AI THERAPY CHATBOT PAPER

Bright Ofori

Siri Gangadharaiah

University of New Haven

2024

1.Abstract

A chatbot is a system that can converse and interact with human users using spoken, written, and visual languages (Adamopoulou E et al, 2020).

Mental Health therapy is essential in our society today, considering suicidal cases and tendencies increasing every year. Young people with MH issues have experienced various types of social support such as appraisal, informational, emotional, and instrumental support from chatbots.(Kim J et al,. 2018). Yet, professional therapy sessions can be very resource-intensive which makes accessibility not available to all. With the rise of Artificial Intelligence impacting problem-solving in every industry, the mental health therapy industry should not be left out. There is a need for more automated and low-cost ways to access mental health therapy by harnessing Artificial Intelligence without trading off empathy and professionalism from traditional therapists. In this work, we present an AI Therapy chatbot that is designed to hold not only empathic conversations but also flag suicide-prone messages and give hotlines for human intervention. We quest for more automated, scalable, and low-cost ways to access mental health therapy by harnessing Artificial Intelligence without trading off empathy and professionalism from traditional therapists. In this work, we present an AI Therapy chatbot that is designed to hold not only empathic conversations but also flag suicide-prone messages and give hotlines for human intervention. In this work we implement the AI therapy Chatbot in two stages: The first approach employs a suicide classifier that was implemented using BERT fine-tuned on a supervised dataset. The classifier checks messages from users for risk levels and flags the message as suicide or no-suicide with the given threshold. The AI-therapy chatbot then displays a default message with the United States Suicide Hotline, intending to get the user to call for human intervention. In the Second stage, we implement a conversation chatbot utilizing the tranformer based Meta LlaMA 3 8B model fine-tuned on real-world therapy conversation datasets gathered from Kaggle. The Transformer architecture underpins modern language models due to its efficient attention mechanism. This approach aims to generate a contextual approach and to simulate real-world therapy dialogues. This AI-therapist does not eliminate the need for traditional therapists but rather complements it. Through several evaluation tests we demonstrate the chatbot's effectiveness in holding real work therapy sessions and the inevitable importance of this AI-driven approach in mental health support.

2.Introduction

Mental Health issues have become a global concern, in 2020, the United States Disease Control Centre recorded 49000 suicide cases. To put that into perspective that is 11 deaths every minute. These statistics indicate the need for Mental Health Care is on an increasing trajectory. But the problem arises when people cannot always trust professional human therapist with their information, long wait times or high cost of professionals' therapy sessions. There is therefore a need for an alternative that is not only equally efficient, but scalable and also cost effective. Conversational AI, driven by advancements in natural language processing (NLP) and machine learning, offers a promising solution to these challenges.

AI chatbots powered by advance large language models (LLMs) that have been trained on large corpus of data is pragmatic solution to this problem. In the context of mental health (MH), chatbots may encourage interaction with those who have traditionally been reluctant to seek health-related advice because of stigmatization (Lee YC et al., 2020). Not only can these chatbots be assessed 24/7 but are also scalable with reduced to no cost incurred. The effectiveness of chatbots has been explored for self-disclosure and expressive writing (Lee YC et al., 2020). The power of LLMs in recent times have the ability to simulate human therapist sessions thus offer advice and emotional support like real professional therapy sessions. In this project we explored distil-bert-uncased, fine-tuned to classify high risk suicidal messaged from user and give a default message with the United States Suicide Hotline. We also explored LlaMA 3 3B and LlaMA 3 8B fine-tuned on real therapy conversations dataset from Kaggle. Due to limited computational resources, we used the 8-bit quantized versions of the model and applied LLoRa for performance efficient finetuning.

