Analysis of zero shot approach on fine-tuned and untrained MISTRAL-7B model

Sai Hemanth Kilaru,
skilaru@arizona.edu

$-> GitHubRepository <-$

Abstract

This research analyzes the performance of the Mistral-7B model in a zero-shot context using a medical chatbot dataset. The approach integrates a transformer-based framework with quantization techniques to evaluate the model's ability to generate accurate medical responses to the patient's questions. The evaluation process involved refining the dataset to eliminate unhelpful responses and subsequently generating outputs for a chosen set of test samples. Metrics such as BERT Score and ROUGE were used to evaluate the quality of the responses relative to the correct answers.The results indicate that although the base Mistral-7B model exhibits significant capabilities without fine-tuning, there exists potential for imrpoved performance through specialized training. These results highlight possible methods for improving model effectiveness and provide insights into the utilization of large language models in the medical field.

1 Introduction

Recent developments in natural language processing (NLP) have made huge changes in the AI applications, especially in specialized fields such as healthcare. Language models such as Mistral-7B have demonstrated remarkable abilities in understanding and generating like a human text, leading to investigations into their effectiveness in particular fields, including medical dialogue systems. The concept of the zero-shot learning, where models predict without task specific training, has gained prominence as a method of evaluating such capabilities, particularly in the case of absence of labeled datasets. Previous research has shown that well trained models can generalize effectively to the new challenges, utilizing their comprehensive knowledge to produce contextually appropriate responses. However, the practical evaluation of these models in specific applications remains largely unexplored.This research aims to evaluate the model's effectiveness and feasibility in real-world medical cases by comparing the quality of generated responses to human created references using metrics like BERT Score and ROUGE.

2 Related Work

Evaluating language models in particular fields such as healthcare increasingly important in recent years. Research shows the importance of domain adaptation in improving model accuracy. Large language models (LLMs) can generate relevant text, yet their performance often improves with the specific fine-tuning.Zero-shot learning has become prominent in evaluating LLMs, as shown with the GPT-3, showing effective task adaptation. Research in the medical field demonstrates that LLMs can handle the medical queries with varying success based on complexity. Evaluation metrics such as BERT Score and ROUGE are widely recognized for evaluating the generated text quality. BERT Score effectively evaluates semantic similarity, whereas ROUGE focuses on the the overlap of n-grams.This study builds on the existing literature by by examining the zero-shot capabilities of the Mistral-7B model in medical dialogue scenarios.

3 Methodology

This study evaluates the performance of the Mistral-7B model in producing responses to medical queries in a zero-shot context. The methodology consists of several components:

3.1 Dataset

The evaluation dataset was taken from the ruslan mv AI Medical Chatbot dataset(in hugging face), comprising of dialogues between patients and doctors. Responses lacking proper information were excluded to maintain relevance, focusing on the generation of informative outputs.

3.2 Fine-Tuning and Model training

The fine-tuning parameters were set with the help of the Lora config.The model training was performed on the base-model with the selected parameters and it was trained with the 500 samples of data(In order to avoid the GPU from crashing) and trained the data with the 225 epoches. The training metrics were tracked in wandb and presented in the form of the graphs in the link generated after the training process. In the another case, no training/fine-tuning is done for the base-model. direct zero-shot evaluation is performed.

3.3 Zero-Shot Evaluation

Here, we follow the zero shot approach that is ; evaluating the fine-tuned model on new test samples it hasn't seen before.The function containing the evaluation metrics such as BERT score and ROGUE is written along with this.BERT Score evaluates semantic similarity, while ROUGE measures n-gram overlaps, providing a comprehensive view of response quality The scores are then calculated and then appended in a dictionary in the specific order ; required for the output presentation.Results were organized into a structured format and saved in a CSV file for the furher analysis.

3.4 Testing

After ensuring that the fine-tuned model performed well by observing the metrics scores , in the csv file; i have written a function such that it made the model in a form of a query based answering system like a chatbot.Then , with the help of that function , i have tested the model on a few samples on my own and it went well.

4 Evaluation Metrics

To evaluate the Mistral-7B model's performance comprehensively, we employed two primary metrics:

1. **BERT Score**: By using contextual embeddings from the BERT model, BERTScore evaluates the semantic similarity between generated and reference responses, offering insights into the model's accuracy in understanding medical terms.

2. **ROGUE**: ROUGE measures the n-gram overlaps between generated and reference texts, including variants such as ROUGE-N , ROUGE-L,etc.., to capture response quality .

5 Results and Discussion

The evaluation of the fine-tuned Mistral-7B model and the base model is stored in the evaluation.csv and base model evaluation.csv files, which include key metrics: prompt, reference, prediction, bert precision, bert recall, bert f1, rouge1, rouge2, rougeL, and rougeLsum.

As there are 50 columns each in the dataset , i cannot briefly tell the result for all the samples.Here are the overview of both te metrics:

1. **BERT Metrics**: The fine-tuned model scores indicated the strong semantic similarity to reference answers, where as the base model indicated the same in most cases.This shows that fine-tuning mostly matched that base model.

2. **ROUGE Scores**: These scores are mostly similar for both the models but the fine tuned model has improvised in few samples.

6 Conclusion

This study successfully evaluated the performance of the Mistral-7B model in generating responses to medical inquiries using a zero-shot methodology. The results indicate that the model shows significant proficiency in understanding and responding to complex medical questions, as told by strong BERT Score and ROUGE metrics. Despite the model's satisfactory performance without fine-tuning, the further research showed the enhancements as fine-tuning the model on targeted medical datasets , improved its efficiency. This research contributes to the existing proof regarding the effectiveness of large language models in specialized fields, especially in healthcare, and paves the way for future developments in healthcare solutions using AI.

References

[1] Anicomanesh. (2023). Unleashing the Power of Mistral-7B: Efficient Fine-Tuning for Medical QA. *Medium*. Retrieved from https://medium.com/@anicomanesh/unleashing-the-power-of-mistral-7b-efficient-fine-tunin

[2] Ruslanmv. (2023). AI Medical Chatbot Dataset. *Hugging Face Datasets*. Retrieved from https://huggingface.co/datasets/ruslanmv/ai-medical-chatbot

[3] Hugging Face. (2023). *GenerationMixin*. Retrieved from https://huggingface.co/

docs/transformers/v4.46.0/en/main_classes/text_generation#transformers.GenerationMixin

[4] Hugging Face. (2023). *Model*. Retrieved from https://huggingface.co/docs/transformers/v4.46.0/en/main_classes/model