資訊技術專案開發與實

綠色供應鏈管理平臺
測試報告智能辨識功能開發

第六組
張士宏、楊倢瑄
矽創電子股份有限公司

計畫目標

- **特定領域圖像辨識引擎**
  以OCR技術為基礎, 針對實驗室有害物質檢測報告, 解析特定表格、文字混合內容。
- **高效能處理**
  優化OCR辨識架構, 運用機器學習技術以及平行處理能力, 能達成解析全文並完成相關內容解析工作。
- **高準確度**
  通過自然語言處理及深度學習, 力求達到一定程度以上的文字識別準確率, 資料處理錯誤率控制越低越好。
- **自動化後處理**
  系統性遞迴更新減少人工修正處理, 提高資料處理效率。
- **系統管理機制**
  建立辨識結果驗證和錯誤報告系統, 支援系統持續優化。

上學期成果

團隊人數：2

每周花費時間：約12小時

- **特定領域圖像辨識引擎**

  以YOLO、Tesseract及小語言模型，解決方案目前已完成初版，其他方案待評估。

- **高效能處理**

  優化OCR辨識架構，目前已完成初版。

- **高準確度**

  預計第二階段完成。

- **自動化後處理**

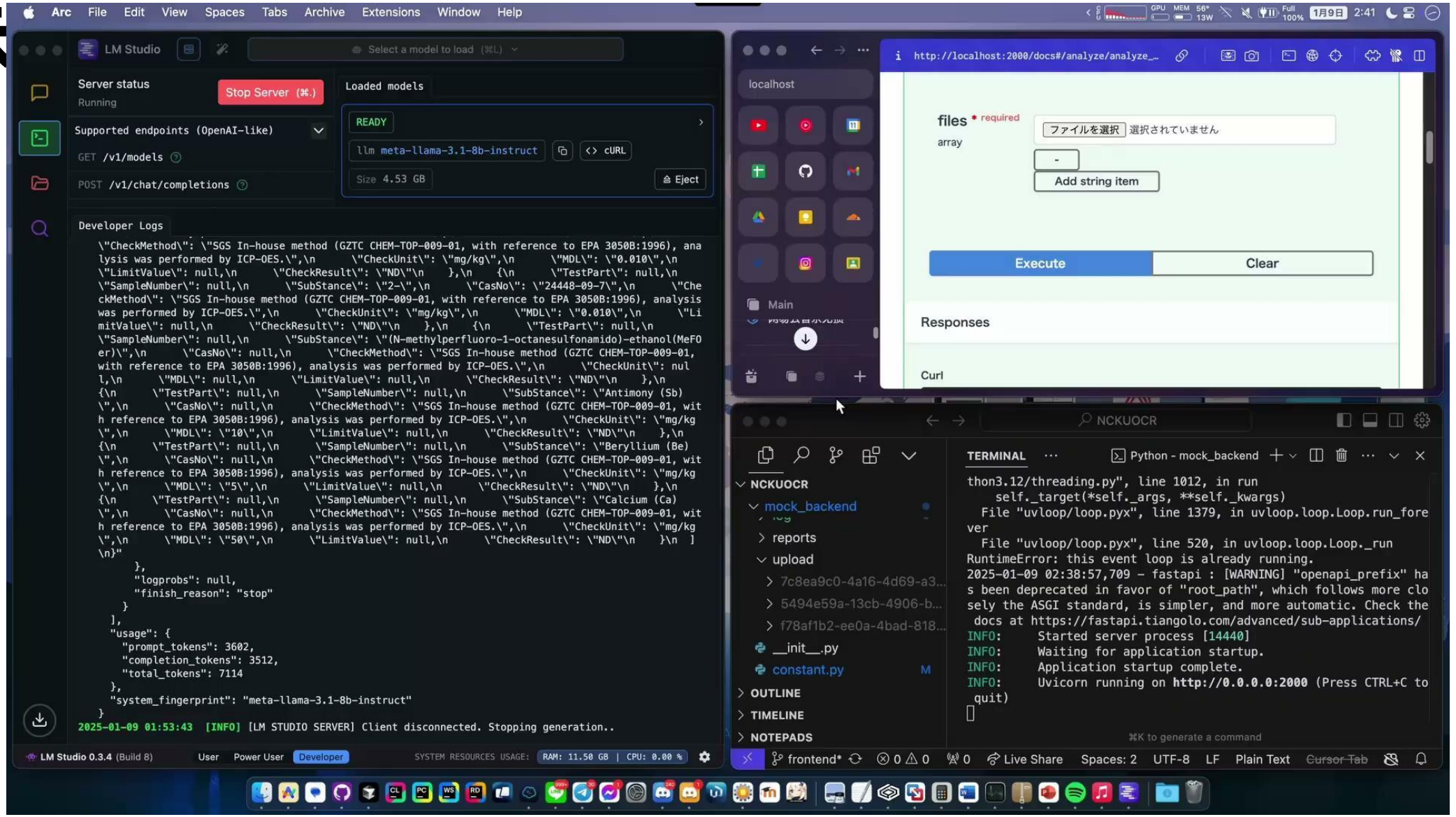
  預計第二階段完成。

- **系統管理機制**

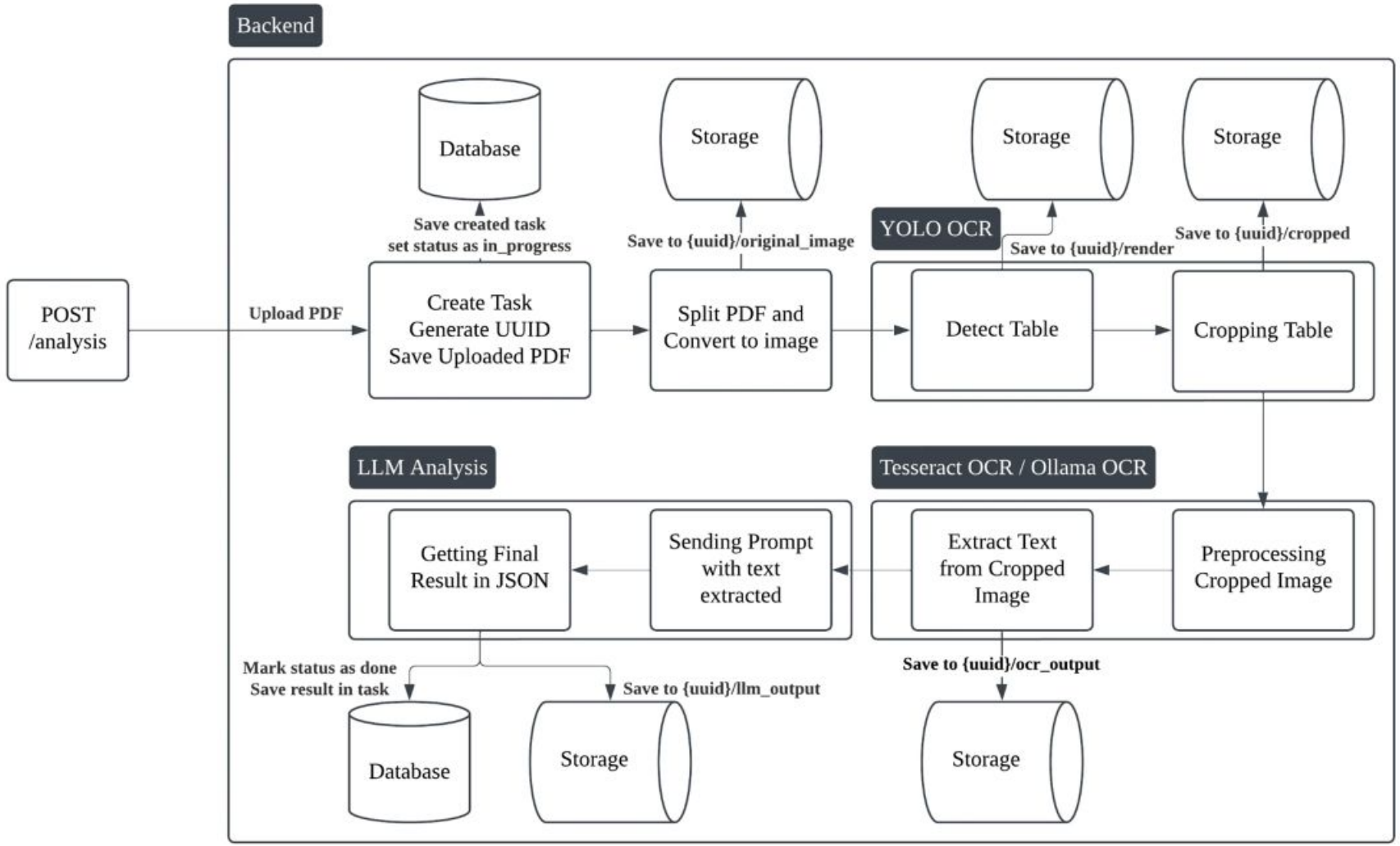
  建立辨識結果驗證，目前已完成。

  錯誤報告系統、支援系統持續優化預計第二階段完成。

成果影片展示


Arc File Edit View Spaces Tabs Archive Extensions Window Help
LM Studio
Select a model to load (⌘L)
Server status
Running
Stop Server (⌘.)
Loaded models
READY
llm meta-llama-3.1-8b-instruct
cURL
Size 4.53 GB
Eject
Supported endpoints (OpenAI-like)
GET /v1/models
POST /v1/chat/completions
Developer Logs
2025-01-09 01:53:43 [INFO] [LM STUDIO SERVER] Client disconnected. Stopping generation..
LM Studio 0.3.4 (Build 8)
User
Power User
Developer
SYSTEM RESOURCES USAGE:
RAM: 11.50 GB | CPU: 0.00 %
http://localhost:2000/docs#/analyze/analyze_...
localhost
files * required
array
ファイルを選択
選択されていません
Add string item
Execute
Clear
Responses
Curl
Main
NCKUOCR
mock_backend
reports
upload
__init__.py
constant.py
OUTLINE
TIMELINE
NOTEPADS
TERMINAL
Python - mock_backend
RuntimeError: this event loop is already running.
INFO: Started server process [14440]
INFO: Waiting for application startup.
INFO: Application startup complete.
INFO: Uvicorn running on http://0.0.0.0:2000 (Press CTRL+C to quit)
frontend*
Live Share
Spaces: 2
UTF-8
LF
Plain Text

上學期後端 Endpoint /analyze 架構分析



總耗時: 116.286s (1min 56.286sec)

效能標的: 7頁 SGS 檢驗報告 PDF 檔案 (CANEC2105310907.PDF)

