Introduction

This project explores whether large language models (LLMs) can support medical information extraction from radiology findings. The dataset contains 1,237 samples. The work was divided into two tasks: first testing baseline performance with prompts (Task 1), then trying parameter-efficient fine-tuning of Qwen models (Task 2). Because of GPU limits, some experiments did not run to completion, but the process still showed both potential and difficulties.

Task 1: Prompt Design and Baseline

For the baseline, I designed a straightforward prompt asking the model to extract only disease names. This gave mixed results. It could capture common diseases but often missed some or added hallucinated ones. On a small validation set, performance was roughly F1 ~0.44, which shows prompting alone is limited.

Task 2: Fine-Tuning Attempts

The plan was to fine-tune Qwen/Qwen2.5-7B-Instruct with LoRA and 4-bit quantization. Key settings: 2 epochs, batch size 2 with gradient accumulation, max sequence length 512. However, Colab T4 GPUs ran out of memory often. Even partial training steps were possible, but no full convergence. This highlighted the practical barriers of training large LLMs with limited compute.

Reflections

Baseline (zero-shot): F1 ~0.44. Fine-tuned: partial runs suggested some improvement, but results were inconclusive. The main lesson was the gap between theory and practice: LoRA is efficient, but still not trivial to run on small GPUs.

Conclusion

Prompting gives a simple baseline but lacks accuracy. Fine-tuning has potential but was not fully achievable under resource limits. Future work should include testing on stronger GPUs, trying smaller models, or exploring hybrid retrieval + LLM methods. Being realistic, the report shows both what was done and where it stopped.

Deliverables

- Notebooks: Task 1 (prompting) and Task 2 (fine-tuning setup) - Code: preprocessing and partial training scripts - Model: attempted LoRA adapter (incomplete) - Results: partial metrics and error cases