Finetune LLMs for Disease Diagnosis

This coding test assesses the ability to fine-tune large language models (LLMs) efficiently and effectively. The task is to generate structured disease labels based on free-text radiology findings. You will be provided with a dataset of 1,236 training and 236 testing examples.

Your submission will include both a baseline using prompt engineering and a fine-tuned LLM model. You'll also design appropriate evaluation metrics and analyze performance and failure cases.

Task 1. Build zero-shot baseline using prompt engineering

Please prompt QWen3 (4B and 8B) to generate the disease labels from findings. Describe how you design the prompt to maximize the baseline performance

- Input: findings (column `input_finding`)

> The liver is normal in size and shape with homogeneous density. A patchy low-density lesion is seen around the liver fissure. The intrahepatic duct system is not obviously dilated, and the course is normal. The gallbladder is not enlarged, with no obvious thickening of the wall, and no clearly abnormal density foci are seen inside. The spleen is normal in size and shape with homogeneous density, and some punctate low-density foci are seen inside. The pancreas is normal in size and shape with homogeneous density, and no clearly abnormal density foci are seen inside. The main pancreatic duct is not obviously dilated, and the peripancreatic fat space is clear. Both kidneys are normal in size and shape with homogeneous density. A round low-density lesion is seen in the right kidney with a diameter of about 16mm. The left adrenal gland is thickened, and a punctate high-density lesion is seen in the right adrenal gland. The renal pelvis-calyx system is not obviously dilated. The perirenal fat space is clear, and no clearly abnormal density foci are seen. No enlarged lymph nodes are seen in the retroperitoneum.

- Expected output: disease labels (column `output_disease`)

> Renal cyst, Adrenal hyperplasia, Adrenal calcification

Please also design a metric to compare model ouput and ground truth labels.

Task 2. Efficient Model Fine-tuning

Please fine-tune the well-known QWen3 on the provided training set for disease diagnosis generation. Candidates can use parameter efficient fine-tuning methods (e.g., LoRA, DoRA) for better compute efficiency or fully fine-tuning.

Notes:

- The objective of this task is to test the applicants' ability of using advanced code repositories without costing too much compute. Thus, the task is designed to be done on freely provided compute resources, such as Google Colab. If you have access to better computing, please feel free to train the larger models.
- We recommend using 4B and 8B Qwen3 models.
- Please upload the fine-tuned model to huggingface

Submission

1. Technical Report (PDF)

The report should include:

- Prompt design and baseline performance.
- Fine-tuning methodology (architecture, hyperparameters, etc.).
- Description of the evaluation metric.
- Comparative results (baseline vs. fine-tuned for 4B and 8B models).
- Analysis of typical failure cases and their potential causes.

2. GitHub Repository

The repository should contain:

- Full codebase for data preprocessing, training, and inference.
- Evaluation scripts
- Download link to trained models
- An Excel file with all the inference results
- A well-documented README with setup and usage instructions. Here is a checklist

Evaluation Criteria

Your submission will be evaluated based on:

- Effectiveness: Accuracy of disease label prediction.
- Efficiency: Use of compute and techniques like LoRA/DoRA.
- Reproducibility: Completeness of code and documentation.
- Insight and Analysis: The depth and clarity of your written report, particularly in your interpretation of results, discussion of failure cases, and justification of design choices.

If you have any questions on the description, please reach out to `jun.ma2@uhn.ca` with the Email Subject: Your_Name- NLP Coding Test