. Automated methods are often employed to sift through these posts and identify users who may be at risk [1]. Many of these automatic methods are black-box, such as models like neural networks. It is difficult to determine how the model makes its decisions. However, some of these models may provide explainability features that can be utilized to interpret their decision. For example, a lexicographical model can provide word importance [2] [3]. However, these explanations can often be complicated to interpret for people without extensive experience in data-related science.

One method that can help model interpretability is natural language generation (NLG). NLG has been utilized before for image classification to help explain the reasoning behind a specific classification using a deep classifier and a recurrent explanation generator [4]. Interpretable text classification for mental health has been recently developed with NLG models such as MentalLLaMA [5]. However, we aimed to test different freely available generative LLMs from the LLaMA family of models [6] on detecting the risk of suicide. The LLaMA family was released online at the start of 2023 and has

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada

*Corresponding author.

✉ sternwill970@gmail.com (W. Stern); nurn@fit.edu (N. Nur)

🌐 <https://www.fit.edu/faculty-profiles/n/nur-nasheen/> (N. Nur)

🆔 0009-0005-8879-2448 (W. Stern)

© 2024 Copyright © 2024 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.reddit.com/r/SuicideWatch/>

gained many users and applications. The LLaMA family is more lightweight than many other generative LLMs. Consumer desktop computers and laptops can run many LLaMA models locally². This is because they utilized smaller models trained on more data. As a result, these models can achieve similar performances to much larger models. Using these models becomes more convenient as a result. Additionally, many generative LLMs are not open-sourced and often collect user-inputted data. This is in contrast to LLaMA 2 [6], which was a permissive license, and the models can be used locally, allowing mental health professionals to control their own data use. For many mental health professionals, sharing their client's data with external companies would violate confidentiality. We plan to evaluate the quality of the generated responses from LLaMA models by comparing them to those written by mental health professionals. To do that, we used a variety of automatic metrics and qualitative analysis of the responses generated.

Here is a list of four key contributions:

- Generating natural language explanations for suicide risk using a generative LLM utilizing annotations directly from mental health professionals.
- Measuring the performance of different models for natural language explanations and suicide risk classifications.
- Measuring the accuracy of generated suicide explanations by comparing them to expert-written explanations.
- Assessing the strengths and drawbacks of using generated explanations of suicide risk with human experts in a user experimentation environment.

2. Related Work

2.1. Explainable Artificial Intelligence (XAI) for Mental Health

In recent years, the gradual increase of clinical and translational science in medicine has resulted in the progression of mental healthcare [7]. However, knowledge of mental disorders in terms of their diagnosis, etiopathophysiology, and treatment has not yet been expanded. According to the authors, even when conducting in-depth research in neuroimaging, no single indicator in human brain biology can accurately differentiate between various mental disorders and their subtypes in patients [8, 9]. Data quality and diversity have increased due to technical innovation, but data collection alone will not help improve our understanding of mental disorders. XAI has

the potential to meaningfully analyze complex sets of data points to understand the complex nature of mental disorders better. XAI determines which features in the dataset contribute most to a specific classification pattern by methods such as saliency maps to visualize the relative weight or importance of features in the data [10, 11, 12]. Combining XAI with social media data, facility data, and human sensory data is intended to improve the understanding of recurring patterns of mental health disorders [7, 9, 13]. This will help accurately predict risk and disease trajectories, leading to the development of scalable detection and prevention tools.

