# Mental Health Counseling Conversational Chatbot

Aditya Jain, Ritwik Harit, Shrey, Vasan Vohra, Vasu Kapoor, Vinayak Sharma
*Dominos*
*Roll Numbers: 2021511, 2021557, 2021562, 2021572, 2021573, 2021574*

***Abstract*—Mental health is a widely recognized crisis globally. It is predicted that depression will become the leading cause of disease burden globally by 2030. With the scarcity of associated professionals and their limited accessibility, we present to build a first-line support system in the form of a conversational chatbot. We have developed an agent capable of identifying mental health conditions and providing emotional support via empathetic responses. We have incorporated the ability of multilingualism to cater to a broad audience easily, along with public deployment on popular platforms.**

## I. Introduction

Mental health issues have become a growing concern globally, with many individuals lacking access to timely and affordable support. Advances in artificial intelligence have paved the way for conversational agents to fill this gap by providing accessible and scalable mental health assistance. A chatbot capable of engaging in mental health conversations must accurately detect emotions and respond empathetically.

This paper presents the results of developing and evaluating a mental health counseling conversational chatbot. The primary goal is to leverage state-of-the-art natural language processing models to address mental health queries empathetically and effectively. The study uses several datasets, which consist of question-answer pairs in English, and evaluates the chatbot's performance across both English and Hindi scenarios.

The meta-llama/Llama-2-7B model was employed with 4-bit quantization to ensure efficient computation. The evaluation focused on Zero-Shot and Few-Shot prompting techniques, examining metrics such as BLEU, BERT, Meteor, ROUGE, TextBlob Sentiment, and VADER Sentiment to measure linguistic and emotional accuracy.

The results underline the challenges in handling low-resource languages like Hindi while showcasing the potential for high-resource languages like English. This work provides insights into improving language generation tasks for multilingual mental health applications.

## II. Dataset Description

We have used following datasets.

- *samhog/psychology-10k*
- *Amod/mental-health-counseling-conversations*
- *ZahrizhalAli/mental-health-conversational-dataset*

We used the *samhog/psychology-10k* dataset which consists of 9,846 question-answer pairs in the English language. The questions consisted of mental health queries and the answers were provided by licensed psychologists. We used the first 3000 samples from the dataset to evaluate the performance of our baseline models.

For the dataset, we created the following scenarios:

- English: Both the questions and the answers were in English Language. The prompts to the model were also in English Language.
- Hindi: We employed three strategies to create prompts for the Hindi language. We used system prompts as templates. The following are the ways we modified the few-shot learning system template:
  - The instructions and the examples were given in English.
  - The instructions and the examples were given in Hindi.
  - The instructions were given in English and the examples were given in Hindi.

We used *Amod/mental-health-counseling-conversations* dataset, it contains anonymized conversations designed to assist in training AI systems for tasks related to mental health and counseling. This dataset consists of conversation or dialogue snippet between a counselor and a client. These conversations cover a wide range of mental health topics, providing diverse examples for training and evaluating natural language processing models.

We used *ZahrizhalAli/mental-health-conversational-dataset*, it focuses on improving conversational AI for mental health support. It contains 175 examples of conversational exchanges, with each entry comprising a question and an answer. The questions typically reflect patient concerns, and the answers simulate responses from healthcare providers, emphasizing anonymity and pre-processing to ensure data privacy.

## III. Related Work

[5] present their findings on how LLMs are more empathetic than humans. Their evaluations were able to show that LLMs can be leveraged for generating empathetic responses to both positive and negative containing dialogues as illustrated by their superior responses. These models as well as human responses were evaluated by human participants that proved LLMs outperforming the human responses.

[6]Vitalk is a chatbot platform which provides free mental health content in a conversational manner. The conversations included lessons on managing stress and some physical exercises such as breathing, meditation and relaxation. Moreover, the chatbot uses emojis, GIFs and some gamification tools to improve user engagement. Vitalk was able to significantly reduce the anxiety, depression and stress symptoms.

[7] MAML: They do meta-training on English and Swahili scripts with english as support set and Swahili as query set to facilitate cross lingual transfer. Additionally, they also do meta-adapting which leverages only the low-resource language to form support and query sets enhancing the model's ability to learn under resource-scarce conditions. Support sets act as examples for the model to learn the task and query sets are testing examples that are evaluated.

[7] In-context Learning using LLMs. This setup utilizes LLMs that were trained on specific tasks by looking at examples and descriptions of what needs to be done. They create a template for prompts for the models to understand and work with. There leveraged zero-shot prompting where the LLMs were required to answer the questions in the new language without prior examples of that language as well as Few-shot prompting where the model sees some examples and finds patterns to answer a new question. They create examples and queries in the same language as well as in different languages which is Swahili in this case. To create prompts, they leverage machine translation, Swahili prompts from native speakers as well as Cross-lingual prompting.