下學期成果

團隊人數：2

每周花費時間：約12小時

- **優化前端**

  加入前端功能，可實際運用。

- **特定領域圖像辨識引擎**

  ML方案，以RapidOCR及小語言模型。

  表格解析與讀取方案，文件分類，以YOLO及RapidOCR。

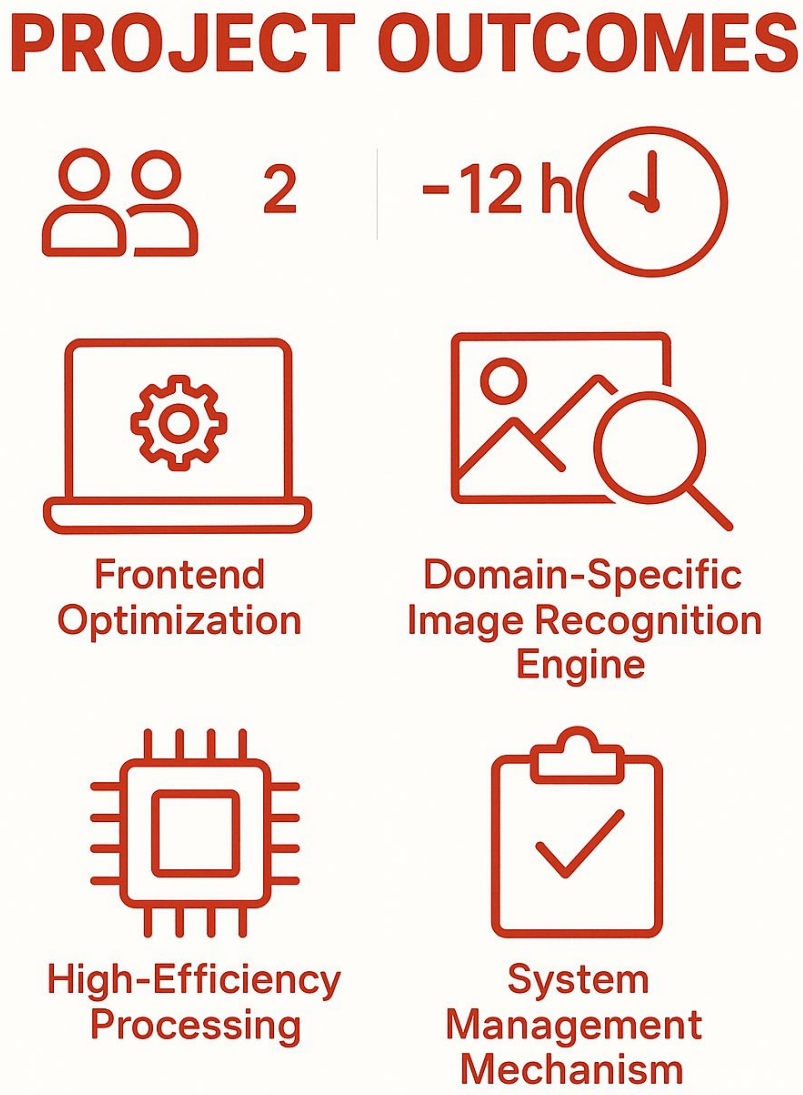
- **高效能處理**

  優化OCR辨識架構。

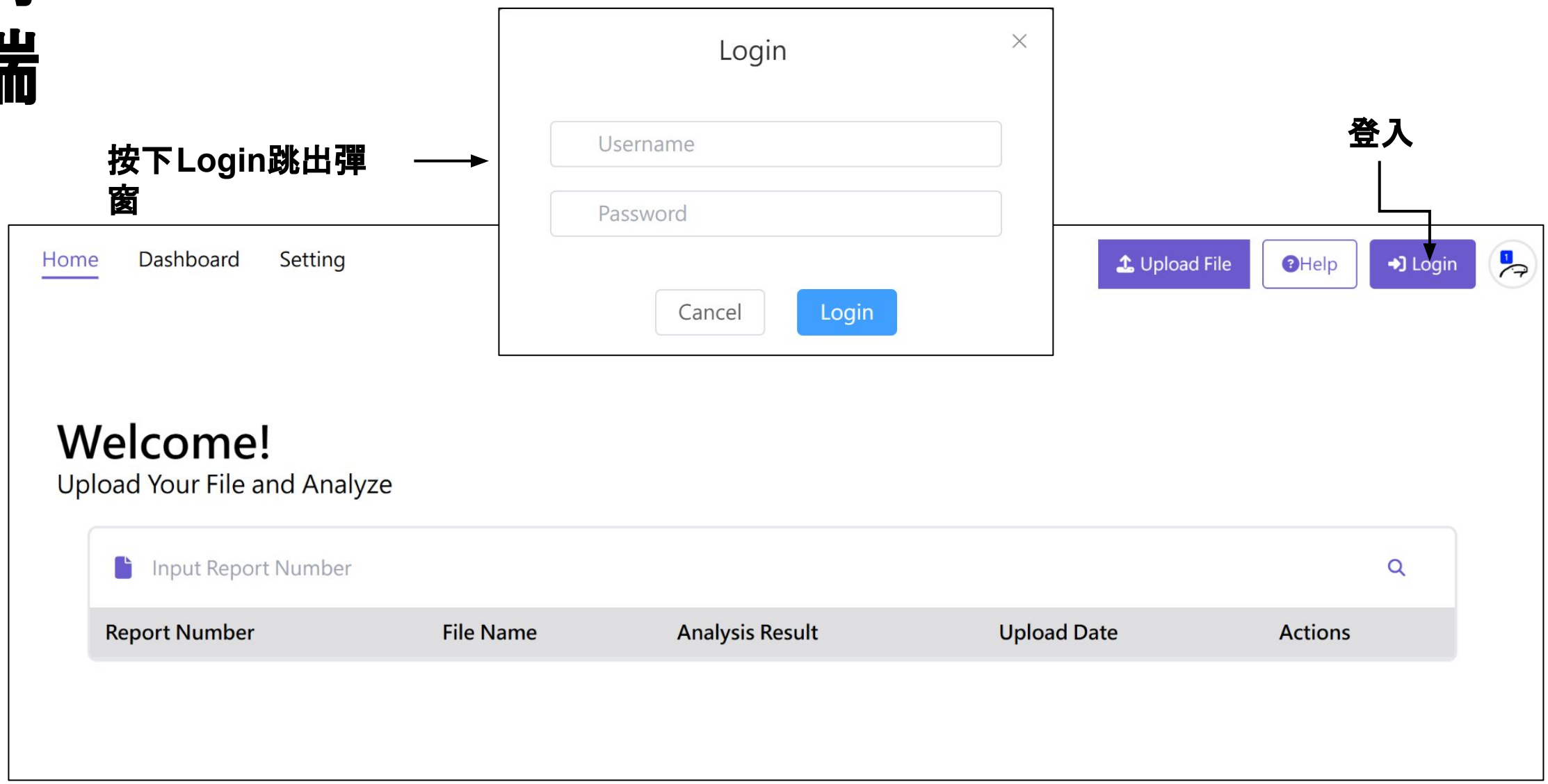
- **系統管理機制**

  建立辨識結果驗證，目前已完成。

  錯誤報告系統、支援系統持續優化。

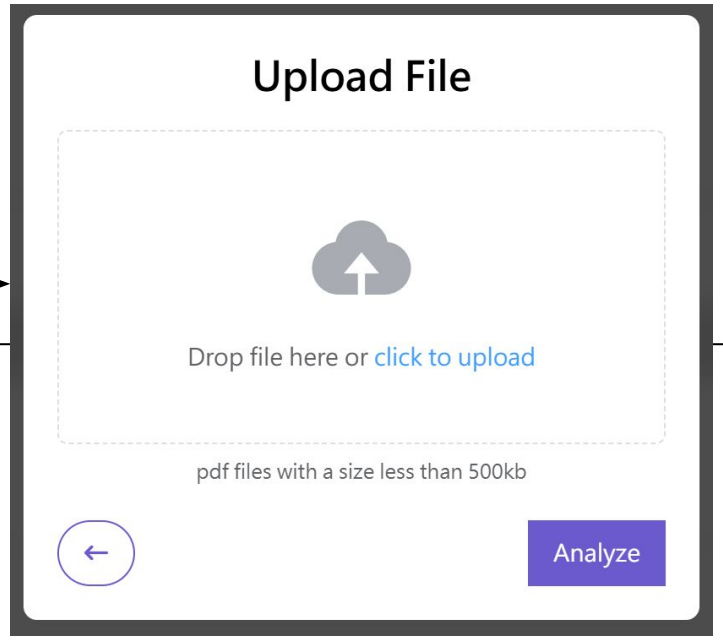


前端


按下Login跳出彈窗
登入
Login
Username
Password
Cancel
Login
Home
Dashboard
Setting
Upload File
Help
Login
Welcome!
Upload Your File and Analyze
Input Report Number
Report Number
