

Mental Health Counseling Conversational Bot

Group 1: Dominos



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI



Introduction



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 propose to build a first-line support system in the form of a conversational chatbot.

We aim to develop an agent capable of identifying mental health conditions and providing emotional support via empathetic responses. We also plan to incorporate the ability of multilingualism, particularly Hindi, to cater to a broad audience easily, along with public deployment on popular platforms.



Dataset Description



Dataset Description



We have used following datasets.

1. **samhog/psychology-10k**
2. **Amod/mental-health-counseling-conversations**
3. **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.

Dataset Description



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 counselling. This dataset consists of conversations or dialogue snippets between a counsellor 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.



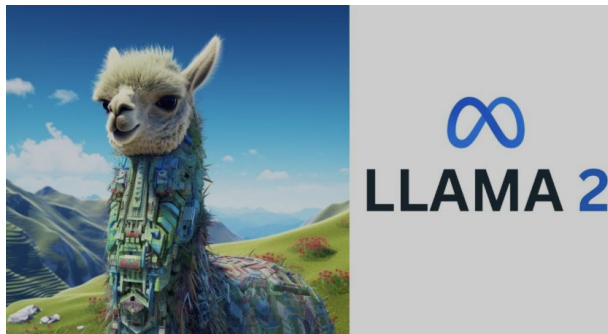
Model Description



LLAMA -2 7b chat



- Llama 2 has 7 billion parameters and has displayed good performances with moderate parameter count.
- Llama 2 models have specifically fine-tuned for conversational environment making it a suitable model for our mental health counseling conversational chatbot.
- It is capable of producing coherent, friendly and meaningful responses to user queries or dialogues.
- It is pre-trained on publicly online data and then fine-tuned using RLHF with rejection sampling and PPO.
- They additionally use grouped query attention that shares key and value vectors across the attention layer decreasing memory load and faster inference.



Evaluation Metrics



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.

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.



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.

- **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.

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

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.

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

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.



Vader



VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool. 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.

Interface



Conversation in English



HealthyChat
bot

November 19

/start 11:14 ✓

Hi! How can I assist you? 11:14

Hi, I have been feeling really stressed lately and don't know what to do about it. 11:15 ✓

It's important to identify the sources of your stress and develop coping mechanisms. Let's work together to identify your stressors and develop a plan to manage them. We can also explore relaxation techniques such as deep breathing, progressive muscle relaxation, or mindfulness meditation. Additionally, we can work on developing a self-care plan that includes healthy habits such as exercise, healthy eating, and social support. Remember that you don't have to go through this alone, and I'm here to help you. How does that sound? Would you like to start by identifying your stressors? 11:15

It's mainly work. I feel like I'm always behind, and no matter how hard I try, it's never enough. 11:16 ✓

It's mainly work. I feel like I'm always behind, and no matter how hard I try, it's never enough. 11:16 ✓

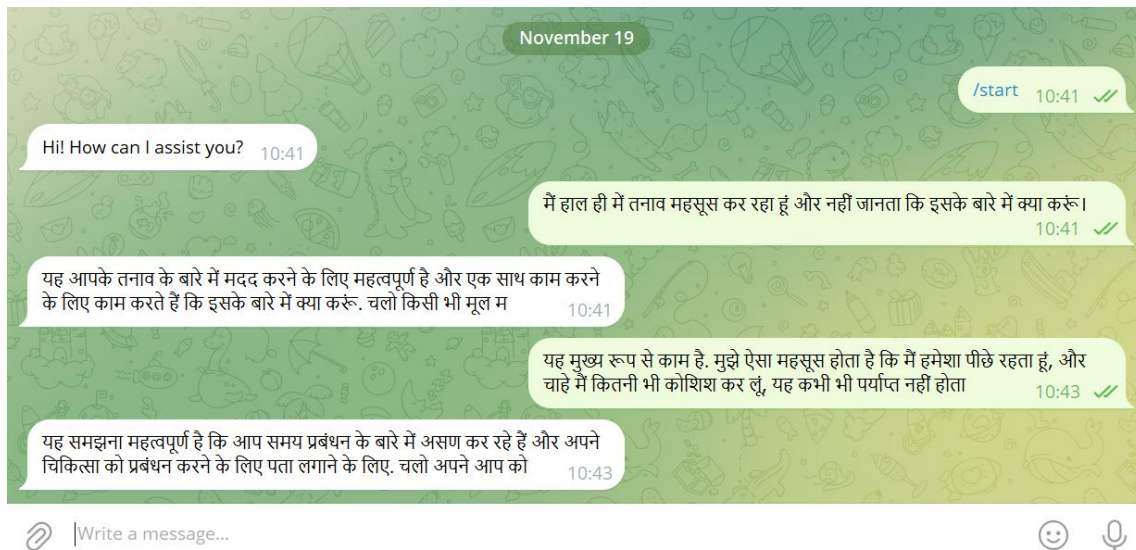
It's important to prioritize self-care and set realistic goals. Let's work on creating a plan to help you manage your workload and reduce stress. We can also explore any underlying issues that may be contributing to your feelings of being behind and unproductive. We can work on developing coping strategies and time management techniques to help you stay on top of your workload. Let's work together to create a plan that works for you. How does that sound? Would you like to start by identifying your most important tasks and prioritizing them? We can also work on breaking down larger tasks into smaller, more manageable chunks. Let's work together 11:16

I think prioritizing tasks could help. 11:17 ✓

It's great that you recognize the importance of prioritizing tasks. Let's work together to develop a plan to help you manage your tasks and workload. Would you like to start by identifying the tasks you need to complete? We can also discuss any challenges you're experiencing with prioritizing tasks and work on finding ways to overcome them. I'm here to listen and help you feel more in control of your workload. What would you like to do first? 😊 11:17

Write a message...

Conversation in Hindi



The responses are more biased towards providing emotional support over processing the entire context of the user's query as explained in the conclusions.

Results





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

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

Conclusion



Conclusion and Discussion



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 on fine-tuning showing better understandability for the model.
 - We see improvement in the BLEU, BERT, ROUGE and the 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 shows the fine-tuning is properly carried out.

Conclusion and Discussion



- **English Vs Hindi :**
 - The English dataset shows great better results and better improvements after the fine-tuning as compared to the hindi datasets.
 - The less 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 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 which are converted, also potentially signaling towards noisy data.



Group members



- Aditya Jain - 2021511
- Ritwik Harit - 2021557
- Shrey IIITD - 2021562
- Vasan Vohra - 2021572
- Vasu Kapoor - 2021573
- Vinayak Sharma - 2021574



GitHub Repository



- GitHub Repository: <https://github.com/Ritwik21557/LLM-PROJECT>



THANK YOU



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

