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**, which contains over 16,000 expert-verified medical QA pairs sourced from trusted institutions such as the **NIH** and **MedlinePlus**.

The research evaluates whether small LLMs, when fine-tuned efficiently, can provide accurate, interpretable, and computationally effective medical responses. The evaluation includes both quantitative metrics and qualitative assessments of contextual relevance and factual accuracy.

Research Objectives

- Fine-tune Qwen3-0.6B using MedQuAD medical QA 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 real-world AI-based healthcare support systems.

Keywords

Medical Question Answering, Qwen3-0.6B, MedQuAD, Large Language Models, Healthcare AI, Natural Language Processing, LoRA, Medical NLP, Fine-Tuning, Knowledge-Based QA.

Dataset Description

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

Project Workflow

The project follows a structured workflow:

1. **Exploratory Data Analysis (EDA)** – Understanding the dataset and preparing it for training.
2. **Model Fine-Tuning** – Using the `model_fine_tuning` script to fine-tune the base Qwen3-0.6B model. This produces the fine-tuned model: `my-qwen-model`.
3. **Model Evaluation** – Evaluating the fine-tuned model using metrics such as BLEU, ROUGE-L, F1-score, Exact-Match, METEOR, Perplexity, and human evaluation.
4. **Model Inference** – Using the fine-tuned model to answer medical questions interactively.

Framework & Implementation Overview

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

Advantages of the Model

- Requires low computational resources.
- Strong performance after fine-tuning on medical QA tasks.
- Generates accurate and interpretable medical responses.
- Suitable for medical chatbots and healthcare support 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 synonymy and paraphrase matching.
Perplexity	Evaluates how well the model predicts the next token (lower is better).
Human Evaluation	Validates medical accuracy and clarity.

Model Testing Instructions

To test the model:

1. Download the repository and upload it to Google Drive.
2. Open the `model_inference.ipynb` notebook in Google Colab.
3. Connect to a runtime environment (GPU recommended, CPU works but slower).
4. Run the first code cell to connect Google Drive to Colab. This allows access to all necessary files, including the fine-tuned model `my-qwen-model`.
5. Run the notebook sequentially until reaching the cell that asks for your question input. Enter the question to get an answer.
   Note: This step may take around 15 minutes on CPU.
6. The base model is accessed through the Hugging Face API; no local file is needed for it.

Key Findings

- Fine-tuned Qwen3-0.6B (`my-qwen-model`) produced contextually accurate medical answers.
- LoRA and quantization made training efficient and resource-friendly.
- Lightweight LLMs are suitable for domain-specific applications like healthcare QA.
- Potential for real-world deployment in medical information systems.

Project File Structure

```
medDataset_processed.csv       # MedQuAD formatted dataset
ds_config.json                 # DeepSpeed configuration file
qwen-medquad-finetuned/        # Fine-tuned model outputs (my-qwen-model)
model_fine_tuning/             # Script to fine-tune base model (outputs my-qwen-mode
research_report.pdf            # Full project analysis
model_inference.ipynb          # Notebook for running inference
README.md                      # Project documentation
```

Future Enhancements

- Integration into healthcare chatbots.
- Multilingual medical question answering.
- Hallucination detection for sensitive medical responses.
- Expand domain using PubMed and clinical literature.

License

This project is released under the **MIT License** and is free for academic and research use.

Acknowledgements

This research uses open-source resources from:

- Hugging Face Transformers
- MedQuAD Dataset
- Qwen by Alibaba Cloud
- NIH and MedlinePlus medical resources