2.2. LLMs in Natural Language Processing (NLP)

LLMs have evolved in many variations, showing their powerful capabilities. It is clear that pre-trained models significantly outperform non-pre-trained models in NLP. The advancement and success of NLP attract researchers to explore and address mental health behaviors. According to [14], the lack of a taxonomy, extensive data for evaluation, and the inability to distinguish between mental health disorder-related behaviors and healthy behaviors in social media text are challenges faced in ML and NLP research. Today, BERT [14] is one of the most popular and best-performing methods for language interpretation tasks. BERT-like transformer models usually undergo supervised pretraining on a large corpus [14]. Research has shown that pretraining a model towards a smaller corpus or a specific task, such as in Domain Adaptation (DAPT) pretraining and Task Adaptation (TAPT) pretraining, can improve downstream performance in BERT-like models [14]. TAPT refers to using a smaller pretraining corpus on the unlabeled training set for a given task. In contrast, DAPT uses a large pretraining corpus of unlabeled domain-specific text [15]. The authors, seeking better performance, used DAPT to propose PsychBERT, a new language model modified from the BERT model for language patterns in psychology, psychiatry, mental health, or behavioral health text [14]. BioGPT, a domain-specific generative pre-trained transformed language model, was created to perform biomedical text mining and knowledge discovery [16]. It can generate distinct descriptions of biomedical terms in various large-scale biomedical literature. Downstream tasks such as end-to-end relation extraction, question answering (QA), and document classification were applied and adapted by analyzing and designing the target sequence format and the prompt for better modeling of the tasks [16]. The authors contended that this model achieves satisfactory performance in better biomedical text generation ability than GPT-2 in text generation tasks [16].

²<https://github.com/ggerganov/llama.cpp>

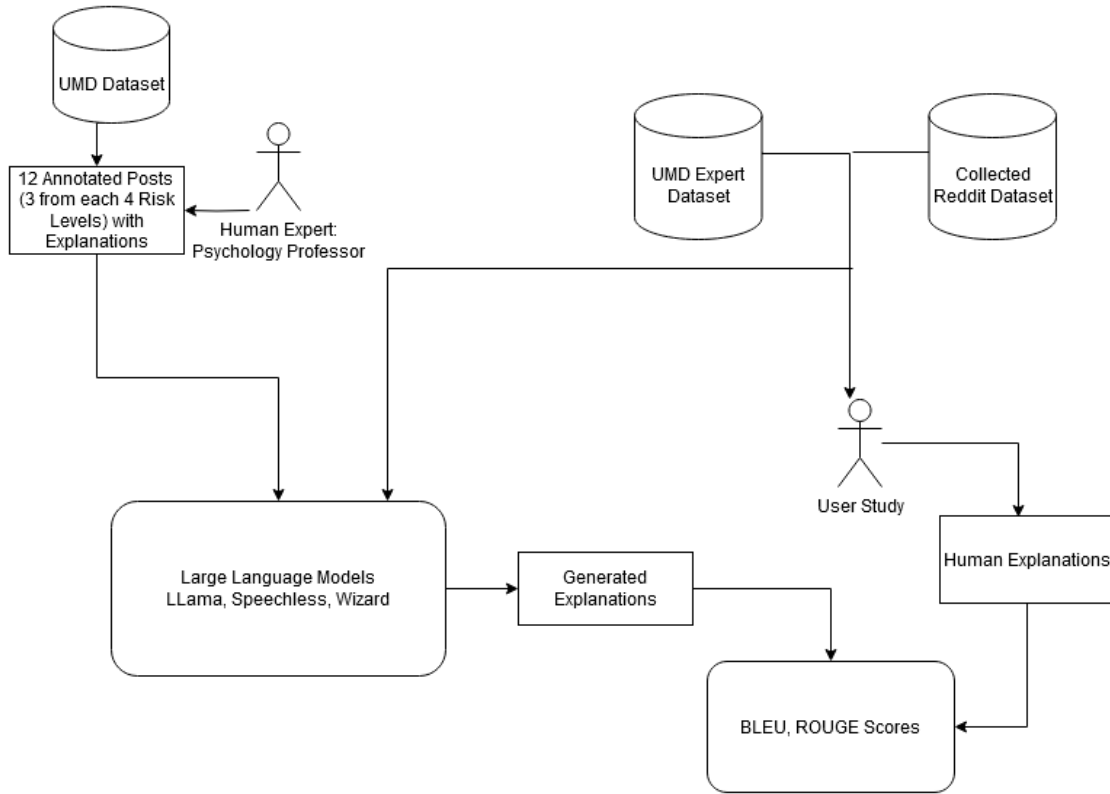


Figure 1: Overall workflow

2.3. Generative Artificial Intelligence in Mental Healthcare

In the realm of NLP, significant strides have recently been made in generative AI. This surge in advancements has spurred innovation, particularly in leveraging generative AI for mental health treatment. One current state-of-the-art model is MentaLLaMA [5]. This is a fine-tuned version of the LLaMA LLM [6] that was trained on mental health data. There has also been research on the use of general-purpose LLMs like ChatGPT for classifying and explaining mental health data such as depression and anxiety [17]. In our work, we are looking at applying and evaluating open-source LLMs to classify and generate natural language explanations.

