Enhancing Medical Question-Answering with Qwen3-0.6B Using the MedQuAD Dataset

Project README Documentation

Project Overview

This project explores the fine-tuning of **Qwen3-0.6B**, a lightweight Large Language Model (LLM), for **medical question answering (QA)** using the **MedQuAD dataset**, containing 16,000+ expert-verified QA pairs from NIH and MedlinePlus. The study evaluates whether small LLMs can provide accurate, interpretable, and computationally efficient medical responses using both quantitative metrics and human evaluation.

Research Objectives

- Fine-tune Qwen3-0.6B using MedQuAD data.
- Evaluate answer quality using BLEU, ROUGE-L, F1-score, Exact-Match, METEOR, and Perplexity.
- Reduce computational overhead using LoRA and quantization.
- Demonstrate feasibility for AI-based healthcare support systems.

Keywords

Medical QA, Qwen3-0.6B, MedQuAD, LLMs, Healthcare AI, NLP, LoRA, Fine-Tuning, Knowledge-Based QA.

Dataset Description

Dataset: MedQuAD (Medical QA Dataset)
Size: 16,000+ QA pairs
Source: NIH, MedlinePlus, National Library of Medicine
Format: CSV (Question, Answer, URL, Category)

Project Workflow

1. **EDA** – Explore and preprocess the dataset.
2. **Model Fine-Tuning** – Run `model_fine_tuning` to fine-tune Qwen3-0.6B; output: `my-qwen-model`.

3. **Model Evaluation** – Metrics: BLEU, ROUGE-L, F1-score, Exact-Match, METEOR, Perplexity, Human Evaluation.
4. **Model Inference** – Use `model_inference.ipynb` for interactive QA.

Framework & Implementation Overview

Model	Qwen3-0.6B
Fine-Tuning	LoRA (PEFT)
Optimization	DeepSpeed, Accelerate
Quantization	BitsAndBytes (4-bit)
Environment	Google Colab (GPU)
Evaluation Metrics	BLEU, ROUGE-L, F1, Exact-Match, METEOR, Perplexity, Human evaluation

Advantages of the Model

- Low computational resources.
- Strong QA performance post fine-tuning.
- Accurate and interpretable medical answers.
- Suitable for chatbots and healthcare systems.

Evaluation Metrics

BLEU	Measures linguistic accuracy and similarity.
ROUGE-L	Measures contextual overlap and meaning retention.
F1 Score	Balances precision and recall for QA.
Exact-Match	Checks if generated answer exactly matches the reference.
METEOR	Measures semantic similarity using synonyms and paraphrases.
Perplexity	Evaluates next-token prediction (lower is better).
Human Evaluation	Validates medical accuracy and clarity.

Model Testing Instructions

1. Download the repository and upload to Google Drive.
2. Open `model_inference.ipynb` in Google Colab.
3. Connect to a runtime environment (GPU recommended, CPU slower).
4. Run first cell to connect Google Drive; access fine-tuned model `my-qwen-model`.
5. Run sequentially until the question input cell; input your question. CPU inference 15 min.
6. Base model accessed via Hugging Face API; no local file needed.

Key Findings

- Fine-tuned model `my-qwen-model` produced contextually accurate medical answers.
- LoRA and quantization optimized training efficiency.
- Lightweight LLMs suitable for healthcare QA.
- Ready for deployment in medical information systems.

Project File Structure

- `medDataset_processed.csv` – Processed MedQuAD dataset.
- `ds_config.json` – DeepSpeed configuration.
- `model_fine_tuning.ipynb` – Scripts to fine-tune base Qwen3-0.6B.
- `my-qwen-model/` – Fine-tuned model output.
- `model_inference.ipynb` – Notebook for inference.
- `research_report.pdf` – Project methodology, results, and analysis.
- `README.md` – This documentation.

Future Enhancements

- Integrate into chatbots.
- Multilingual QA.
- Hallucination detection for sensitive answers.
- Expand domain using PubMed and clinical literature.

License

Released under **MIT License** for academic and research use.

Acknowledgements

Open-source resources from Hugging Face Transformers, MedQuAD, Qwen (Alibaba Cloud), NIH, and MedlinePlus.