## IV. Model Description

We used LLaMA 2 of 7 billion parameters from Meta which has been specifically fine-tuned for conversational environments. With moderate parameter count, it has displayed strong performances. It has the capability to understand and adapt to input data and generate friendly, coherent and meaningful responses. These features are fitting for mental health counseling conversational chatbot and hence we opted for this model as the mode for conversation with the users.

[3] LlaMA models have displayed impressive results on various NLP tasks such as text generation, question answering, etc. It is a publicly available model, allowing developers and researchers to use the model for their use case. Their training involved pre-training on publicly available online resources as well as RLHF(Reinforcement Learning with Human Feedback) which includes rejection sampling and PPO(Proximal Policy Optimisation). They also use Grouped Query Attention that improves scalability of large models where they share projections of key and value vectors while maintaining impressive performances[4].

## V. Evaluation Metrics

### A. BLEU Score

It is the geometric average of the modified n-gram precisions $p_n$, using n-grams up to length $N$ and positive weights $w_n$ summing to one. Let $c$ be the length of the predicted sentence and $r$ be the ground truth sentence length. The brevity penalty $BP$ is calculated as follows:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

We will use BLEU-1, BLEU-2, BLEU-3, and BLEU-4 by adjusting $N$ and applying uniform weights $w_n = 1/N$. The BLEU score is :

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where $p$ is the penalty.

### B. BERT Score

BERT Score leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. It has been shown to correlate with human judgment on sentence-level and system-level evaluation. Moreover, BERT Score computes precision, recall, and F1 measure, which can be useful for evaluating different language generation tasks.

### C. Meteor Score

This score goes beyond simple n-gram matching by considering word-to-word matches, including synonyms, stemming (matching different forms of a word), and paraphrasing. This allows METEOR to better account for semantically similar but not identical words or phrases.

METEOR takes into account both the precision and recall while evaluating a match [1].

- **Precision:**
$$P = \frac{m}{w_t}$$
where: $m$: Number of unigrams in the candidate translation also found in reference
$w_t$: Number of unigrams in candidate translation.
- **Recall:**
$$R = \frac{m}{w_r}$$
where $w_r$: Number of unigrams in reference translation.
- **F-mean:**
$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

The Chunk Penalty is computed as follows:

$$p = 0.5 \left(\frac{c}{u_m}\right)^3$$

where:
$c$: Number of chunks in candidate
$u_m$: Unigrams in candidate.

The final meteor score combines the F-score computed from precision and recall with the chunk penalty.

$$M = F_{\text{mean}}(1 - p)$$

### D. ROUGE Score

ROUGE focuses more on recall and is particularly useful for comparing the generated summaries with reference summaries. Here's an overview of ROUGE and its variants:

- **ROUGE-N**: Measures the overlap of n-grams between the generated summary and reference summaries.

$$\text{ROUGE 1} = \frac{\text{Number of overlapping unigrams}}{\text{Total number of unigrams in reference}}$$

- **ROUGE-L**: Measures the longest common subsequence (LCS) between the generated summary and reference summaries.

$$\text{ROUGE L} = \frac{\text{Length of LCS}}{\text{Total number of words in reference}}$$

### E. TextBlob

TextBlob is a Python library used for Natural Language Processing Tasks. It returns two things: polarity and subjectivity. The polarity score is within the range -1 to 1. If the score is higher than 0 then it shows positivity, equal or close to 0 shows neutral, else it shows negativity. Subjectivity score is within the range 0 to 1. If the score is close to 1 then the sentence is subjective, if it is close to 0 then it is objective. For this project we are interested in using the polarity score.

### F. VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool[2]. It evaluates how negative or positive a sentiment is. It returns the output in a Python dictionary format consisting of four keys: negative, positive, neutral, and compound (overall sentiment). The range of compound scores is within -1 to 1, while the other three scores are the proportion of sentiment that falls in these categories and sums to 1.

## VI. Results

Tables VI demonstrate performance across English and Hindi datasets.

## VII. Conclusion

Insights and Conclusions from the provide valuable information.

General Conclusions:

- Metric Wise :
  - The process of fine-tuning is seen to improve the metrics generally over the data.
  - We see improvements in the semantic similarities in fine-tuning, showing better understandability for the model.
  - We see improvements in the semantic similarities in fine-tuning, showing better understandability for the model.
  - We see improvement in the BLEU, BERT, ROUGE, and METEOR scores on fine-tuning.
  - The METEOR score is more robust compared to BERT for semantic similarity. The BERT score was already good, and we see a significant improvement in the METEOR score, which shows the fine-tuning is properly carried out.
- English Vs Hindi :
  - The English dataset shows better results and better improvements after the fine-tuning as compared to the Hindi datasets.
  - The fewer improvements in the Hindi datasets show the need for more robust Hindi datasets and some language-specific enhancements.
  - We also observed that there might be some information loss in the case of the Hindi datasets, which we made by translating the English to Hindi datasets, leading to lower results in Hindi.
  - Thus, better tuning happens in the case of structured datasets as compared to the ones that are converted, also potentially signaling toward noisy data.