3.Related Work

Previous NLP research in mental health has covered different facets of AI chatbots. [Abhishek Pandey et al., 2024] developed a mental health support chatbot that applies NLP to treat conditions such as anxiety and depression. This work, while emphasizing the effectiveness of chatbots, also indicates further development in terms of usability and integration with healthcare systems as a limiting factor for the wider diffusion of such technologies. Similarly, a recent scoping review by (Mirko Casu et al., 2024) investigated AI chatbots for mental health interventions and reaffirmed their potential to improve emotional well-being but also highlighted challenges concerning user engagement and the necessity for more personalized and context-sensitive systems. In contrast, our project extends this work by including Meta LLaMA for therapeutic conversations and by applying quantization and PEFT for better computational efficiency. Moreover, our work puts great emphasis on subtle and empathetic responses, filling the gap of conversational depth, which is not much emphasized in the other works. Few of the works have explored BERT for different mental health-related tasks such as MentalBERT by (Shaoxiong Ji et al., (2021) which depicted a fine-tuned model on data from social media platforms, such as Reddit, to work on detecting depression and suicidal ideation. While BERT-based models performed commendably in these tasks, the importance of domain-specific pretraining to achieve optimal performance was asserted. Similarly, Mental Health and Stress Prediction Using NLP and Transformer-Based Techniques by IEEE in 2024 used BERT for stress and anxiety predictions but pointed out the challenges that lie in the variability of datasets and the need for fine-tuning on specific mental health issues. In contrast, our project extends these approaches by combining BERT with advanced models like Meta LLaMA, focusing not only on classification but also on enabling deeper, therapeutic conversations. Besides, our use of PEFT and gradient clipping techniques optimizes the computational feasibility of large models, addressing scalability and real-time application challenges.

Advances in conversational AI (e.g., Meta LLaMA).Recent large language model improvements have massively improved the landscape for conversational AI. The models, from 7B to 65B parameters, achieve performance competitive with state-of-the-art results on many tasks, from free-formn generation to multiple-choice questions. This is evinced through the work done by Meta on these foundational models- (Hugo Touvron et al., 2023). Improvements in architecture and efficiency are very critical in bringing even better and more scalable conversational systems.

According to recent studies, patient safety has rarely been evaluated, health outcomes have been inadequately quantified, and no standardized evaluation procedures have been used (Abd-Alrazaq AA et al., 2019). There is therefore a growing concern for health outcomes of effectiveness in already deployed chatbots. Despite recent work's growing use, efficiency and reliability is yet to be researched, especially when it comes to embodiment and modality.

4.Methodology

4.1 Suicidal Classification

Kaggle Dataset consisting of 232704 samples of text scraped from user post with labels or suicide or no-suicide.

-Model: Pre-trained BERT (bert-base-uncased), fine-tuned on mental health data.

-Approach

For suicide classification, we used this encoder-based model due to its effectiveness in classification task. The BERT model was fine-tuned on our dataset and it outputs either suicide or non-suicide.

The model training was conducted on a CUDA-enabled GPU due to the large nature if the model involved and for faster processing times.

Training

- **Optimizer**: The training process we also utilized AdamW optimizer with a learning rate of 10e-5. We wanted to ensure the learning is done in very little steps and no warmup was employed.

- **Learning Rate Scheduler**: A linear scheduler was used with zero warm-up steps to adjust the learning rate gradually over time so ensure stable training and optimization.

- **Number of Epochs:** The model was trained for 2 epochs was used for training , ensuring a balnce betwween computational cost and time.

- **Batch Size**: A batch size of 16 was selected for efficiant memory usuage since the dataset has **232074 unique text**.

- **Loss Function**: Cross-entropy loss was used to evaluate the model's performance during training, as it is well-suited for binary classification tasks.

Evaluation Metrics

The classifier was evaluated on a test dataset of 20% samples, balanced across the two classes for suicide and non-suicide. This approach was to prevent skewness Standard metrics such as precision, recall, and F1-score were used to assess model performance, with results as follows:

Results

	Precison	**Recall**	**F1 score**	**Support**
suicide	0.98	0.98	0.98	11604
Non suicide			0.98	11604
accuracy	0.98	0.98	0.98	23208
Macro avg	0.98	0.98	0.98	23208
Weighted avg	0.98	0.98	0.98	23208

The model achieved a 98% accuracy across both classes, indicating high reliability in distinguishing between suicide-prone and non-suicidal messages. The macro and weighted averages for precision, recall, and F1-score also highlighted the model's balanced performance across the two categories, ensuring minimal bias toward any single class.

