Fine-Tuning a Large Language Model for Medical Question Answering

Presented by: Atharva | Prompt Engineering



Project Overview

Goal

Create specialized medical QA system using Flan-T5-base

Dataset

13 curated medical QA pairs covering diverse topics

Model

Google's Flan-T5-base with 250M parameters

Rationale

Small, high-quality data + instruction-tuned foundation model



Implementation Approach

Dataset Preparation

Formatted, structured medical QA pairs ready for fine-tuning

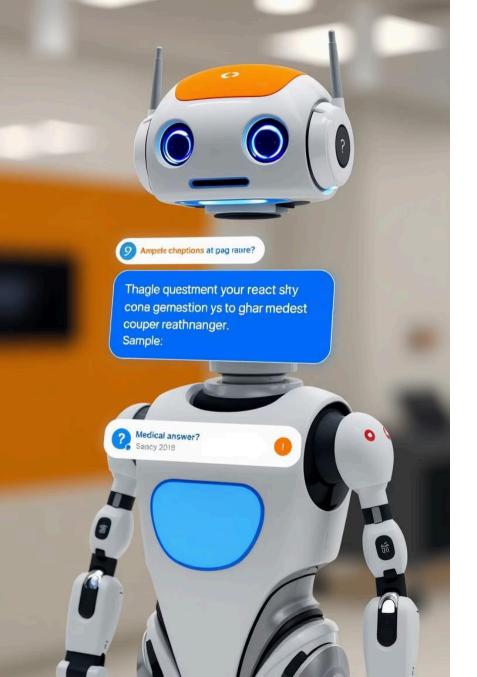
Training Configuration

• Learning rate: 1e-4

Batch size: 4

Weight decay: 0.01

• Epochs: 10



Model Evaluation & Results

Qualitative Testing

Expert review of generated answers for accuracy

Example Q&A

Clinical questions with precise model responses

Limitations

Some inaccuracies & domain-specific gaps observed

Evaluation Style

Informal expert assessment without formal metrics

Inference Pipeline

Process

- Input question preprocessing
- Model inference
- Answer generation & output

Generation Parameters

- Temperature: controls randomness
- Beam search: improves answer quality

Interactive Interface

Screenshot showing user query input and generated reply

Limitations & Future Directions

Current Limitations

Small dataset, single hyperparameters, qualitative evaluation

Proposed Improvements

Expand dataset, add formal metrics, tune hyperparameters

Lessons Learned

Importance of data quality and careful evaluation



Demo & Results Summary

Demo Overview

Shows model response generation to medical queries

Achievements

Functional specialized medical QA system using Flan-T5

Conclusion

Promising approach for medical LLM applications in education and research



Key Takeaways & Next Steps

Successful Fine-Tuning

Instruction-tuned Flan-T5 adapts well to medical QA

Data Quality is Crucial

High-quality curated pairs maximize performance

Future Work

3

4

Broader datasets, automated metrics, interface enhancements

Potential Impact

Assist medical education, reduce clinician workload