In our work, we are investigating the use of instruction-tuned, openly available general LLMs to generate explanations for suicide risk. We also conducted a user study with psychology experts to evaluate and verify the quality of the explanations generated from different models.

3. Methodology

Our data pipeline begins with the LLM prompts. To craft these prompts, a psychology professor evaluated 12 samples from the University of Maryland Reddit Suicidality Dataset, Version 2 [18, 19] (UMD dataset). The professor determined the level of suicide risk (no, low, medium, or high risk) for each post and provided explanations for why each post indicated that specific level of suicide risk. One sample was randomly selected from each risk level for the four-shot prompt input, resulting in four samples from the posts annotated by the professor. This four-shot prompt was then utilized as input for different LLMs. The LLMs, along with the UMD expert-labeled dataset and the collected Reddit data, were provided with this prompt to generate explanations for suicide risk. The same datasets were also given to participants in the user study to generate human explanations. The quality of the explanations provided by the users was assessed by comparing them with LLM-generated explanations using Bilingual Evaluation Understudy (BLEU) [20], and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [21] scores. This workflow can be seen in Figure 1.

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...

Response: This post is no risk because...

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...

Response: This post represents high risk because...

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 into post context.

Input: I am feeling...

Response: This post is low risk because...

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...

Response: This post represents medium risk because...

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: (Post from dataset)

Response: This post represents (Will be completed by generative LLM)

Figure 2: The few-shot prompt used for generating explanations

3.1. Data Sources and Pre-processing

We used two social media datasets mined from Reddit; one was labeled by mental health experts (UMD dataset) and has been widely used by researchers, and the other was collected by researchers in our team for other ongoing research work [3]. The detailed descriptions of datasets are provided below:

- UMD Dataset:** The University of Maryland Redd-it Suicidality Dataset, Version 2 Task A. It consists of the post history of 850 users of the r/-SuicideWatch subreddit on the social media website Reddit. Among these 850 posts, 245 posts were classified by mental health experts, and the rest were annotated using a crowdsourcing process. The experts annotated the posts with four classes: no, low, medium, and high risk of suicide. We randomly selected 12 posts annotated from the crowd-sourced section of the UMD dataset (3 from each label). Then, these annotated posts were given to a human expert, a professor of psychology, to classify into four labels and to provide an explanation for each classification. The UMD dataset has been used to help researchers study the risk of suicide in posts that have been uploaded to Reddit. It consists of posts that were uploaded to Reddit prior to 2015. The dataset includes two different sections: a crowd-

sourced testing and training set and a golden standard set labeled by mental health experts. In this dataset, each user is given a label: no risk, low risk, medium risk, or high risk. These risk factors correspond to the risk of suicide for each user based on their writing. It consists of several different tasks. These include using only SuicideWatch subreddit data, using only non-SuicideWatch Reddit data, and using users' complete Reddit history, including SuicideWatch and non-SuicideWatch data. Task A is where one tries to predict users' risk of suicide based only on posts made on the SuicideWatch subreddit, the largest suicide-related subreddit on Reddit. For our study, we used the methodology of Task A.

- Scraped Reddit Data:** This unlabeled data was collected by our research team from the r/SuicideWatch subreddit. Due to the huge amount of data, we used Reddit posts posted to that subreddit during February 2023, which contained the most recent data collected. We generated the dataset by filtering user posts only from /r/SuicideWatch and classifying them individually without considering the user ID. Once again, we limited the text size to 1000 characters to fit within the context limit and to standardize the texts.