4.2 Therapeutic Conversations

- **Datasets**: Conversations between psychologists and patients; labeled Q&A dataset.

  The first Dataset is from Kaggle dataset that contained real world therapy conversations. The second dataset is also from Kaggle that is a collated question answering therapy sessions on reddit. Both datasets were concatenated.

- **Models**: Meta LLaMA (8B parameters).

Quantization

- 8-bit quantization using bitsandbytes library to optimize model memory footprint and usage

```
bnb_config = BitsAndBytesConfig(load_in_8bit=True)
```

Low-Rank Adaptation (LoRA)

LoRA was implemented to enable efficient fine-tuning with minimal parameter updates: it nodified about 5% of model parameters. LoRA Rank (r): 8.
Dropout Rate: 0.1
Target Modules: Query and Value projection layers
Adaptation Alpha: 16

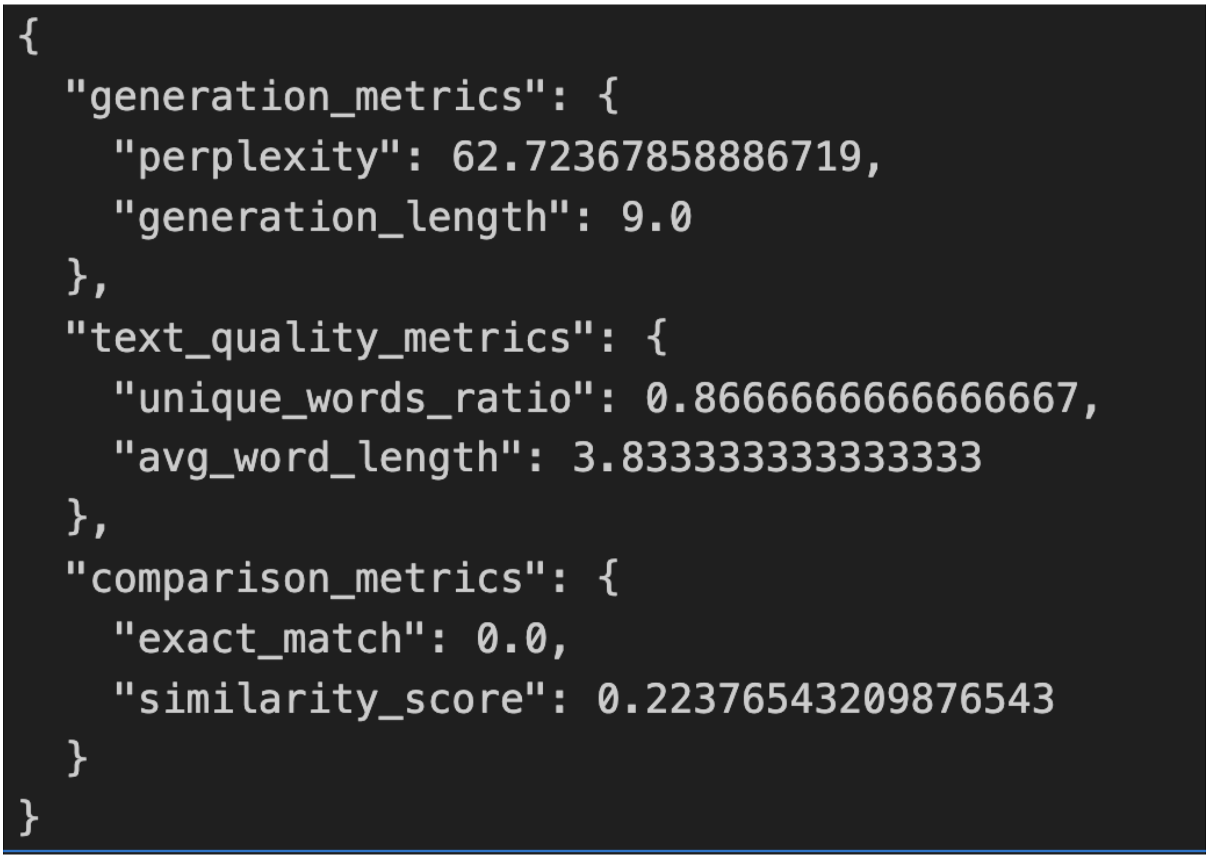
```
peft_config = LoraConfig(
   lora_alpha=16,
   lora_dropout=0.1,
   r=8,
   bias="none",
   task_type="CAUSAL_LM",
   target_modules=["q_proj", "v_proj"]
)
```

