Task 2: Receipt Digitization and Data Extraction

The essence of the problem

Develop a system that automatically processes receipt images and extracts text and key data.

Real-world document digitalization tasks require quickly and massively converting images into structured data. Documents can vary in quality, structure, and content.

System must :

- extract text from images ;
- find key fields ;
- work on stream documents ;
- provide a balance between accuracy and speed.

The challenge is not just to recognize text, but to build **a robust and efficient document processing pipeline** .

Input data

Dataset: SROIE

Formats:

- `.jpg , .png`

Documents can contain :

- noise ;
- slope ;
- various quality ;
- different structure .

Requirements for the solution

- OCR / VLM / hybrid ;
- HTTP API:

```
POST / scan_receipts
```

- batch processing ;
- automatic Job .

Format result

```
{
"results": [
{
" doc_id ": "",
"status": "ok",
" full_text ": "",
```

```
"fields": {
"company": "",
"date": "",
"total": "",
"address": ""
},
"metadata": {
"method": "",
" duration_ms ": 0
}
}
]
}
```

Criteria assessments

- Accuracy extraction - 40%
- Productivity - 30%
- Completeness processing - 15%
- Efficiency resources - 10%
- Justification solutions - 5%

Expected result

A system capable of quickly and accurately extracting data from receipt images and scalable to large volumes. Improvements that enhance the convenience and practical value of the solution are welcome.