File Name
Analysis Result
Upload Date
Actions

前端

按下upload file 跳出彈窗



上傳檔案

Home Dashboard Setting

Upload File Help Logout

Welcome!

Upload Your File and Analyze

輸入report number查詢

Input Report Number

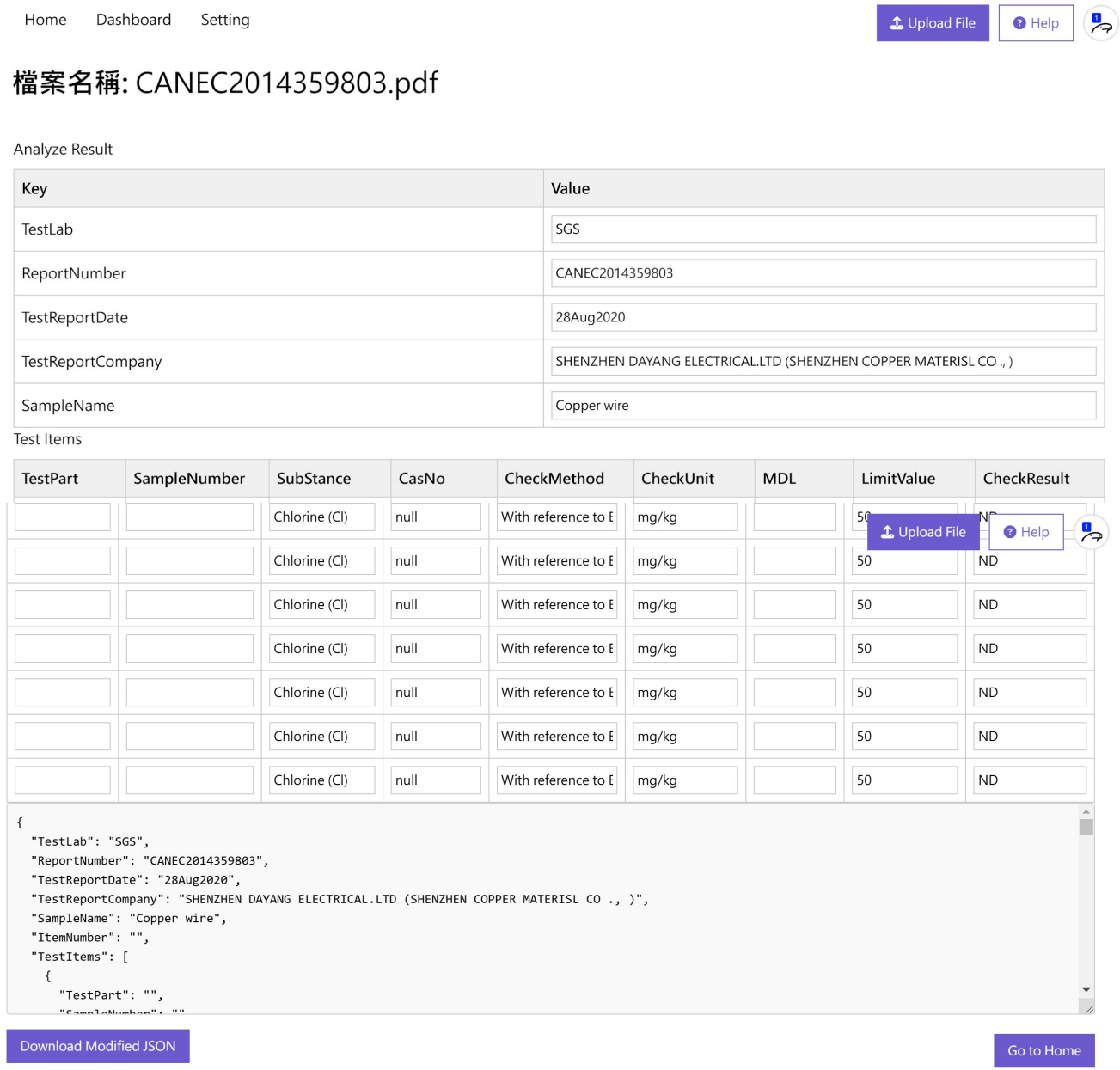
Report Number	File Name	Analysis Result	Upload Date	Actions
R069520	T05.jpg	Success	2025/5/9	Download Review
R029074	T04.jpg	Success	2025/5/9	Download Review
R010667	T02.jpg	Success	2025/5/9	Download Review
R970648	T01.jpg	Success	2025/5/9	Download Review