Our training strategy priortised computational efficiency and generalization:
Epochs: 3
Batch Size: 5 (per device)
Gradient Accumulation Steps: 8
Learning Rate: 5e-4
Mixed Precision Training: Enabled
Warmup Steps: 50

Computational Efficiency

- **Memory Efficiency:** 8-bit quantization achieved approximately 50% memory footprint reduction
- **Training Speed**: Gradient accumulation enabled stable training with limited computational resources
- **Parameter Efficiency**: LoRA adaptation modified less than 5% of model parameters

Results for llama 3 8B



```
{
  "generation_metrics": {
    "perplexity": 62.72367858886719,
    "generation_length": 9.0
  },
  "text_quality_metrics": {
    "unique_words_ratio": 0.8666666666666667,
    "avg_word_length": 3.8333333333333333
  },
  "comparison_metrics": {
    "exact_match": 0.0,
    "similarity_score": 0.22376543209876543
  }
}
```

5. Discussion

Perplexity: 62.72 (Moderate performance) suggests the model's predictions are somewhat close to the actual distribution but can be improved).
Generation Length: 9.0 tokens (Short responses; might need adjustment for longer, more detailed answers). **Unique Words Ratio of** 0.87 (High diversity in word usage; positive for variety).

The main goal of this project is to develop a therapy chatbot to assist the mental health of people, focusing on engaging in therapeutic conversations and suicidal tendencies. The results depicted that our approach for classification using models like BERT and Meta LLaMA for therapeutic conversation, achieved reliable results. The larger model which has 8 billion parameters demonstrated more promising accuracy and also the ability to handle nuanced queries than the model which has 3 billion parameters. Nevertheless, computational limitations were a thing to tackle, however, we could handle it better by addressing PEFT and quantization techniques. Our classification model also showed some nuance in output. This can furthur be investigated by getting more complex user queries. Despite these advancements, the generalization of the system could further be improved, specifically for handling responses that hold complex emotions.

Furthermore, there is room for improving its empathetic and sentimental perception and ensuring its adaptability in the real-time-based context given by the user. Future work can be done on refining the dataset with more nuanced emotional scenarios, improvising real-time response speed and accuracy, and integrating explainability methods like LIME or SHAP to understand the decisions made the model better.

6. Future Work

Data Augmentation: Increasing the dataset with more diverse examples, particularly those reflecting a mixture of empathetic scenarios and emotions, can also improve the model's precision in classifying ambiguous cases, such as borderline suicidal tendencies.

Model Optimization: Latency can be reduced by Further tuning the model with more epochs and studying parameter-efficient inference techniques, which will also optimize performance for real-world applications.

Improved Empathy and Real-Time Adaptability: By integrating user feedback and sentiment analysis, the chatbot could pick up on emotional context more precisely and change in real-time to provide more empathetic responses.

Explainability: The intervention of interpretability tools like SHAP or LIME would allow the model's decision-making process to be understood, a crucial component in trust-building about mental health applications.

Ethical and Legal Considerations: Privacy concerns and ethical considerations of the AI will have to be addressed as the system gets closer to being deployed in the real world.

