

AI 輔助生成保障事件決定書稿概 念驗證服務成果報告

中華民國 115 年 3 月 10 日

版本變更紀錄 (Change History)

制/修訂版次 (Version)	制/修訂日期 (Date)	制/修訂說明 (Change Description)	作者 (Author)
1.0	1/29	初版	余健豪
2.0	3/10	修正版	余健豪

目錄

壹、摘要與前言.....	8
一、專案名稱.....	8
二、專案緣起.....	8
三、專案摘要說明.....	8
(一) 驗證自動化去識別化與資料結構化之可行性.....	8
(二) 驗證智慧化爭點偵測與語意檢索之輔助效益.....	9
(三) 探索基於實務案例之草稿生成輔助模式.....	9
四、專案期程.....	10
(一) 第一階段：專案規劃與前期準備.....	10
(二) 第二階段：核心功能驗證與開發.....	10
(三) 第三階段：結案與成果報告.....	10
五、專案工作項目.....	11
(一) 資料前處理與去識別化機制驗證.....	11
(二) 卷宗內容解析與結構化處理.....	11
(三) 爭點萃取與相似案例檢索驗證.....	11
(四) 決定書草稿生成流程驗證.....	11
(五) 成果彙整與可行性評估.....	12
六、交付項目與各項進度.....	12
貳、系統架構及工作項目說明.....	13
一、系統架構說明.....	13
(一) 光學字元辨識 (OCR) 服務.....	13
(二) 大型語言模型分析模組.....	14
(三) 檢索增強生成系統.....	14
(四) 服務前端程式.....	16
(五) 服務後端程式.....	16
(六) 資料儲存與檔案管理機制.....	17
二、系統資源說明.....	17
(一) 最低資源配置 (可執行 PoC 驗證).....	17
(二) 建議資源配置 (較佳效能與穩定性).....	18
(三) 配置說明與選型考量.....	18
三、系統技術範圍及說明.....	19
(一) 系統整合開發架構.....	19
(二) 核心作業技術流程實施.....	20

參、系統管理及操作說明文件	24
一、帳號註冊與登入驗證	24
(一) 註冊與登入流程說明	24
二、卷宗上傳與文件檢視	25
(一) 卷宗上傳與佇列管理	25
(二) 處理狀態即時監控	26
三、AI 卷宗分析與爭點判斷	27
(一) 文本摘要與事實提取	27
(二) 智慧爭點偵測與法理分析	28
(三) 人機協作之動態調修機制	29
四、相似決定書查詢與結果統計	30
(一) 向量語意檢索	31
(二) 搜尋參數設定與進階檢索功能	31
(三) 判決趨勢之視覺化統計	32
(四) 智慧容錯與人工過濾機制	32
五、五段式決定書草稿生成與匯出	33
(一) 決定書結構化自動填充	33
(二) 判決結果切換機制	34
(三) 文書自動化匯出	35
六、系統可同時支援使用者人數與文件處理時間說明	35
(一) 系統可同時支援使用者數量	35
(二) 卷宗文件處理時間說明	36
(三) OCR 處理效能	36
(四) AI 分析處理效能	36
(五) 整體文件處理時間	36
(六) GPU 加速之效能推估	37
七、重新解析機制與錯誤回饋說明	37
肆、測試結果與可行性分析	38
一、測試目標與測試原則	38
二、系統功能與流程驗證成果	38
(一) 卷宗文件上傳與處理流程	38
(二) 案件分析與結構化成果產出	38
三、AI 分析輔助成果說明	38
四、人工介入與修正機制驗證	39
五、系統穩定性與例外處理測試	39
六、測試成果與可行性分析總結	40

七、測試資料來源與資料安全控管機制說明	40
八、AI 分析結果差異說明	41
九、測試案件處理紀錄補充說明	41
伍、結論與建議.....	42
(一) 導入分段式、流程化之案件分析機制	42
(二) 強化人工介入與判斷輸入之設計	42
(三) 擴充判例參考來源與人工指定機制	43
(四) 決定書草稿之人機協作生成模式	43
(五) 建立可回饋之案例累積與知識擴充機制	44
(六) 發展方向總結	44
(七) 未來系統實作所需系統架構圖及軟體規格	45
(八) 未來系統實作所需硬體設備規格與參考價格	47