下載或檢視先前檔案

前端調用API頁面

以表格呈現分析結果並且
可以修正錯誤數 值

Json與表格內容同步



後端｜文件分析

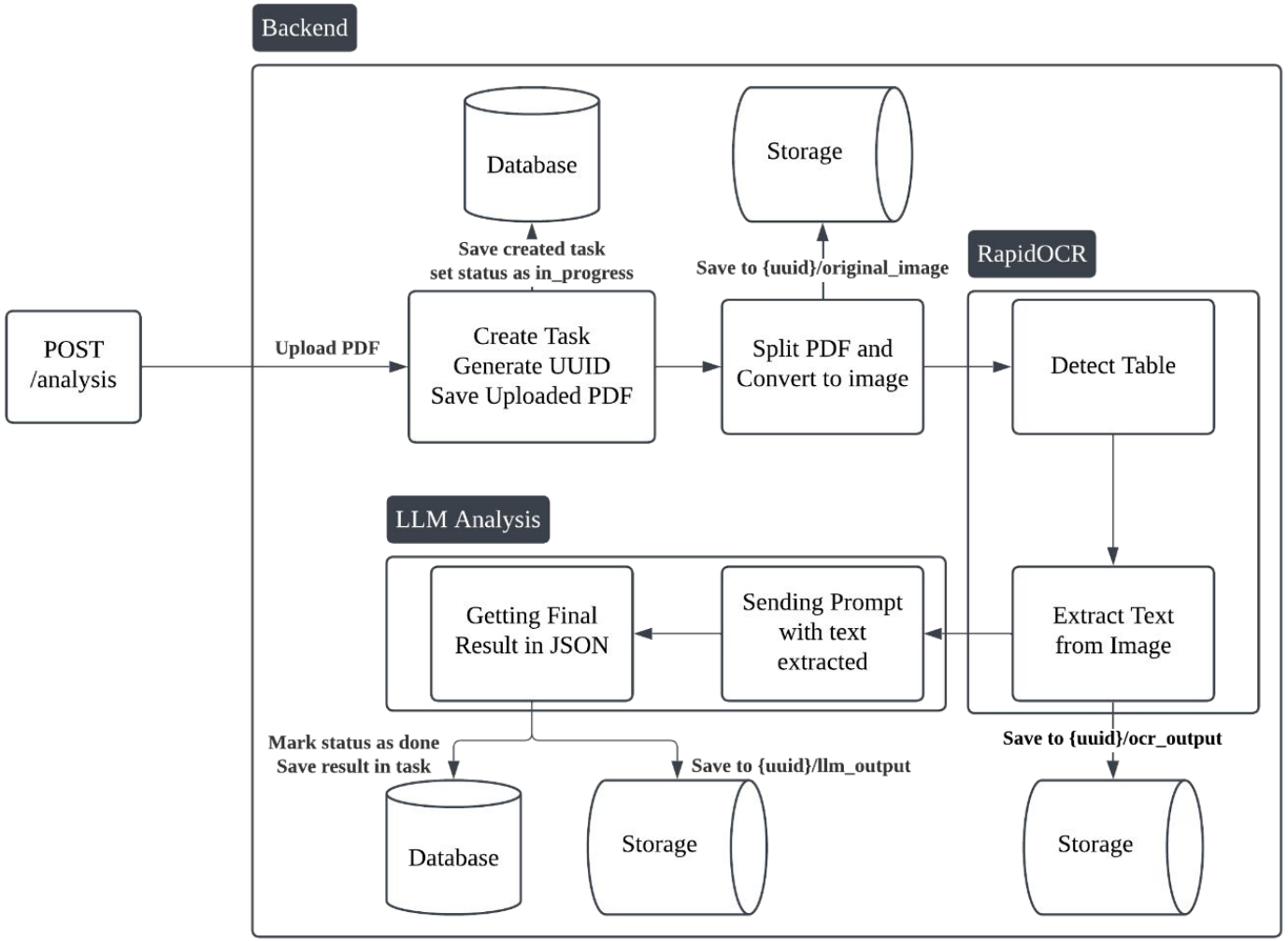
本次實作有兩個方案

MACHINE LEARNING | **表格解析與讀取**

後端

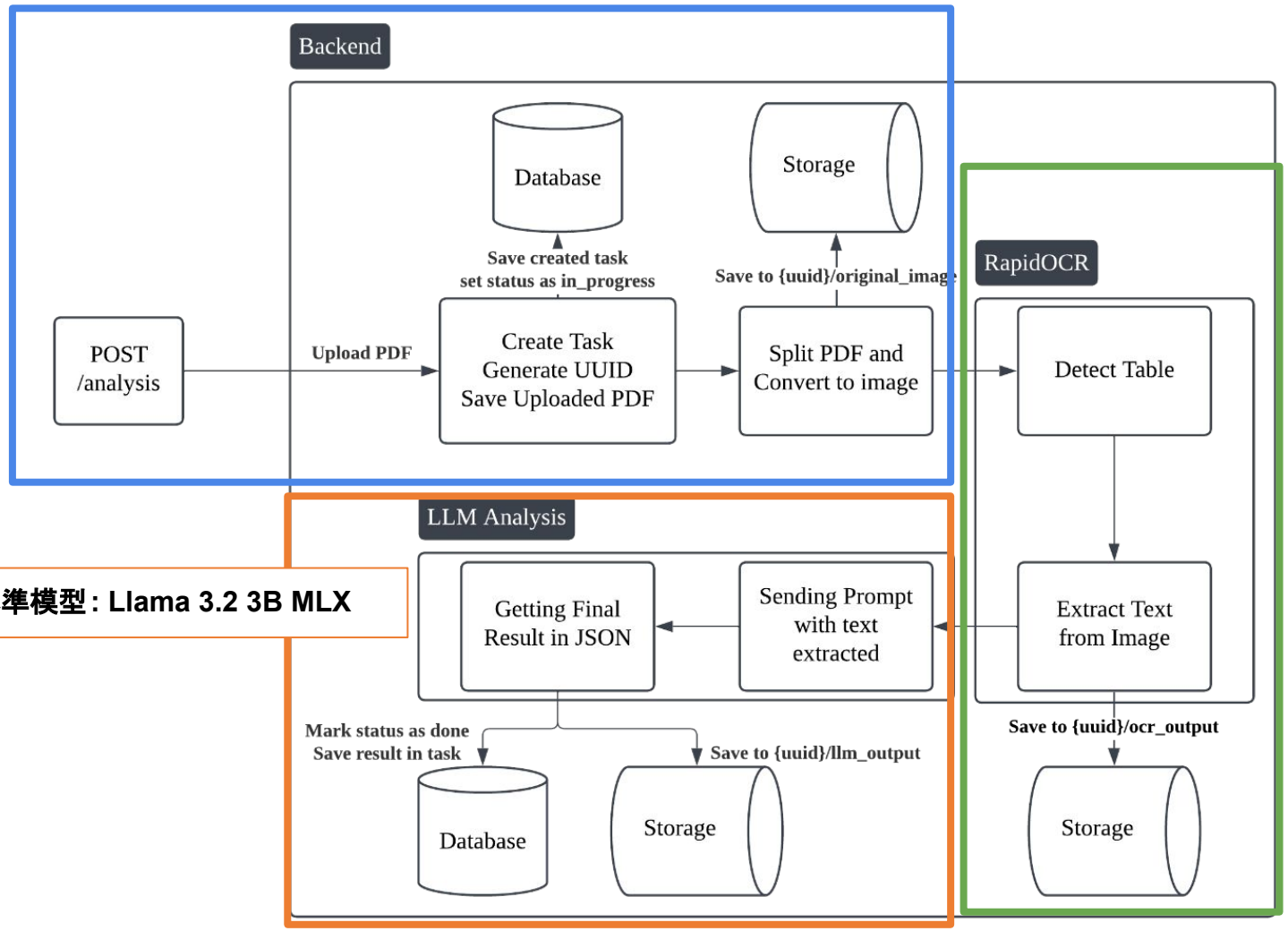
	方案不同之處	總耗時	精確度
上學期	利用YOLO、Tesseract及小語言模型	116.286s	82.6%
MACHINE LEARNING	RapidOCR及小語言模型 要求 LLM 運用截取文字，放入指定結構	31.8s	78.4%
表格解析與讀取	YOLO及RapidOCR 先用YOLO切出表格再OCR 直接使用OCR結果(HTML資料)填入結構	7.123s	62.7%