Input Text: I have been feeling very sad lately. I have had some thoughts of suicide. I feel very alone. Nobody seems to want to talk to me. School is not going well, and I feel like I am failing always.
Speechless 13B: This post represents medium risk. The person mentions they've "had some thoughts of suicide," which indicates a direct connection with self-harm or ending their life, even though the context doesn't specify if it's the imminent actionable plan. Additionally, feelings of sadness and loneliness contribute to this categorization as well.
WizardLM 13B: This post represents low risk. The poster mentions feelings of isolation, academic struggles, and a history of sadness, but there's no explicit mention or plan for suicide or any immediate danger. It seems they are looking for some form of support/connection, which can be seen as an attempt to prevent escalating distress into more serious issues like self-harming behaviors.
LLaMA Chat 13B: This post represents low risk. The poster mentions feelings of sadness, as well as a history of non-suicidal self-injury in the form of cutting. However, there are no explicit statements about suicide plans or intentions and instead seems to be seeking support from others.
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.

Figure 3: Generated explanations from different LLMs

3.2. LLMs for Text Generation

The Generative LLM we used was the LLaMA model developed by Meta AI research [6]. After weighing performance benefits and resource limitations, we chose to use quantized versions of the models using the llama.cpp GGUF format³. We used the Q5_K_M quantization for each model. This allowed us to run the models at a high speed using only the CPU while maintaining most of the performance of the original non-quantized model. We used the recent LLaMA 2 model as our foundational model. To improve LLaMA's performance across tasks, it needs to be instruction-tuned [22]. This means it must be fine-tuned on instructions and outputs that are similar to what people might ask it to do. The LLaMA 2 Chat model, from Meta AI, was fine-tuned using 27,540 instructions. It also utilizes Reinforcement Learning with Human Feedback (RLHF) [23]. We tested both the 70 billion parameter (70B) and the 13 billion parameter variations (13B) of the chat models.

The other models we tested were WizardLM [24] and SpeechlessLM⁴. WizardLM is an instruction-tuned version of the LLaMA 2 foundational model. Its fine-tuning dataset was generated using a special generation method called Evol-Instruct [24] to generate complex instructions. We tested the 13B model of WizardLM 1.2. The third model we tested was SpeechlessLM. It is also an instruction-tuned version of the LLaMA 2 foundational model. This model is a merge of WizardLM-13B-V1.2 and OpenOrca-Platypus2-13B, which is another popular instruction-tuned LLaMA 2 model. At the time of test-

³<https://github.com/ggerganov/ggml/blob/master/docs/gguf.md>

⁴<https://huggingface.co/uukuguy/speechless-llama2-13B>

ing, it was ranked as one of the top LLM on a variety of LLM benchmarks using the Language Model Evaluation Harness [25]. We chose to use these models since they represent some of the most common and best performing LLMs available at the time of testing. We think that if these LLMs are able to classify and generate explanations for suicide risk successfully, then any more advanced models or fine-tuning will be able to do so as well.

For all the models, we used few-shot prompting to generate the explanations. The posts were taken from the UMD dataset and were labeled with risk level and explanation from a mental health expert. An example of the few-shot prompt can be seen in Figure 2. This prompt was inputted to the LLM. The classification result from the LLM was scraped from the generated text using regular expressions.

3.3. User Study

We conducted a user study by including five PhD students studying psychology. Four of them completed at least their second year of study, and one of them is in his/her first year. Three of the five participants were female, and the rest were male. All participants confirmed that they use social media regularly. We want to clarify that the psychology professor mentioned earlier is not a participant in the study. In addition, he is a member of the research team and one of the authors. The user study consists of three rounds. First, they were randomly presented with two posts from posts originally annotated and explained by our psychology professor to understand their level of understanding and expertise compared to

the professor. In the second round, three different Reddit users’ posts from the UMD Expert dataset were shown to them. They were instructed to classify the risk of suicide (no, low, medium, or high risk) for each of the posts sequentially and provide an explanation for their reasoning. The posts they were given came from a subset of 25 Reddit users’ posts. These Reddit users’ posts came from the UMD expert-labeled dataset that was sampled without replacement. In the third round, we repeated the same process with our second data source, the Reddit data mined by our research team. Table 2 and Table 3 provide the evaluation metrics for rounds 2 and 3, respectively.