圖目錄

圖 1 系統架構圖.....	13
圖 2 多維度案件特徵檢索邏輯圖.....	15
圖 3 系統整合服務組件與數據流向圖.....	20
圖 4 卷宗上傳與多模態 AI 分析時序流程圖	21
圖 5 相似案例語意向量檢索時序流程圖.....	22
圖 6 決定書草稿生成與結構化產出時序流程圖.....	23
圖 7 系統帳號註冊介面圖.....	24
圖 8 法律卷宗資料上傳介面圖.....	25
圖 9 文件處理狀態監控介面圖.....	26
圖 10 卷宗彙整與 AI 分析結果介面圖	27
圖 11 AI 自動化法律標籤提取與適用法源分析介面圖.....	29
圖 12 AI 自動化爭點判斷與深度分析建議介面.....	30
圖 13 相似案例檢索與多維度權重調整介面圖.....	30
圖 14 AI 自動生成之決定書段落草稿介面圖.....	34
圖 15 決定書草稿之主文與結論生成結果介面圖.....	34
圖 16 系統整合服務組件與數據流向圖.....	45

表目錄

表 1 交付項目	12
表 2 最低資源配置表	17
表 3 建議資源配置表	18
表 4 相似案例判決結果分類與視覺化標示說明表	32
表 5 建議 GPU 硬體規格參考表	48
表 6 建議伺服器設備參考表	48

壹、摘要與前言

一、專案名稱

專案名稱為公務人員保障暨培訓委員會 114 年度「AI 輔助生成保障事件決定書稿概念驗證服務案」，以下簡稱本專案。

二、專案緣起

面對申訴案件卷宗內容繁雜、人工閱卷與比對過往案例耗時之現況，本案以概念驗證為核心目標，嘗試導入 AI 技術，驗證其於行政案件輔助分析情境中的可行性與應用成效。

本專案聚焦於驗證：是否能透過 LLM 分析既有決定書內容，整理並萃取具參考價值之案例資料，作為新案卷宗分析時的輔助參考依據，進而嘗試產製決定書草稿或分析建議內容，提供使用者作為撰寫與審議時之參考資料。

系統整合 OCR 文字辨識、向量相似度檢索與 LLM 分析技術，驗證從卷宗資料整理、關鍵爭點對應到參考案例彙整之整體流程可行性。本案成果將作為後續評估 AI 技術於行政審議流程中輔助應用之重要依據，並不涉及取代人工判斷或形成具決定效力之內容。

三、專案摘要說明

本專案以概念驗證 (PoC) 為導向，聚焦於驗證 AI 技術於行政案件輔助分析與文書撰擬情境下之可行性，其具體目標如下：

(一) 驗證自動化去識別化與資料結構化之可行性

建構符合資安與個資保護要求之資料處理流程，驗證 AI 技術於卷宗資料前處理階段的應用效果。系統透過 OCR 與語意模

型，嘗試自動識別並遮蓋卷宗內之姓名、身分證字號、地址等敏感資訊，以降低人工處理負擔，並作為後續資料運用之安全前提。同時，驗證將非結構化文本拆解為案件摘要、事實、爭點及法律依據等結構化欄位之可行性，作為後續分析與檢索之基礎資料來源。

（二） 驗證智慧化爭點偵測與語意檢索之輔助效益

透過向量語意比對技術，評估 AI 於實務案例檢索與比對情境中的輔助成效。系統嘗試自動萃取案件之核心爭點，並結合向量資料庫，針對摘要、事實、爭點及法源等多個面向進行相似度檢索，協助承辦人員快速定位具高度關聯性之參考案例，並提供相似度數值作為輔助判斷之依據。

