INVOICE INFORMATION EXTRACTION

NLU Project

PROJECT IDEA

Our project aims to develop a fine-tuned Large Language Model (LLM) capable of extracting structured information from invoices in JSON format. We use a fine-tuned microsoft/phi-2 model, leveraging Low-Rank Adaptation (LoRA) to enhance model performance on this specific task.

The ultimate goal is to create a model that efficiently converts raw invoice text into structured data, making it highly useful for automated invoice processing.

DATASET INFORMATION

Dataset link: https://www.kaggle.com/datasets/ananthakrishnanpv/7000-invoice-images-with-json

Data Preprocessing:
We used 500 invoice samples, applied OCR to extract text, and paired each with its corresponding JSON annotation. These were merged into a unified JSON format (input-output pairs) for model fine-tuning.

BASE MODEL INFORMATION

- Model: microsoft/phi-2
- Architecture: Decoder-only Causal Language Model (LLM)
- Capabilities:
  - Strong performance on reasoning and language tasks
  - Lightweight and efficient for fine-tuning
  - Quantization: 4-bit quantized using BitsAndBytes to save memory during fine-tuning

FINE-TUNING WITH LORA

Low-Rank Adaptation (LoRA) is a technique that adds low-rank matrices to model weights, enabling fine-tuning with minimal parameter updates.

Injects trainable rank-decomposition matrices into transformer layers.
Only a small number of parameters are trained → lower compute & memory cost.

Quantization Settings:

- 4-bit, with double quantization
- Quant type: nf4, compute dtype: float16

LoRA Configuration:

- Rank (r): 8, Alpha: 32, Dropout: 0.05
- Applied to fc1 and fc2 layers only

Trainable Parameters:

- 6.5M trainable / 2.78B total parameters
- Only 0.24% of the model is fine-tuned

EVALUATION METHODS

- Metrics:
  - Overall Accuracy: Fraction of correctly extracted fields.
  - Exact Match Count: Full match between predicted and ground truth JSON fields.
  - Fields Found: How many fields were detected.

- Approach:
  - Parsed and flattened both expected and predicted JSON outputs
  - Matched keys and values for exactness
  - Used 5 validation samples for qualitative and quantitative comparison

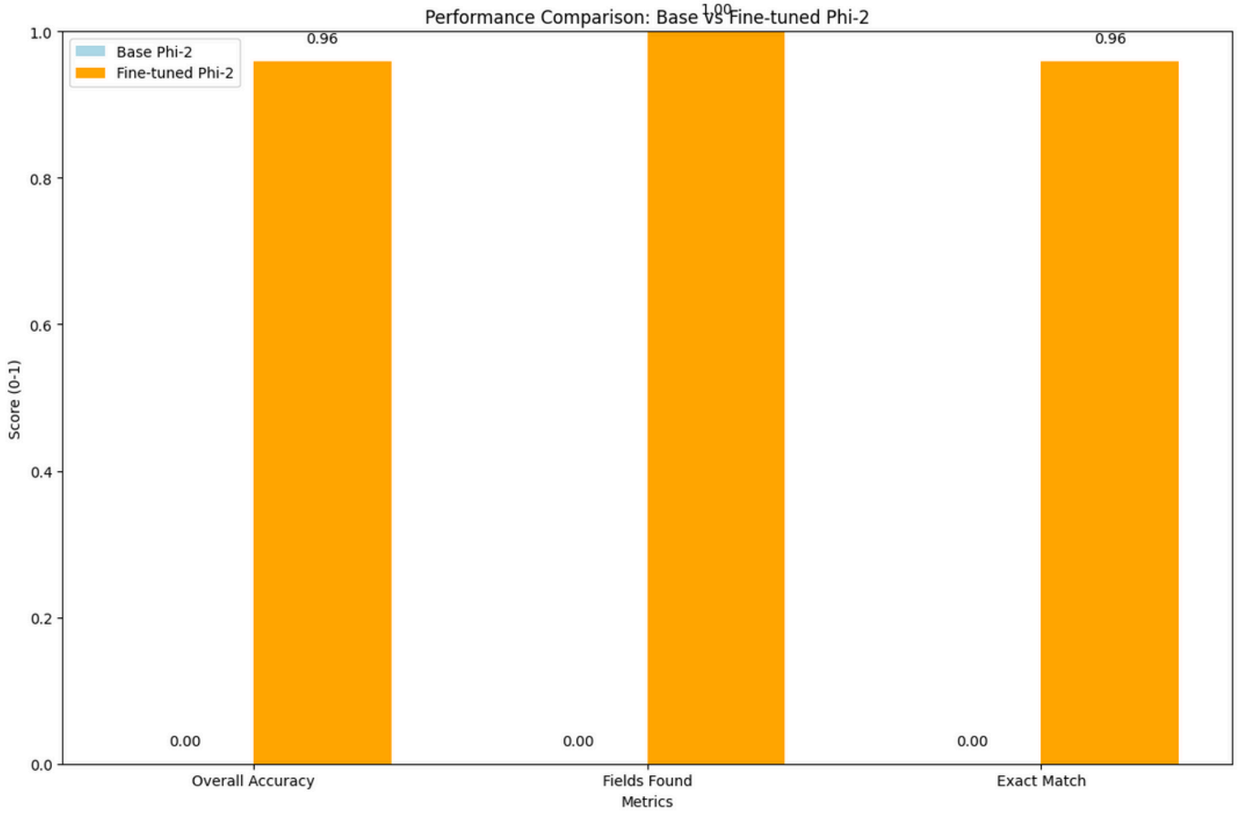
PROJECT RESULTS AND PLOTS

Base Phi-2 Model Results:

- Avg. Fields Found: 0.00
- Avg. Exact Match: 0.00
- Overall Accuracy: 0.00

Fine-tuned Phi-2 Results:

- Avg. Fields Found: 95.97%
- Avg. Exact Match: 100%
- Overall Accuracy: 95.97%



STRUCTURED OUTPUT

The model outputs structured data in JSON format, demonstrating its capability to accurately extract key invoice information.

Example:

Extracted JSON:

{ "buyer": { "address": "123 Client Street Clientville, CA 90210", "email": "john.smith@example.com", "name": "John Smith" },
"invoice": { "date": "April 15, 2025", "due_date": "May 15, 2025", "number": "INV-2025-051" }, "payment": { "method": "credit card", "net_amount": 30 },
"products": [{ "amount": 99.99, "description": "Cloud Storage: Premium tier", "quantity": 1, "price": 99.99 }, { "amount": 199.99, "description": "Technical Support: 24/7", "quantity": 1, "price": 199.99 }, { "amount": 1499.95, "description": "Software License: Enterprise", "quantity": 5, "price": 299.99 }],
"seller": { "address": "123 Innovation Drive Silicon Valley, CA 94025" },
"tax": { "amount_excluding_tax": 1,879.93, "amount_including_tax": 1,952.92, "tax_amount": 152.99 } }