3.4. Evaluation

BLEU and ROUGE methods were used to evaluate the generated explanations by comparing them to those the participants in our user study wrote. The human explanations of the user study were used as the baseline and then compared with those generated from the model.

On the crowdsourced data portion of the UMD dataset, we fine-tuned the MentalBERT model to provide a classification baseline. The MentalBERT model is a commonly used model for classification tasks in mental health. The MentalBERT model is fine-tuned on the UMD suicide crowdsourced training dataset for 6 epochs. The training dataset consists of 484 user posts.

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

Table 1
Classification accuracy of LLMs on the UMD dataset

4. Results and Evaluations

This section describes the results of the experimentation and user study. Quantitative results include model accuracy and explanation evaluation scores using BLEU and ROUGE scores. We also discuss the quality of the generated explanations based on human evaluation.

4.1. Quantitative Results

We investigated the classification accuracy of the different LLMs. The classification F1, precision, and recall can be found in Table 1. The model with the top F1 was the WizardLM 13 billion (B) parameter version and Speechless 13B. This model also had the highest precision. The

highest recall was a tie between MentalBERT and SpeechlessLM. For suicide risk detection, recall is essential as we want to minimize the number of cases that go undetected. We suspect that the unique training data that is used for the WizardLM and Speechless models helps them achieve slightly higher accuracy than the LLaMA Chat models. Compared to MentalBERT, we can see that the LLMs were able to classify the risk of suicide fairly effectively.

We also wanted to compare the quality of the generated explanations across the different LLMs by generating explanations for the UMD expert-labeled dataset. An example of how each model generates different explanations can be found in Figure 3. Table 2 lists the average scores grouped by four labels based on ground truth for the UMD dataset (round 2 of the user study). Here, the reference explanations are posts from the UMD dataset annotated by participants in the user study. According to the scores, the LLM was able to most closely match the users’ responses on the low and medium risk posts, with some models, such as LLaMA 2 Chat 70B, performing better than the no-risk classifications. Moreover, the models of which the explanations most closely matched the users were the WizardLM and the LLaMA Chat 70B model; however, there was not a significant difference from the other models. The LLaMA Chat 70B and the WizardLM models performed the best on the entire dataset.

We used the same few-shot prompts that we used with the UMD dataset for round 3, where the participants in the study labeled and wrote explanations for our self-collected Reddit data. The results of the explanation evaluation metrics can be found in Table 3. The results show that the LLaMA Chat 70B has the highest BLEU and ROUGE scores, with the WizardLM model being a close second.

BLEU scores are known to penalize longer texts than the reference ones. In our case, the BLEU scores in both rounds (both tables) did not exceed 0.3. We suspect that one factor causing this may be that the model-generated explanations are more detailed than the human explanations collected during the user study. We observed that during the user study, participants sometimes provided explanations in phrases and incomplete sentences, causing a greater difference between the AI-generated and human explanations, leading to a lower BLEU score. For that reason, we chose to use the ROUGE-1 score, as it will help mitigate the effects of the different lengths of explanations between the participants and the LLM. Similarly to the BLEU score, the ROUGE score also has limitations. It may not fully capture the semantic meaning or coherence of the text summary, as it relies solely on 1-gram overlap. We suspect that the lower BLEU and ROUGE scores can be partially attributed to the small sample size and the short user study explanations. In addition, our evaluation results show that the instruction-tuning

	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

Table 2
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

Table 3
Generation evaluation metrics for collected Reddit dataset

method used by the WizardLM and SpeechlessLM models and the larger parameter size of the LLaMA 70B model give them an advantage in both explanation quality and classification accuracy.

It is often very difficult to judge the risk level of a social media post because suicide risk assessment differs between people. In order to measure agreement between people in our user study, we calculated the Krippendorff alpha coefficient [26, 27]. The alpha value is 0.612, meaning there was a moderate agreement of suicide risk.

