Medical Question Answering Project Report

# 1. Dataset Loading and Preprocessing

We used a medical QA dataset, tokenized the inputs and labels using a pre-trained tokenizer, and split the dataset into training, validation, and test sets in an 80/10/10 ratio. This ensures effective training and fair evaluation.

# 2. Model Training

We fine-tuned a pre-trained transformer model using the HuggingFace Trainer API. The model was trained using the following configuration:  
- Learning rate: 2e-5  
- Batch size: 4  
- Epochs: 10  
- Weight decay: 0.01

# 3. Hyperparameter Optimization

We tested the model with three different hyperparameter configurations. The selected configuration provided the best tradeoff between training time and evaluation metrics. ROUGE scores were used for evaluation.

# 4. Evaluation

Evaluation was conducted on validation and test sets. ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum) were calculated to assess the quality of model predictions. Additionally, qualitative evaluation was done by generating responses to sample questions.

# 5. Deployment Interface

A Gradio interface was created to allow users to enter medical questions and receive AI-generated answers. The interface provides an interactive experience for testing the model.

# 6. Qualitative Results

Examples of both accurate and inaccurate model responses were documented. The model performs well for diseases present in the dataset, but struggles with those not seen during training.

# 7. ROUGE Scores Output

ROUGE Scores (on test set):  
ROUGE-1 F1: 0.3940  
ROUGE-2 F1: 0.2526  
ROUGE-L F1: 0.3475  
ROUGE-Lsum F1: 0.3476

# 8. Conclusion

The fine-tuned QA model performs reasonably well for known medical terms. It shows promise in real-world deployment with proper dataset extension and fine-tuning. Hyperparameter tuning and evaluation metrics helped in optimizing performance. Final deployment via Gradio provides an accessible user experience.