文件分析 | MACHINE LEARNING



效能標的 : 7頁 SGS 檢驗報告 PDF 檔案 (CANEC2105310907.PDF)

文件分析 | MACHINE LEARNING



使用者上傳 PDF
儲存 PDF 檔案
將PDF分頁並將分頁儲存為照片
耗時 0.593s

使用 RapidOCR
自動辨識及 OCR 表格內容
耗時 7.775s (處理 7 張照片)~ 1.11s/張

用調整後 Prompt
指示 LLM 擷取文字
輸出 JSON 分析結果
耗時 22.687s

基準模型 : Llama 3.2 3B MLX

總耗時 : 31.05sec

效能標的 : 7頁 SGS 檢驗報告 PDF 檔案 (CANEC2105310907.PDF)

文件分析 | MACHINE LEARNING

文件標示表格，圖片轉文字

Test Method : With reference to EN 14582:2016, analysis was performed by IC.

Test Item(s)	Unit	MDL	*001*
Fluorine (F)	mg/kg	50	ND
Chlorine (Cl)	mg/kg	50	ND
Bromine (Br)	mg/kg	50	ND
Iodine (I)	mg/kg	50	ND

Elementary Analysis

RapidOCR 內部自動判斷，提取

Test Item(s) Unit MDL 001 Fluorine (F) mg/kg 50 ND Chlorine (Cl) mg/kg 50 ND Bromine (Br) mg/kg 50 ND Iodine (I) mg/kg 50 ND

文件分析 | MACHINE LEARNING

LLM 資料分析及後處理

為了使 LLM 可以一次分析，並避免前後結果落差
送給 LLM 資料有經過以下處理

cropped_0_page_2.txt
cropped_1_page_5.txt
…
…

合併

cropped_0_page_2.txt
Halogen
Test Method: With reference to EN 14582:2016,
analysis was performed by IC.
Test Item(s) Unit MDL 001
Fluorine (F) mg/kg 50 ND

cropped_1_page_5.txt
Test Item(s) CAS NO. Unit MDL 001
Perfluorooctanoic acid (PFOA) and its salts+ 335-67-1
mg/kg 0.010 ND
….

prompt_data.txt

文件分析 | MACHINE LEARNING

Prompt Optimizing

```
IMPORTANT: DO NOT GENERATE CODE. EXTRACT DATA DIRECTLY INTO JSON FORMAT.
DO NOT ADD ANY INTRODUCTORY TEXT LIKE "Here is the extracted data in JSON format:".

Task: Extract data from OCR text of a testing report into JSON format.
```

對於小模型很重要的明確提示，以防模型生成額外的東西
(提示詞 or Styling)

文件分析 | MACHINE LEARNING

Prompt Optimizing

```
Output format (JSON only, no code, no introductory text):
{
    "TestLab":"",
    "ReportNumber":"",
    "TestReportDate":"",
    "TestReportCompany":"",
    "SampleName":"",
    "ItemNumber":"",
    "TestItems": [
      {
        "TestPart": "",
        "SampleNumber": "",
        "SubStance": "",
        "CasNo": "",
        "CheckMethod": "",
        "CheckUnit": "",
        "MDL": "",
        "LimitValue": "",
        "CheckResult": ""
      }
    ]
}
```

JSON 格式定義

文件分析 | MACHINE LEARNING

Prompt Optimizing

```
Field definitions:
- TestLab: Testing company (e.g., "SGS", "SGS-CSTC", "Intertek", "全國公證")
- ReportNumber: Report number (look for patterns like "Report No.", "ReportNo.", "測試報告
- TestReportDate: Report date (look for patterns like "Date:", "報告發行日期")
- TestReportCompany: Company requesting testing (look for patterns like "申請廠商", "Compa
- SampleName: Sample name (look for patterns like "樣品名", "sample(s) was/were submitted'
- ItemNumber: Sample item number (look for patterns like "批號", "Model No.")
- TestItems: List of test results
  - TestPart: Test category (e.g., "重金屬", "二氧化硫", "黃毒素", "農藥殘留")
  - SampleNumber: Sample number (if available)
  - SubStance: Testing substance (e.g., "鉛", "鎘", "汞", "砷")
  - CasNo: CAS number (if available)
  - CheckMethod: Testing method (look for patterns like "驗方法", "Test Method:")
  - CheckUnit: Testing units (e.g., "ppm", "ppb", "mg/kg")
  - MDL: Method Detection Limit (look for patterns like "定量極限", "MDL")
  - LimitValue: Limit value (look for patterns like "限值", "Limit")
  - CheckResult: Result (e.g., "未出", "ND", "Not Detected")
```

JSON 各項的「清楚」說明，對小模型很重要

文件分析 | MACHINE LEARNING

Prompt Optimizing

```
OCR tips:
1. Words may be merged or have missing spaces
2. Look for patterns even with missing spaces
3. First page usually contains report metadata
4. Test results are typically in tables or structured sections
5. For Chinese reports:
   - Look for key terms like "測試報告號碼", "報告發行日期", "樣品名", "批號"
   - Test categories are often marked with numbers in parentheses like " (一) ", " (二) "
   - Results may be marked as "未出" for not detected
6. For English reports:
   - Follow the existing pattern recognition rules
   - Look for "Test Part Description:", "Test Method:", etc.
```

針對 RapidOCR 可能產出的東西給的提示

文件分析 | MACHINE LEARNING

Prompt Optimizing

```
Example input (Chinese format):
測試報告號碼：TWNC01023308
報告發行日期：2021年10月05日
申請廠商：廣漢貿易有限公司
樣品名：紅片
批號：215668

（一）重金屬
驗方法：臺灣中藥典第三版(107年)(THP3001)
定量極限（ppm） 检出值(ppm) 限值(ppm)
鉛(Pb) 0.5 未出 5.0
```

```
Example output (JSON only, no introductory text):
{
  "TestLab": "全國公證",
  "ReportNumber": "TWNC01023308",
  "TestReportDate": "2021年10月05日",
  "TestReportCompany": "廣漢貿易有限公司",
  "SampleName": "紅片",
  "ItemNumber": "215668",
  "TestItems": [
    {
      "TestPart": "重金屬",
      "SampleNumber": "",
      "SubStance": "鉛",
      "CasNo": null,
      "CheckMethod": "臺灣中藥典第三版(107年)(THP3001)",
      "CheckUnit": "ppm",
      "MDL": "0.5",
      "LimitValue": "5.0",
      "CheckResult": "未出"
    }
  ]
}
```