4.2. Qualitative Results

We found that our method produces very promising results qualitatively. Each of the models is able to provide a reasonable explanation for the classification. The generated explanations start by mentioning the risk level ("This post represents medium risk of suicide."). Then, the explanations provide reasoning behind the classification using parts from the users' posts as evidence. Many explanations, particularly in the low-risk or medium-risk category, include reasoning for why the post is not in a different risk category. As an example, an explanation of a medium risk category post may include reasoning for why the post is not high risk and why the post is not low risk. Word clouds generated on the explanations for low (Figure 4) and medium risks (Figure 5) are also in line with our observations. For example, the predicted explanations for low-risk posts (Figure 4 right) emphasize the high importance of non-suicidal thoughts with uncertainty by words such as 'feeling', 'mention', and 'indicating' implying low risk that the person may be suicidal. However, in Figure 5, for the medium risk, the strength of the words is similar for both human participants and generated explanations. This demonstrates a moderate

to high risk of suicidal ideation. Some examples of how the different explanations look for the different versions of the LLMs can be found in Figure 3.

One issue we observed in the evaluation was the occurrence of generated explanations that were hallucinations. Hallucinations occur when the model starts providing inaccurate information. While this is a common issue in LLMs, it is highly undesirable [28]. One example of that can be seen in Figure 3 in the LLaMA Chat 13B models explanations. The model mentions that the poster has been cutting themselves. However, that is not mentioned anywhere in the input post.

The mental health expert professor reviewed some LLM classifications and generated explanations. According to him, one of the most significant indicators of risk of suicide was previous suicide attempts, and the LLM did not take that much into consideration. The LLM puts a substantial focus on non-suicidal self-harm. The presence of this factor can be a contributing factor to suicidal behavior, but it does not always correlate directly with it. However, he agreed that the model and explanations successfully distinguished between suicidal thinking and depression. As he observed, neither the model nor the experts could differentiate between "no risk" behavior and "low risk" behavior.

5. Discussion and Limitations

In this work, we see that we are able to use LLMs to predict the risk of suicide and generate natural language explanations for the classification.

There are several limitations in using this method, however. LLMs are very resource intensive, and some mental health facilities may not have the resources. As a result, proper infrastructure, including training and large amounts of computing, needs to be developed for LLMs to benefit mental health professionals. Another limitation is that the classification of the risk of suicide is very subjective as interpreted from natural language. Each expert may find the risk of suicide to be different, thus making it very difficult to benchmark. We saw this quite frequently in our work. Many of the mental health experts we consulted differed in opinion on the risk of

of our abilities to reduce bias and the spread of personal information, and only members of the research team had access to them. Please note that the research is still a work in progress, and none of the models guarantee that the results presented are accurate. These evaluations should not be used as a substitute for evaluations by mental health professionals. Experts and users, such as mental health professionals, are responsible for verifying any results.

8. Acknowledgment

This work is supported by Florida Institute of Technology's Institutional Research Initiatives (<https://www.fit.edu/institutional-research/institutional-research/>) and AccessComputing (<https://www.washington.edu/accesscomputing/>).

Dr. Nasheen Nur is the project's primary investigator and the corresponding author of this paper.