（三） 探索基於實務案例之草稿生成輔助模式

本專案嘗試驗證「以案引案」之文書撰擬輔助流程，評估 LLM 結合當前案情與過往決定書案例後，產製決定書草稿之可行性。

系統將依據參考案例之論證結構與邏輯，嘗試產出具備既定格式（如主文、事實、理由等）之草稿內容，提供使用者作為撰寫與審議時之參考資料，以評估其在降低初稿撰擬負擔、提升作業一致性方面的實務價值；惟最終判斷與內容確認仍由承辦人員負責。

四、 專案期程

此時程規劃以專案啟動日（2025 年 9 月 9 日）為起點，將所有關鍵里程碑與交付物清晰地標示出來。

（一） 第一階段：專案規劃與前期準備

- 日期：至 2025 年 9 月 29 日
- 內容：包含專案目標、範圍、核心功能細節、驗證指標、技術架構規劃與初步時程規劃等。
- 交付項目：專案計劃書、專案人員名冊、保密切結書、保密同意書及著作財產權與著作人格權同意書。

備註：此階段為規劃及分析階段。

（二） 第二階段：核心功能驗證與開發

- 資料去識別化服務驗證：驗證模型在小規模數據集上的準確性與效率。
- 爭點生成與法規匹配服務驗證：測試 LLM 標註能力與 RAG 檢索結果的相關性。
- 決定書草稿生成服務驗證：評估草稿的生成品質與對效率的提升幅度。

備註：此為專案的主要開發與驗證階段。

（三） 第三階段：結案與成果報告

- 日期：至 2025 年 2 月 8 日
- 內容：總結 PoC 階段的驗證成果、技術可行性分析、未來正式導入的建議。

五、 專案工作項目

（一） 資料前處理與去識別化機制驗證

針對納入概念驗證之歷史決定書及相關卷宗資料，建置資料前處理流程，包含 OCR 文字辨識、文件格式轉換及內容清理等作業，並驗證自動化去識別化技術於實務資料中的可行性，嘗試識別並遮蔽姓名、身分證字號、地址等個人敏感資訊，以符合個人資料保護與資訊安全相關規範。

（二） 卷宗內容解析與結構化處理

本案針對非結構化卷宗文本，驗證運用語意分析與大型語言模型進行內容解析之可行性，嘗試將資料拆解為案件摘要、事實描述、爭點說明及相關法規依據等結構化欄位，作為後續爭點比對與案例檢索之基礎資料。

（三） 爭點萃取與相似案例檢索驗證

透過語意分析與向量化技術，嘗試自卷宗資料中萃取案件核心爭點，並建置相似案例檢索流程，結合向量相似度比對方式，針對歷史決定書資料進行多面向比對，驗證系統於協助承辦人員快速定位具參考價值案例之實務效益，並提供相似度指標作為輔助判斷依據。

（四） 決定書草稿生成流程驗證

在前述資料解析與案例檢索基礎上，本案嘗試驗證以「以案引案」方式輔助文書撰擬之可行性，評估大型語言模型結合當前案件資料與歷史案例後，產製決定書草稿之能力。生成內容以既有決定書架構為參考，產出主文、事實及理由等段落，僅作為草稿或參考資料使用，不涉及取代人工判斷或形成具法律效力之內容。

(五) 成果彙整與可行性評估

本專案彙整各項概念驗證成果，針對 AI 技術於保障事件決定書輔助分析與草稿生成情境中的適用性、限制條件與風險進行整理，作為後續是否進行正式系統導入、功能擴充或流程優化之評估基礎。