## Team Details

**Team Name:** Dominos Github link

**Team Members:**

| Name | Roll Number |
|---|---|
| Aditya Jain | 2021511 |
| Ritwik Harit | 2021557 |
| Shrey | 2021562 |
| Vasan Vohra | 2021572 |
| Vasu Kapoor | 2021573 |
| Vinayak Sharma | 2021574 |

## References


[1] "Meteor Score," *Machine Learning Interview*, 2023. [Online]. Available: https://machinelearninginterview.com/topics/machine-learning/meteor-for-machine-translation/.

[2] "VADER Sentiment Analysis," 2016. [Online]. Available: https://github.com/cjhutto/vaderSentiment.

[3] "Brief Introduction to Llama 2." [Online]. Available: https://medium.com/@florian_algo/brief-introduction-to-llama-2-cec2d59fc13f.

[4] "Demystifying GQA — Grouped Query Attention for Efficient LLM Pre-training." [Online]. Available: https://towardsdatascience.com/demystifying-gqa-grouped-query-attention-3fb97b678e4a.

[5] "Welivita, A., Pu, P. (2024). Are large language models more empathetic than humans? arXiv preprint arXiv:2406.05063." .

[6] "Daley, K., Hungerbuehler, I., Cavanagh, K., Claro, H. G., Swinton, P. A., Kapps, M. (2020). Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Frontiers in digital health, 2, 576361."

[7] "Lifelo, Z., Ning, H., Dhelim, S. (2024). Adapting mental health prediction tasks for cross-lingual learning via meta-training and in-context learning with large language models. arXiv preprint arXiv:2404.09045." .

TABLE I
INTEGRATED METRICS TABLE ACROSS ENGLISH DATASETS

| **Metric** | **ENGLISH Test Data** | | **AMOD** | | **English Conversational** | |
|---|---|---|---|---|---|---|
| | **Zero Shot** | **Fine Tuned** | **Zero Shot** | **Fine Tuned** | **Zero Shot** | **Fine Tuned** |
| avg_bleu | 0.04 | 0.05 | 0.00 | 0.00 | 0.00 | 0.026 |
| avg_bert | 0.83 | 0.89 | 0.80 | 0.84 | 0.81 | 0.86 |
| rouge1_fmeasure | 0.20 | 0.37 | 0.20 | 0.31 | 0.20 | 0.36 |
| rouge2_fmeasure | 0.008 | 0.15 | 0.007 | 0.005 | 0.007 | 0.10 |
| rougeL_fmeasure | 0.10 | 0.24 | 0.10 | 0.16 | 0.10 | 0.20 |
| avg_meteor | 0.16 | 0.35 | 0.11 | 0.16 | 0.10 | 0.20 |
| avg_textblob_sentiment | 0.11 | 0.17 | 0.13 | 0.17 | 0.01 | 0.16 |
| avg_vader_sentiment | 0.48 | 0.69 | 0.40 | 0.65 | 0.38 | 0.49 |

TABLE II
INTEGRATED METRICS TABLE ACROSS HINDI DATASETS

| **Metric** | **HINDI Test Data** | | **Hindi Conversational** | | **Translated to Hindi Test** | |
|---|---|---|---|---|---|---|
| | **Zero Shot** | **Fine Tuned** | **Zero Shot** | **Fine Tuned** | **Zero Shot** | **Fine Tuned** |
| avg_bleu | 0.0193 | 0.02 | 0.00 | 0.00 | 0.01 | 0.02 |
| avg_bert | 0.9036 | 0.92 | 0.90 | 0.91 | 0.87 | 0.91 |
| rouge1_fmeasure | 0.0009 | 0.00 | 0.00 | 0.00 | 0 | 0.00 |
| rouge2_fmeasure | 0.0000 | 0.00 | 0.00 | 0.00 | 0 | 0.00 |
| rougeL_fmeasure | 0.0009 | 0.24 | 0.00 | 0.00 | 0 | 0.00 |
| avg_meteor | 0.02 | 0.18 | 0.01 | 0.10 | 0.07 | 0.09 |
| avg_textblob_sentiment | -0.0027 | 0.00 | 0.00 | 0.00 | 0 | 0.00 |
| avg_vader_sentiment | -0.0001 | 0.00 | 0.00 | 0.00 | 0 | 0.00 |