一些範例

文件分析 | MACHINE LEARNING

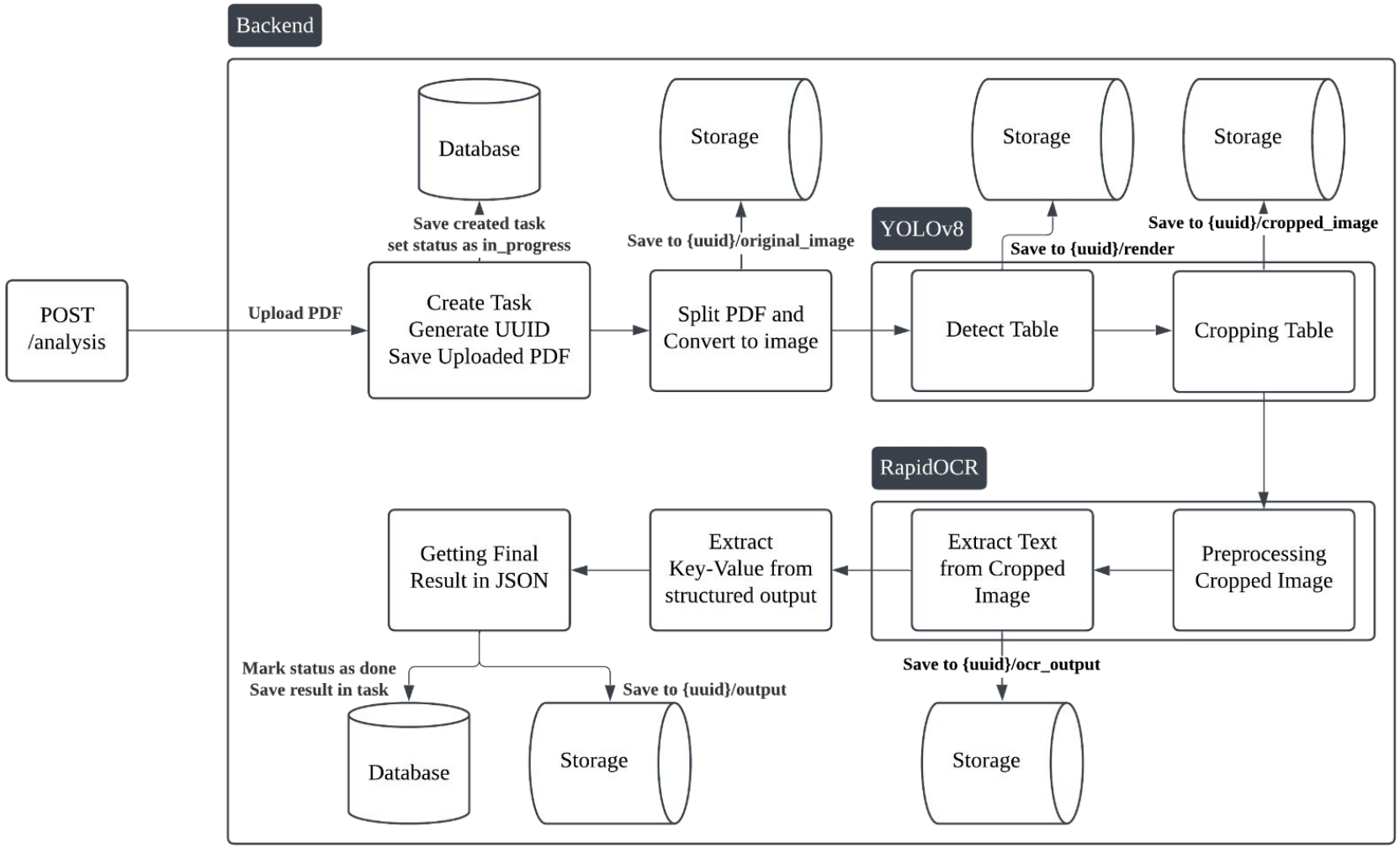
Prompt Optimizing

```
Use "null" for missing data. Use "uncertain" if unsure about a value.

REMEMBER: DO NOT GENERATE CODE. EXTRACT DATA DIRECTLY INTO JSON FORMAT.
DO NOT ADD ANY INTRODUCTORY TEXT LIKE "Here is the extracted data in JSON format:".
```

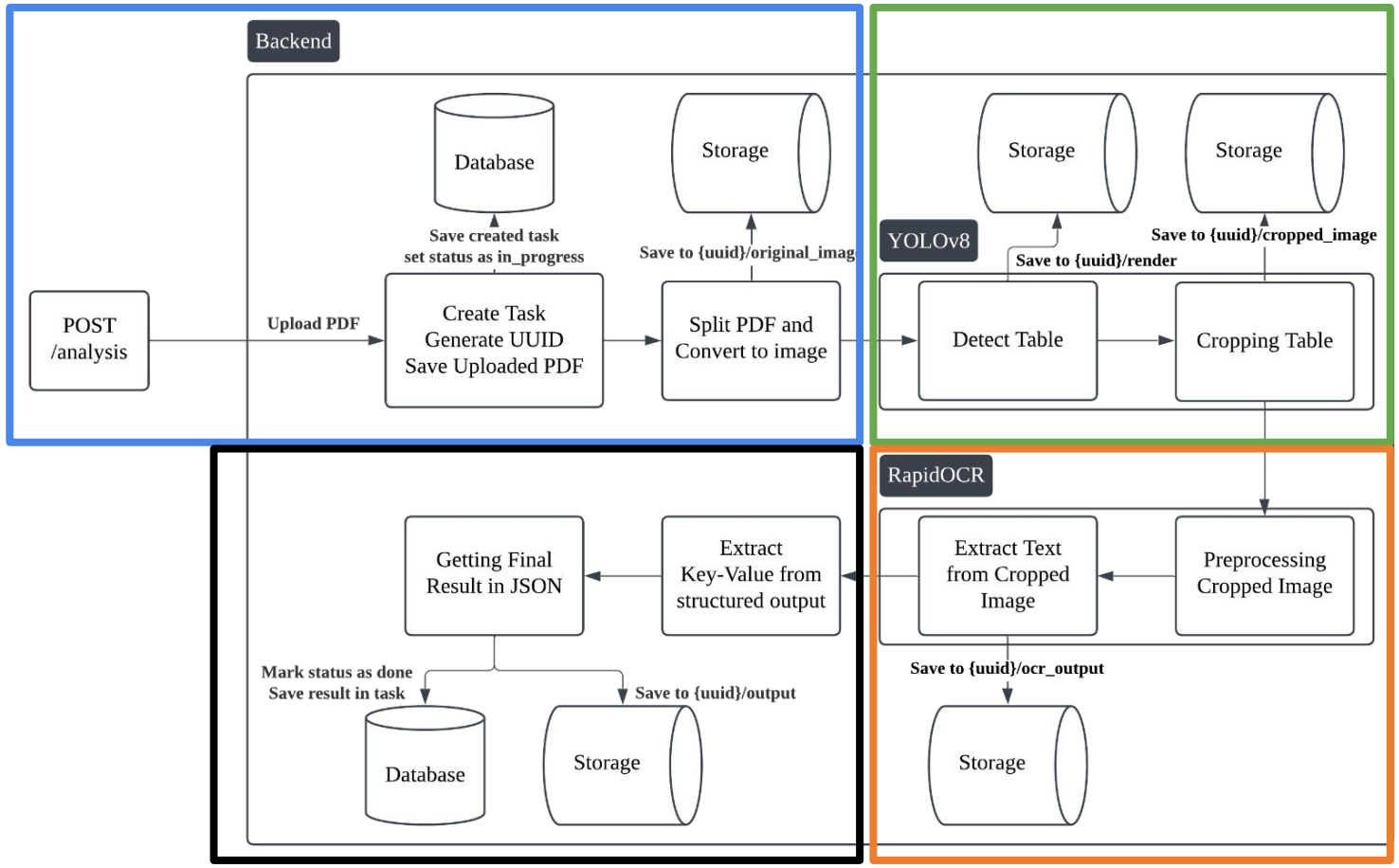
最後還是要再提醒一下，不然小模型都亂 產東西

文件分析 | 表格剖析與讀取



效能標的 : 7頁 SGS 檢驗報告 PDF 檔案 (CANEC2105310907.PDF)