六、 交付項目與各項進度

表 1 交付項目

項次	交付項目	時程	進度
1	專案計畫書	114 年 9 月 29 日	已完成
2	結案報告	115 年 2 月 8 日	已完成

貳、系統架構及工作項目說明

一、系統架構說明

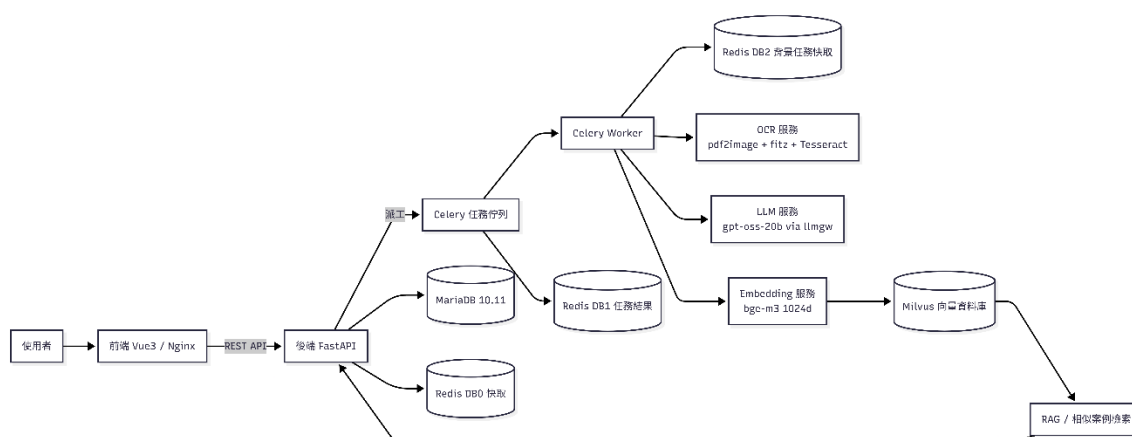


圖 1 系統架構圖

(一) 光學字元辨識 (OCR) 服務

本服務負責將使用者上傳之卷宗文件 (PDF) 轉換為可供後續 AI 分析的純文字資料，並於流程中完成基本影像前處理與去識別化作業，作為系統後續分析 (爭點、標籤、法律依據、草稿生成) 的輸入來源。

在 PDF 類型文件處理上，系統採用 pdf2image + fitz 進行頁面轉圖，並以批次方式處理 (例如每批 50 頁)，同時支援影像旋轉方向偵測與灰階化等前處理，以提升 OCR 辨識穩定度。文字辨識引擎以 EasyOCR 為主，並保留 GPU 加速可選配置。

完成文字擷取後，系統進入去識別化階段，透過 LLM 識別敏感資訊 (如 PERSON、ID、PHONE、FAX、BIRTHDAY 等)，並替換為[欄位_索引]格式；去識別化流程支援批次並行 (例如 5 併發) 以提升處理效率。最終輸出為去識別化之純文字，供後續 LLM 分析與向量化使用。

(二) 大型語言模型分析模組

本模組為系統的語意理解與結構化分析核心，主要將 OCR 去識別化後的文本轉換為「可被處理與檢索」的結構化結果。

系統採用 gpt-oss-20b 作為主要分析模型，並透過 LangChain + AsyncClient 進行非同步呼叫；LLM 呼叫經由 LLM Gateway 統一介接與控管。同時，系統使用 eland-embedding 作為 embedding 模型，產生 1024 維向量，供後續 RAG 與相似案例檢索。

分析流程包含：

- 案件段落切分：將文本切分為案件摘要、案件事實、案件理由。
- 屬性萃取：抽取案件標題、類型、決定字號、管轄法院、上訴人屬性等。
- 爭點分析：針對常見爭點面向進行歸納（如考績公平性、懲處程序瑕疵、病假限制等）。
- 標籤生成：自預定義標籤中挑選 5-10 個，用於後續檢索與分類。
- 法律依據分析：先選擇適用法規，再細到具體條文。
- 輸出結果為結構化案件分析資料，可供 UI 呈現、RAG 檢索、以及草稿生成流程使用。

(三) 檢索增強生成系統

RAG 系統負責將「案件內容」與「歷史決定書/法規」之語意關聯轉換為可量化、可排序的相似案件列表，提供後續草稿生成與承辦審閱之參考依據。向量資料以 Milvus 管理，索引採 HNSW (M=16, efConstruction=200)，相似度計算使用 COSINE，嵌入維度為 1024。

向量集合 (Collections) 依資料型態分流：

- 案件內容匹配：summary_vector、facts_vector
- 爭點匹配：issue_description_vector、issue_analysis_vector
- 法律依據匹配：legal_basis_description_vector
- 完整決定書匹配：vector

