Task 3: PCTR: Photographed Chinese Table Reasoning (Optional)

Methodology

Introduction

The primary objective of this approach is to **extract text from images** of documents, tables, or forms, and subsequently perform **intelligent reasoning and response generation** using a large language model (LLM). The pipeline integrates **PaddleOCR**, an optical character recognition tool, with **Qwen LLM**, a transformer-based language model fine-tuned with LoRA for domain-specific text comprehension. The methodology can be divided into **three main stages**: data acquisition and preprocessing, OCR-based text extraction, and LLM-based reasoning.

Stage 1: Preprocessing and setting up PaddleOCR

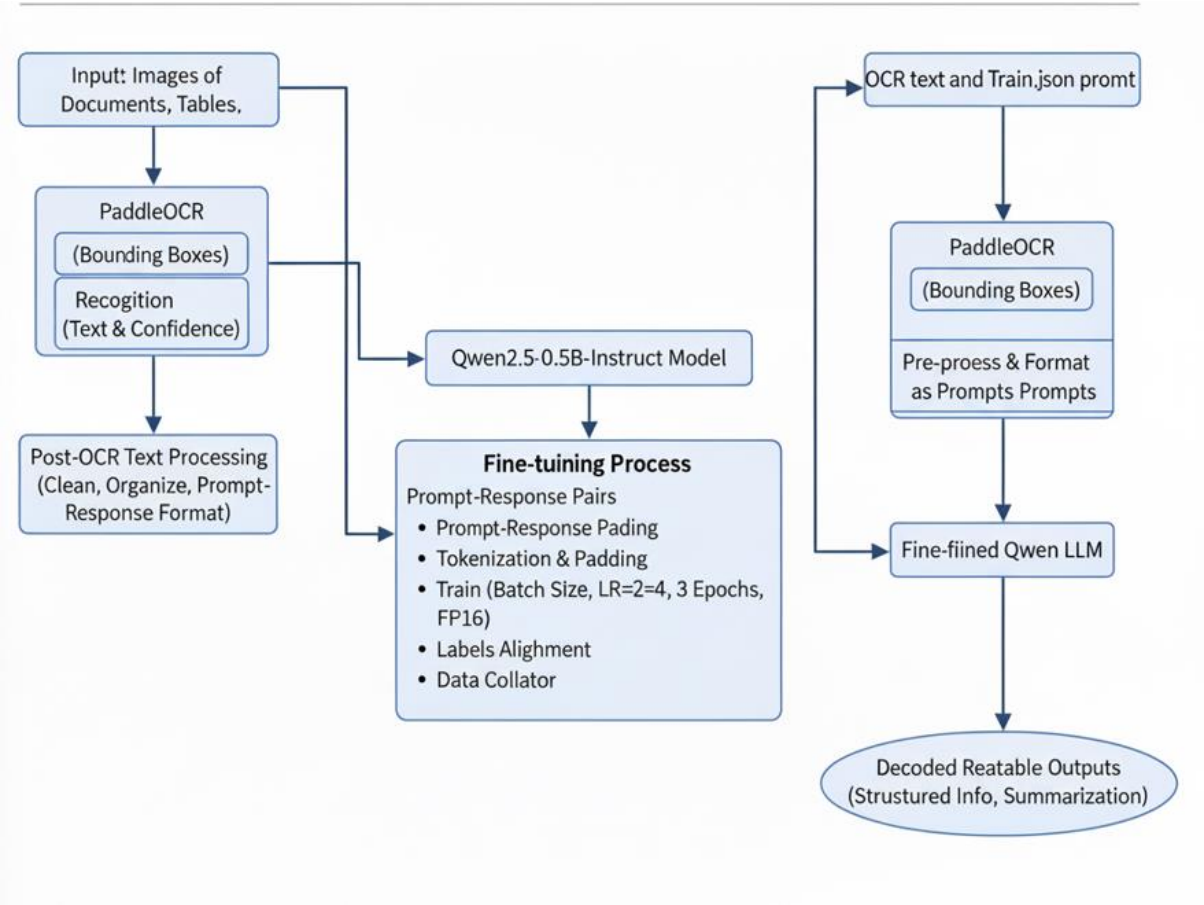
In the OCR stage, **PaddleOCR** was employed to detect and recognize text within the images. The detection module identified bounding boxes around text regions, while the recognition module converted these regions into string outputs along with confidence scores. Post-OCR processing involved cleaning the extracted text to remove noise, special characters, and excess whitespace, and organizing the text into a structured prompt-response format suitable for model training.

Stage 2: Fine tuning QWEN

The next stage involved fine-tuning the Qwen2.5-0.5B-Instruct model using the extracted text. Qwen was chosen for its instruction-following capabilities and compatibility with parameter-efficient fine-tuning through LoRA. The OCR outputs were transformed into prompt-response pairs, where each prompt contained the extracted text and the response contained the expected answer or reasoning. These pairs were tokenized, padded, and truncated to a consistent sequence length to ensure uniform model input. LoRA was applied to specific projection layers in the model to allow efficient adaptation to the task without retraining the entire network. Training was performed using a small batch size with gradient accumulation, a learning rate of 2e-4, three

epochs, and mixed-precision (fp16) for efficient GPU usage. Labels were aligned with input tokens to enable causal language modeling, and a specialized data collator ensured proper handling of padding and labels during batch processing.

Flowchart and Interference



During inference, new images were passed through PaddleOCR to extract text, which was then pre-processed and formatted as prompts for the fine-tuned Qwen LLM. The model generated responses or reasoning based on the extracted text, which were then decoded into readable outputs. This approach enabled the extraction of structured information, automated reasoning, and summarization from complex documents.

Conclusion

Overall, this methodology effectively combined vision-based text extraction and instruction-based language understanding, providing a scalable and efficient framework for automated document analysis. The integration of

PaddleOCR ensured high-quality text recognition, while LoRA fine-tuning of Qwen allowed the model to generate contextually accurate and domain-specific responses. While OCR errors and GPU limitations pose challenges, the proposed pipeline demonstrates significant potential for tasks such as table reasoning, report generation, and multimodal document comprehension.