文件分析 | 表格剖析與讀取



使用者上傳 PDF
儲存 PDF 檔案
將PDF分頁並將分頁儲存為照片
耗時 0.593s

根據 YOLO 給出的邊界切出一張一張的細部照片
耗時 3.56s (處理 7 張照片)~ 0.51s/張

照片前處理
RapidOCR 自動辨識及還原結構
耗時 7.775s (處理 7 張照片)~ 1.11s/張

從前面 RapidOCR 產生的 HTML 剖析資料填入設定好的資料區塊中
耗時 1.422s

總耗時: 13.35sec

效能標的 : 7頁 SGS 檢驗報告 PDF 檔案 (CANEC2105310907.PDF)

文件分析 | 表格剖析與讀取

文件表格標記處理

SGS

Test Report No. CANEC2105310907 Date: 15 Apr 2021 Page 2 of 7

Test Results :

Test Part Description :

Specimen No.	SGS Sample ID	Description
SN1	CAN21-053109.001	"SILICON WAFER"

Remarks :

(1) 1 mg/kg = 0.0001%
(2) MDL = Method Detection Limit
(3) ND = Not Detected (< MDL)
(4) "-" = Not Regulated

YOLO 標記

Halogen

Test Method : With reference to EN 14582:2016, analysis was performed by IC.

Test Item(s)	Unit	MDL	001
Fluorine (F)	mg/kg	50	ND
Chlorine (Cl)	mg/kg	50	ND
Bromine (Br)	mg/kg	50	ND
Iodine (I)	mg/kg	50	ND

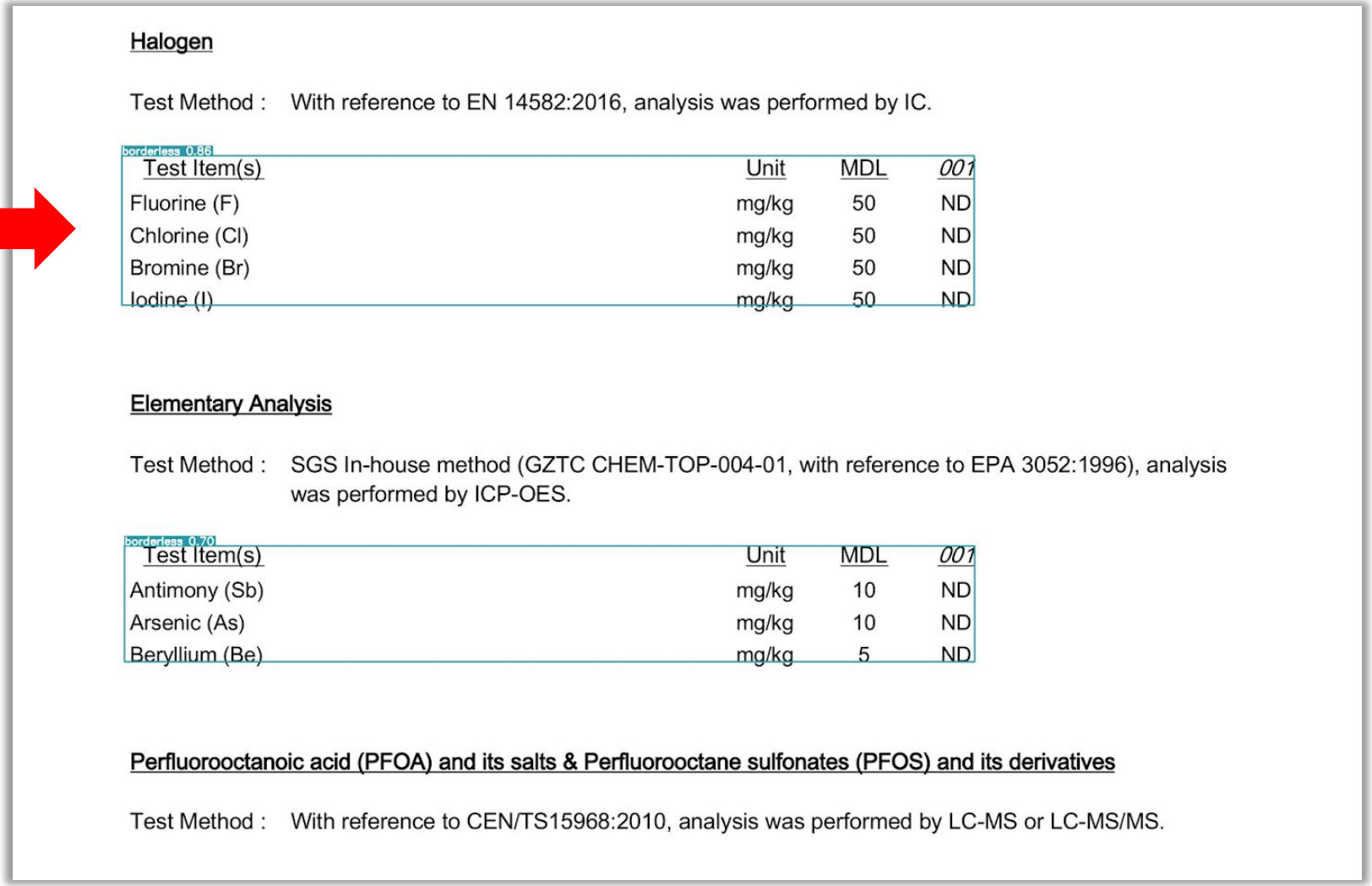
Elementary Analysis

Test Method : SGS In-house method (GZTC CHEM-TOP-004-01, with reference to EPA 3052:1996), analysis was performed by ICP-OES.

Test Item(s)	Unit	MDL	001
Antimony (Sb)	mg/kg	10	ND
Arsenic (As)	mg/kg	10	ND
Beryllium (Be)	mg/kg	5	ND

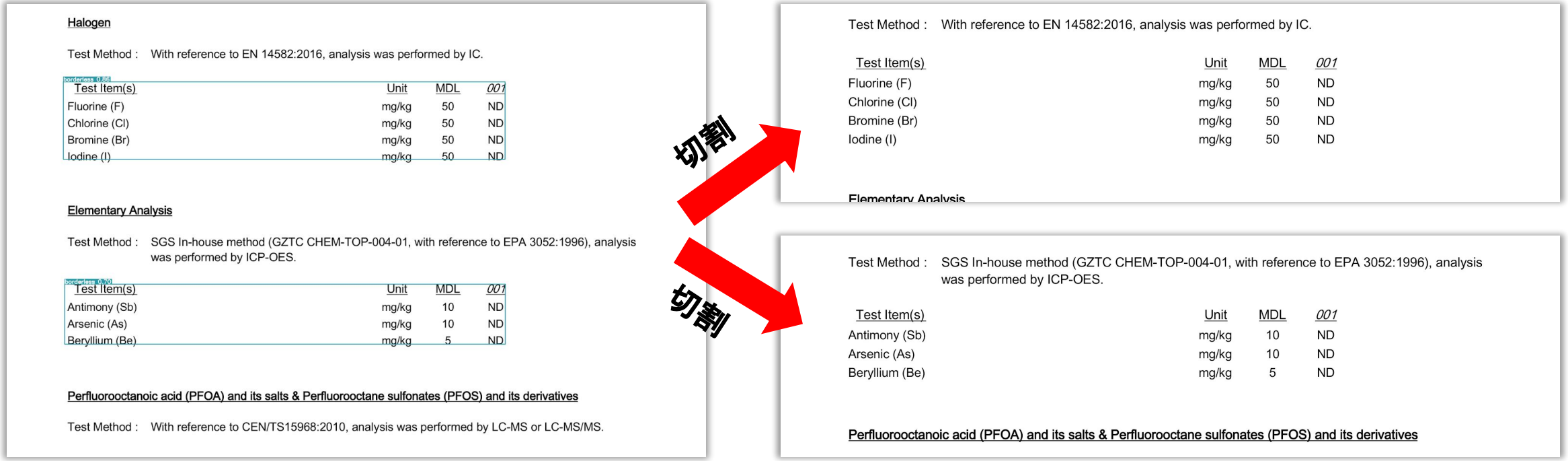
Perfluorooctanoic acid (PFOA) and its salts & Perfluorooctane sulfonates (PFOS) and its derivatives