查詢時，系統會先由 Embedding Service 生成查詢向量，再並行執行多路徑向量檢索（摘要/事實/爭點/法源等）。檢索結果進入聚合階段：計算來源權重、產出加權分數（原始分數 × 權重）、針對缺失來源做分數填補（例如以最低分的 90% 補齊），最後依最終加權分數排序輸出相似案件清單，並保留各來源的分數構成以利追溯。

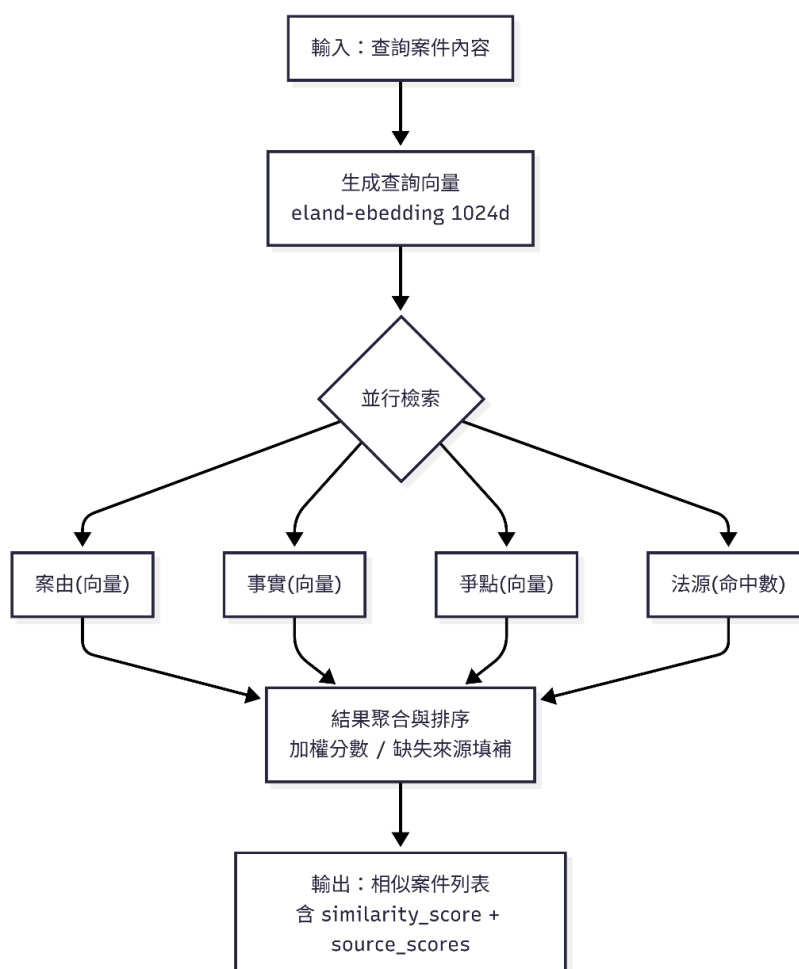


圖 2 多維度案件特徵檢索邏輯圖

(四) 服務前端程式

前端採用 Vue 3 建置，開發/打包工具為 Vite，樣式框架 Tailwind CSS，狀態管理 Pinia，HTTP 客戶端 Axios (含 JWT 自動注入)，路由 Vue Router。

前端主要任務是提供使用者完成：登入/註冊、卷宗上傳、處理狀態監控、案件分析結果檢視、相似案例搜尋與決定書草稿生成/編輯，以及 DOCX 匯出。由於後端存在長耗時任務，前端針對草稿生成等流程採非同步輪詢進度方式，確保使用者操作流暢並可視化掌握處理狀態。

(五) 服務後端程式

後端使用 FastAPI 提供 API 服務，ORM 採 SQLAlchemy；認證採 JWT (HS256)，Token 有效期 30 分鐘。

為處理 OCR、LLM 分析、向量化、草稿生成等長耗時計算，系統以 Celery 建立非同步任務架構，使用 Redis 作為 Broker 與結果儲存 (Backend)，並透過任務容錯設定提升任務可靠度。

