PHR-CHAT: A Llama 2 BASED CHATBOT FOR IMPROVING MENTAL HEALTH

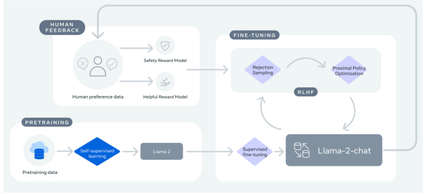
Vibhor Agarwal: 2020349, Manik Arora: 2020519
BTP Advisor: Dr. Tavpritesh Sethi
BTP Commitee Members - Dr. Ganesh Bagler, Dr. Vibhor Kumar

Abstract

- Integrates a mental health chatbot into the Public Health Record Android application.
- Utilizes fine-tuned Llama-2-7b and Llama-2-13b chat models for mental therapy.
- Uses roberta-base for detecting suicidal tendencies and hate speech, acting as model guardrails. This ensures human handoff.
- Prompt Engineering and fine-tuning to ensure cheerful responses.
- Made use of quantization techniques like QLoRA to effectively fine-tune LLM with memory constraints.
- Deployed model as an OpenAI compatible API using HF-TGI Server for fast inference. Made REST endpoint for Guardrails.

Introduction

- Supervised Fine Tuning (SFT) used to develop llama-2-7b-chat models for mental therapy purposes.
- A synthetically generated mental therapy dataset using GPT-3.5 was used for SFT using QLoRA.
- Necessary guardrails were developed using BERT models for suicide and hate-speech classification fine-tuned using datasets from Reddit.



Training and Finetuning process of Meta LLaMA-2 models

Methodology

- **Training and Deployment Setup**: RTXA5000 24GB GPU and Intel Xeon CPU. Training using HF Trainer.

Llama-2-7b-chat-hf-phr_mental_therapy_v2 model

- **Dataset:** Mental Therapy conversational dataset synthetically generated using GPT3.5 API and preprocessed to have a max of 1024 tokens.
- Model chat-template applied to get single text column from multi-turn conversations.
- Supervised fine-tuning (SFT) using quantization techniques like QLoRa for GPU memory-constrained scenarios. Did 4-bit NF4 quantisation.
- **Hyperparameters**:
  - num_epochs = 1
  - max_seq_length = 1024
  - learning_rate:2e-05
  - train_batch_size =1, eval_batch_size = 8
  - early stopping: (5 patience, 0.001 threshold)

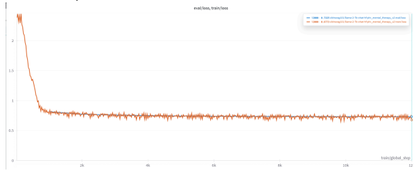
Roberta-base-suicide-prediction-phr-v2 model

- **Suicide Watch Dataset**: Taken from Kaggle and contains ~228k examples, sourced from Reddit.
- Preprocessed by removing numbers, URLs, and emojis. Stopwords were not removed, as BERT needs them. samples with more than 512 tokens dropped.
- Binary class classification problem using BERT model.
- **Hyperparameters:**
  - num_epochs: 3
  - learning_rate: 2e-05
  - train_batch_size =16, eval_batch_size = 32
  - early stopping: (5 patience, 0.001 threshold)

Model Results

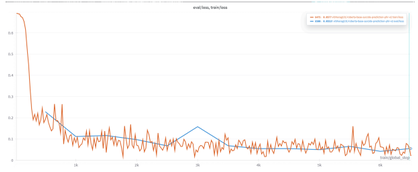
Llama-2-7b-chat-hf-phr_mental_therapy_v2 model

- The fine-tuned model achieved an eval loss of 0.7325.
- A stopping criterion of eval_loss (patience 5, 0.001 threshold) was used.



Roberta-base-suicide-prediction-phr-v2

- Model Achieved the following metrics on the suicide-watch dataset, eval split:
  - Eval loss: 0.0553
  - Accuracy: 0.9869
  - Recall: 0.9846
  - Precision: 0.9904
  - F1: 0.9875
- A stopping criterion of eval_F1 (patience 5, 0.001 threshold) was used.



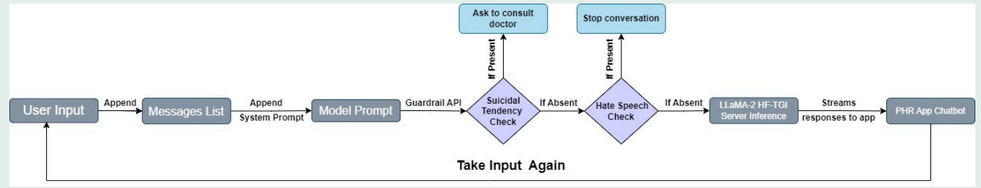
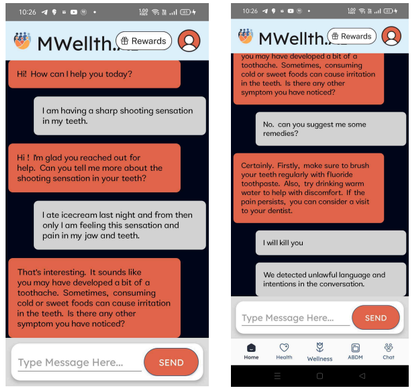


Figure 1: Chatbot Architecture

PHR App Integration

- Deployed Llama model on a local server using Hugging Face text generation inference server for production-ready fast inference.
- The guardrail API is deployed using Flask and Pytorch. This ensures safety and human handoff.
- The chatbot is integrated using these APIs in the PHR Android application for a seamless user experience and enables easy access to basic mental therapy.



Future Work

- Conduct the safety testing of the chatbot after getting IRB Approval for rollout in the PHR app.
- Collect human preference data for aligning the model using Direct Preference Optimisation.
- Work on newly announced LLaMA-3 Models.
- Work on improving the guardrails like suicide and hate-speech classification.

Models and Datasets available at https://huggingface.co/vibhorag101