Test Method : With reference to CEN/TS15968:2010, analysis was performed by LC-MS or LC-MS/MS.



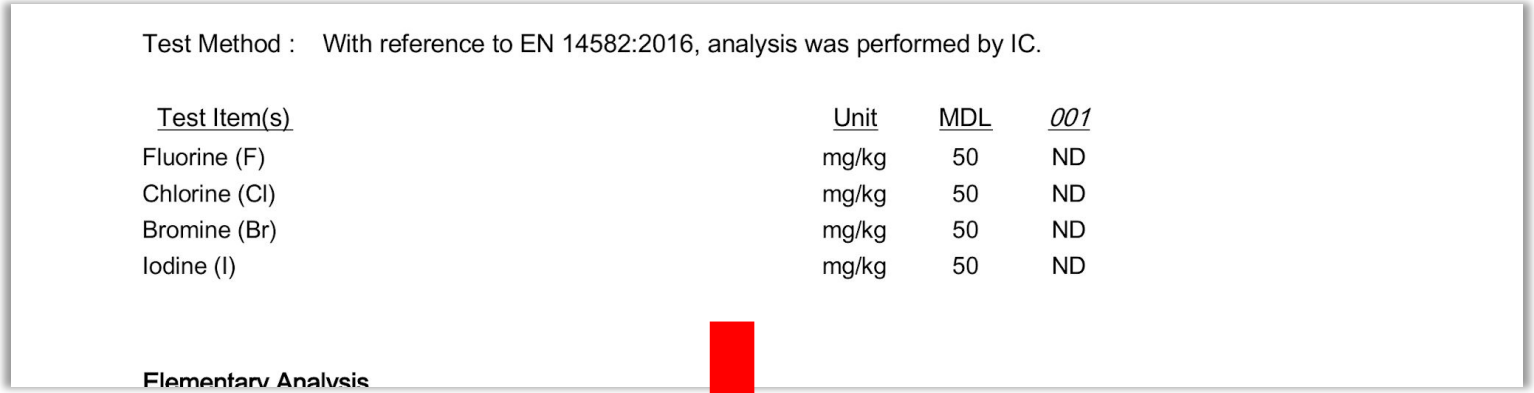
文件分析 | 表格剖析與讀取

文件表格標記處理



文件分析 | 表格剖析與讀取

文字辨識及結構還原



Test Method : With reference to EN 14582:2016, analysis was performed by IC.

Test Item(s)	Unit	MDL	001
Fluorine (F)	mg/kg	50	ND
Chlorine (Cl)	mg/kg	50	ND
Bromine (Br)	mg/kg	50	ND
Iodine (I)	mg/kg	50	ND

Elementary Analysis

RapidOCR 內部自動判斷，提取

TestMethod: WithreferencetoEN14582:2016,analysiswasperformedbyIC.

Test Item(s)	Limit	Unit	MDL	014
Fluorine (F)	900	mg/kg	50 50	ND
Chlorine (CI)	006	mg/kg		ND
Bromine(Br) lodine (l)		mg/kg mg/kg	50 50	ND ND
Comment				PASS

效能標的 : 7頁 SGS 檢驗報告 PDF 檔案 (CANEC2105310907.PDF)

文件分析 | 表格剖析與讀取

第一頁處理

SGS

測試報告

Test Report

號碼(No.) : CE/2020/83834　　日期(Date) : 2020/08/28

昇貿科技股份有限公司
SHENMAO TECHNOLOGY INC.
桃園市觀音工業區工業二路12-1號
NO. 12-1, GONGYE 2ND RD., GUANYIN INDUSTRIAL AREA, TAOYUAN CITY 328, TAIWAN

以下測試樣品係由申請廠商所提供及確認 (The following sample(s) was/were submitt
behalf of the applicant as):

樣品名稱(Sample Description)	:	PURE TIN
樣品型號(Style/Item No.)	:	H99.95S
收件日期(Sample Receiving Date)	:	2020/08/24
測試期間(Testing Period)	:	2020/08/24 to 2020/08/28


```
{
    "TestLab": "SGS Taiwan Ltd.",
    "ReportNumber": "CE/2020/83834",
    "TestReportDate": "2020/08/28",
    "TestReportCompany": "昇贸科技股份有限公司",
    "SampleName": "PURE TIN",
    "ItemNumber": "H99.95S"
}
```

文件分析 | 表格剖析與讀取

文字辨識及結構還原

TestMethod: WithreferencetoEN14582:2016,analysiswasperformedbyIC.

Test Item(s)
Fluorine (F)
Chlorine (CI)
Bromine(Br) lodine (l)
Comment

Limit	Unit	MDL	014
900	mg/kg	50 50	ND
006	mg/kg		ND
	mg/kg mg/kg	50 50	ND ND
			PASS


```
{
  "TestPart": "",
  "SampleNumber": "",
  "SubStance": "",
  "CasNo": null,
  "CheckMethod": "",
  "CheckUnit": "",
  "MDL": "",
  "LimitValue": "",
  "CheckResult": ""
}
```


效能標的 : 7頁 SGS 檢驗報告 PDF 檔案 (CANEC2105310907.PDF)

後端｜文件分析

MACHINE LEARNING

優點

- 準確度高 (錯誤率低)
- 相容度高 (針對不同廠商報告或格式彈性較大)

缺點

- 算力需求高 (需要顯卡 or Apple M系列晶片)
- 速度較慢 (普遍需要 20 secs 以上)
- 部署難度較高 (需要另外架設 LLM Server)

後端｜文件分析

表格剖析與讀取

優點

- 速度快
- 算力需求低
- 部署容易

缺點

- 準確率較低 (容易資料誤植)

未來展望

後端

- MACHINE LEARNING
  - 可以再優化 Prompt，讓更小的模型能產生相同或更好品質的結果 (降低算力需求)。現階段僅從 Llama 8B 優化至 3B，速度加快 4 倍
- 表格剖析與讀取
  - 可以針對 YOLO 及 RapidOCR 使用的模型進行 Fine Tuning，讓表格辨識及剖析更準確
  - 優化 HTML 剖析模組，讓資料誤植機率降低

開發工具

Frontend:
- Vue.js (Language)
- Nuxt.js (Framework)
- Tailwind CSS

Backend:
- Python 3.12.7
- FastAPI (API Router)

- LLM solution
  - Llama 3.2 3B on MLX
  - RapidOCR/RapidTable

- Structure Teardown Solution
  - YOLOv8 & Pillow (Table Detection & Cropping)
  - RapidOCR/RapidTable

資訊技術專案開發與實

綠色供應鏈管理平臺
測試報告智能辨識功能開發

報告結束

感謝您的聆聽！