文件處理流程，建立 Document 記錄並送入 Celery；Worker 依序執行 OCR 擷取、去識別化、LLM 分析、向量嵌入、寫入 Milvus，並以 ProcessingStatus 狀態碼回寫供前端監控。

（六） 資料儲存與檔案管理機制

系統採「關聯式資料 + 向量資料 + 檔案儲存」的分層策略：

- MariaDB 10.11：存放使用者、文件、案件、爭點、標籤、法律依據、草稿任務與草稿段落（JSON）等結構化資料，支援關聯查詢與權限控管。
- Milvus：存放各類語意向量集合以支援相似檢索。
- 檔案儲存：上傳檔案集中，以 UUID 命名保存原始檔與擷取文字；並包含 EasyOCR 語言資料。
- Redis：依 DB 分區使用（例如 DB0 一般快取、DB1 Celery 任務結果、DB2 背景任務快取）。
- 上傳限制與檢核：單檔最大 50MB，支援 PDF，並使用檔案雜湊進行重複檢測。

二、 系統資源說明

為確保 CSPTC 智慧決策系統於「卷宗 OCR → 去識別化 → LLM 分析 → 向量化 → 相似案例檢索 → 草稿生成」等流程中具備足夠之運算與儲存能力，本案建議之硬體資源配置如下，並依使用情境提供「最低可運行」與「建議配置」兩級規格，供後續實際系統參考評估。

（一） 最低資源配置（可執行 PoC 驗證）

適用情境：少量案件、單機部署、功能驗證為主，允許較長處理時間。

表 2 最低資源配置表

項目	配置
GPU	L40s × 2
CPU	16 cores
RAM	64GB
DISK	2 TB SSD

（二）建議資源配置（較佳效能與穩定性）

適用情境：多使用者併發、較大卷宗量、需較穩定的處理時間與任務吞吐。

表 3 建議資源配置表

項目	配置
GPU	H100 ×1 + L40s ×1
CPU	32 cores
RAM	128GB
DISK	4 TB SSD

（三）配置說明與選型考量

- **GPU**：主要負載集中於 OCR 影像處理（若啟用 GPU 加速）、LLM 推論、Embedding 向量生成等；建議配置以 H100 承擔主要推論負載，L40s 分擔 OCR/向量化或作為尖峰緩衝。
- **CPU/RAM**：PDF 轉圖、文字處理、任務佇列與 API 服務會消耗 CPU；同時大量頁面轉換與併發任務會拉高記憶體需求。
- **DISK (SSD)**：用於儲存上傳原始檔、轉換中間檔、擷取文字、任務暫存與資料庫成長；建議採 SSD 以降低 I/O 成為瓶頸。

三、 系統技術範圍及說明

本系統透過分層式架構設計以確保整體運作效率與系統穩定性，並針對卷宗處理、AI 分析與相似案例檢索等核心作業流程進行技術整合與優化。

(一) 系統整合開發架構

本系統採分層設計原則，確保各功能模組間具備低耦合之特性，以利後續維護、擴充與效能調整。各層級技術運用說明如下：

■ 前端 UI 層 (VueJS & Nginx)

前端介面採用 VueJS 建置單頁式應用 (SPA)，提供直覺且具回應性的操作介面；並透過 Nginx 擔任系統入口與反向代理角色，執行請求轉發、靜態資源服務及基礎負載控管，以確保使用者操作流程之順暢性與穩定性。

■ 應用服務層 (FastAPI & Celery)

應用服務層以 FastAPI 作為 API 核心，負責整體作業流程控制與系統服務介接；針對 OCR 辨識、語意分析、向量化與草稿生成等高運算與長耗時作業，則透過 Celery 非同步任務機制進行排程與執行，以避免主流程阻塞，確保系統在多案件處理情境下仍能維持良好回應能力。

■ 資料層 (MySQL & Milvus)