References

- [1] M. Parsapoor Mah Parsa, J. W. Koudys, A. C. Ruocco, Suicide risk detection using artificial intelligence: the promise of creating a benchmark dataset for research on the detection of suicide risk, *Front Psychiatry* 14 (2023) 1186569.
- [2] S. Kaur, R. Bhardwaj, A. Jain, M. Garg, C. Saxena, Causal categorization of mental health posts using transformers, in: *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '22, Association for Computing Machinery, New York, NY, USA, 2023*, p. 43–46. URL: <https://doi.org/10.1145/3574318.3574334>. doi:10.1145/3574318.3574334.
- [3] V. M. Nguyen, N. Nur, W. Stern, T. Mercer, C. Sen, S. Bhattacharyya, V. Tumbiolo, S. J. Goh, Conceptualizing suicidal behavior: Utilizing explanations of predicted outcomes to analyze longitudinal social media data (2023). URL: <https://arxiv.org/abs/2312.08299>. doi:10.48550/ARXIV.2312.08299.
- [4] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, 2016. arXiv:1603.08507.
- [5] K. Yang, T. Zhang, Z. Kuang, Q. Xie, S. Ananiadou, Mentallama: Interpretable mental health analysis on social media with large language models, 2023. arXiv:2309.13567.
- [6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.
- [7] V. Roessner, J. Rothe, G. Kohls, G. Schomerus, S. Ehrlich, C. Beste, Taming the chaos?! using explainable artificial intelligence (xai) to tackle the complexity in mental health research, *European Child & Adolescent Psychiatry* 30 (2021) 1143 – 1146. URL: <https://api.semanticscholar.org/CorpusID:235772369>.
- [8] J.-M. Fellous, G. Sapiro, A. Rossi, H. Mayberg, M. Ferrante, Explainable artificial intelligence for neuroscience: behavioral neurostimulation, *Frontiers in neuroscience* 13 (2019) 1346.
- [9] S. Ali, Abdullah, T. P. T. Armand, A. Athar, A. Husain, M. Ali, M. Yaseen, M.-I. Joo, H.-C. Kim, Metaverse in healthcare integrated with explainable ai and blockchain: enabling immersiveness, ensuring trust, and providing patient data security, *Sensors* 23 (2023) 565.
- [10] M. Nagahisarchoghaei, N. Nur, L. Cummins, N. Nur, M. M. Karimi, S. Nandanwar, S. Bhattacharyya, S. Rahimi, An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives, *Electronics* 12 (2023) 1092.
- [11] N. Nur, Developing temporal machine learning approaches to support modeling, explaining, and sensemaking of academic success and risk of undergraduate students, Ph.D. thesis, The University of North Carolina at Charlotte, 2021.
- [12] D. L. Arendt, N. Nur, Z. Huang, G. Fair, W. Dou, Parallel embeddings: a visualization technique for contrasting learned representations, in: *Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020*, pp. 259–274.
- [13] S. Arisdakessian, O. A. Wahab, A. Mourad, H. Otrouk, M. Guizani, A survey on iot intrusion detection: Federated learning, game theory, social psychology, and explainable ai as future directions, *IEEE Internet of Things Journal* 10 (2022) 4059–4092.
- [14] V. Vajre, M. Naylor, U. Kamath, A. Shehu, Psychbert: A mental health language model for social media mental health behavioral analysis, in: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE Computer*

- Society, Los Alamitos, CA, USA, 2021, pp. 1077–1082. URL: <https://doi.ieeecomputersociety.org/10.1109/BIBM52615.2021.9669469>. doi:10.1109/BIBM52615.2021.9669469.
- [15] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, arXiv preprint arXiv:2004.10964 (2020).
- [16] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, BioGPT: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics* 23 (2022). URL: <https://doi.org/10.1093/bib/bbac409>. doi:10.1093/bib/bbac409.
- [17] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, S. Ananidou, Towards interpretable mental health analysis with chatgpt, 2023. arXiv:2304.03347.
- [18] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, P. Resnik, Expert, crowdsourced, and machine assessment of suicide risk via online postings, in: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 25–36.
- [19] A. Zirikly, P. Resnik, Ö. Uzuner, K. Hollingshead, CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts, in: *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019.
- [20] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [21] C. Lin, Rouge: A package for automatic evaluation of summaries. text summarization branches out, 74–81, 2004.
- [22] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, G. Wang, Instruction tuning for large language models: A survey, 2023. arXiv:2308.10792.
- [23] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, 2023. arXiv:1706.03741.
- [24] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, D. Jiang, Wizardlm: Empowering large language models to follow complex instructions, 2023. arXiv:2304.12244.
- [25] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muenighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2021. URL: <https://doi.org/10.5281/zenodo.5371628>. doi:10.5281/zenodo.5371628.
- [26] K. Krippendorff, *Content analysis : an introduction to its methodology*, SAGE, 2013.
- [27] S. Castro, Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure, <https://github.com/pln-fing-udelar/fast-krippendorff>, 2017.
- [28] M. A. Ahmad, I. Yaramis, T. D. Roy, Creating trustworthy llms: Dealing with hallucinations in health-care ai, 2023. arXiv:2311.01463.