LloRa Experiments: Exploring dynamic Lora and Qauntitization techniques using bigger moidels like Meta LlaMA 130B

6.Conclusion

This project is all about developing a therapy chatbot for mental health, which focuses on the detection of suicidal tendencies and performing therapeutic conversations based on the user's input. By incorporating models such as BERT and Meta LLaMA, we proved improvement in accuracy and user interaction, especially with the larger model with 8 billion parameters. Despite coming across the challenges related to computational demands, we could handle them by addressing optimization techniques, further improvements can be done to enhance real-time empathy, adaptability, and response time. Our work and findings emphasize the potential of Natural Language Processing(NLP) in mental health assistance and aid while highlighting the need for legal and ethical consideration and further research in this domain

This project is all about developing a therapy chatbot for mental health, which focuses on the detection of suicidal tendencies and performing therapeutic conversations based on the user's

input. By incorporating models such as BERT and Meta LLaMA, we proved improvement in accuracy and user interaction, especially with the larger model with 8 billion parameters. Despite coming across the challenges related to computational demands, we could handle them by addressing optimization techniques, further improvements can be done to enhance real-time empathy, adaptability, and response time. Our work and findings emphasize the potential of Natural Language Processing(NLP) in mental health assistance and aid while highlighting the need for legal and ethical consideration and further research in this domain

References

1.Adamopoulou E, Moussiades L. An overview of chatbot technology. Proceedings of the 16th International Conference on Artificial Intelligence Applications and Innovations; AIAI '20; June 5-7, 2020; Neos Marmaras, Greece. 2020. pp. 373–83. https://link.springer.com/chapter/10.1007/978-3-030-49186-4_31 . [DOI] [Google Scholar]

2.Lee YC, Yamashita N, Huang Y. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. Proc ACM Hum Comput Interact. 2020 May 29;4(CSCW1):31. doi: 10.1145/3392836. https://dl.acm.org/doi/10.1145/3392836 . [DOI] [Google Scholar]

3.Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform. 2019 Dec;132:103978. doi: 10.1016/j.ijmedinf.2019.103978. https://eprints.whiterose.ac.uk/151992/ S1386-5056(19)30716-6 [DOI] [PubMed] [Google Scholar]

4Abhishek Pandey; Sanjay Kumar**.**Mental Health and Stress Prediction Using NLP and Transformer-Based Techniques | IEEE Conference Publication | IEEE Xplore **Published in:** 2024 IEEE Symposium on Wireless Technology & Applications (ISWTA)
Date of Conference: 20-21 July 2024Date Added to IEEE *Xplore*: 03 September 2024 DOI: 10.1109/ISWTA62130.2024.10651843 Publisher: IEEE Conference Location: Kuala Lumpur, Malaysia

5.Mirko Casu, Sergio Triscari 2, Sebastiano Battiato 1,3, Luca Guarnera 1, and Pasquale Caponnetto
AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications
Published in: Applied SciencesDate Received: 31 May 2024 **Date Published:** 5 July 2024
DOI: [10.3390/app14135889](https://www.mdpi.com/2076-3417/14/13/5889)

6.[2302.13971] LLaMA: Open and Efficient Foundation Language Models.

Hugo Touvron, Thibaut Lavril[1], Gautier Izacard[1], Xavier Martinet \ANDMarie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal AND Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin \ANDEdouard Grave[1], Guillaume Lample[1] \AND Meta AI Equal contribution. Correspondence: {htouvron, thibautlav,gizacard,egrave,glample}@meta.com

7. Kim J, Kim Y, Kim B, Yun S, Kim M, Lee JS. Can a machine tend to teenagers' emotional needs? A study with conversational agents. Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems; CHI EA '18; April 21-26, 2018; Montreal, Canada. 2018. p. LBW018. https://dl.acm.org/doi/10.1145/3170427.3188548 . [DOI] [Google Scholar]

8.Lee M, Ackermans S, van As N, Chang H, Lucas E, IJsselsteijn W. Caring for Vincent: a chatbot for self-compassion. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; CHI '19; May 4-9, 2019; Glasgow, Scotland, UK. 2019. p. 702. https://dl.acm.org/doi/10.1145/3290605.3300932 . [DOI] [Google Scholar][Ref list]

9. Ji, Shaoxiong; Ji, Shaoxiong; Zhang, Tianlin; Ansari, Luna; Fu, Jie; Tiwari, Prayag; Cambria, Erik MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare.Publication Forum https://jfp.csc.fi/jufoportaali?Jufo_ID=70622