系統採雙軌資料儲存架構：由 MySQL 管理案件基本資料、結構化分析結果與系統紀錄等關聯式資料；同時使用 Milvus 向量資料庫存放語意向量，支援高維度相似度搜尋需求，作為檢索增強生成 (RAG) 機制之基礎。

■ AI 模型層 (GPUStack)

AI 模型層由 GPUStack 統一管理與調度各項模型資源，包含大型語言模型 (GPT-OSS)、語意嵌入模型 (Embedding) 與 OCR 引擎 (EasyOCR)，提供去識別化處理、爭點分析、語意理解與文字辨識等核心 AI 能力，並確保運算資源能依任務需求進行彈性分配。

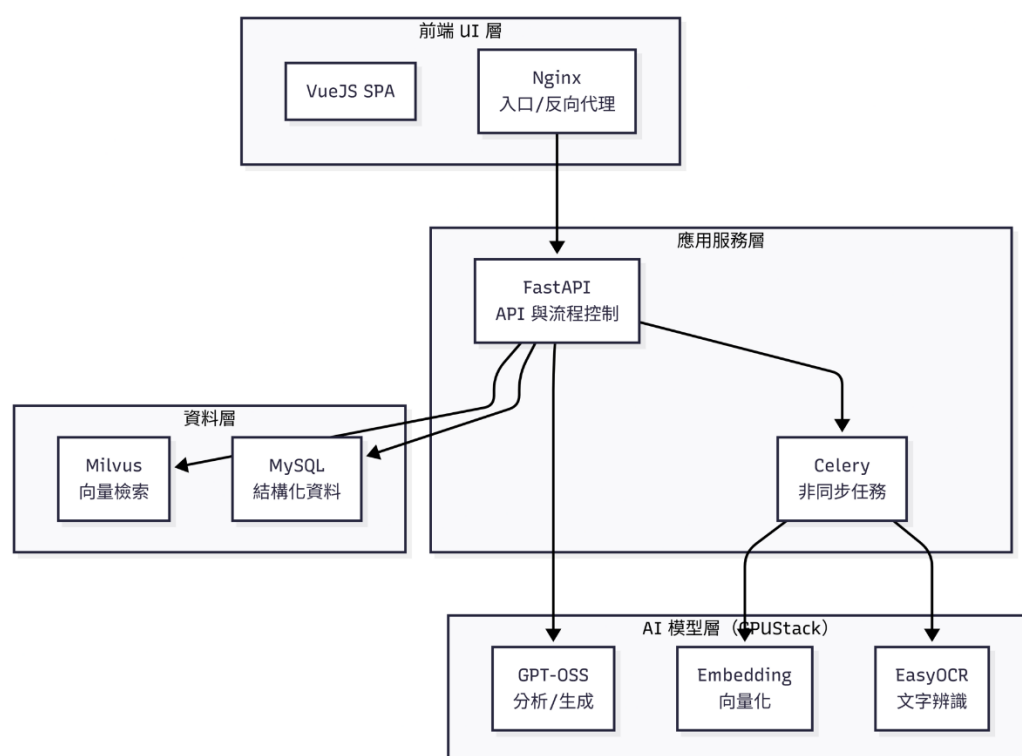


圖 3 系統整合服務組件與數據流向圖

(二) 核心作業技術流程實施

本系統針對法律實務中最重要三項作業場景，設計了嚴謹的技術實作流程，確保數據處理之精確性與運算效能。

1. 卷宗上傳與 AI 自動化分析流程

當使用者上傳 PDF 卷宗後，系統透過非同步任務機制啟動分析鏈路，將原始影像轉譯為高品質之法律結構數據：

- 任務建立與分發：由 FastAPI 接收上傳請求並於 MySQL 建立案件紀錄，隨即將高耗能任務送往 Celery 佇列。
- OCR 佈局辨識：由 GPUStack 調度 EasyOCR 模型進行文本還原與版面佈局分析。
- 語意提取與去識別化：透過 GPT-OSS 進行文本摘要、重點提取、法規過濾及敏感資訊去識別化處理。
- 數據向量化與同步：將分析結果透過 Embedding 模型轉化為向量，並同步存入 MySQL 與 Milvus。

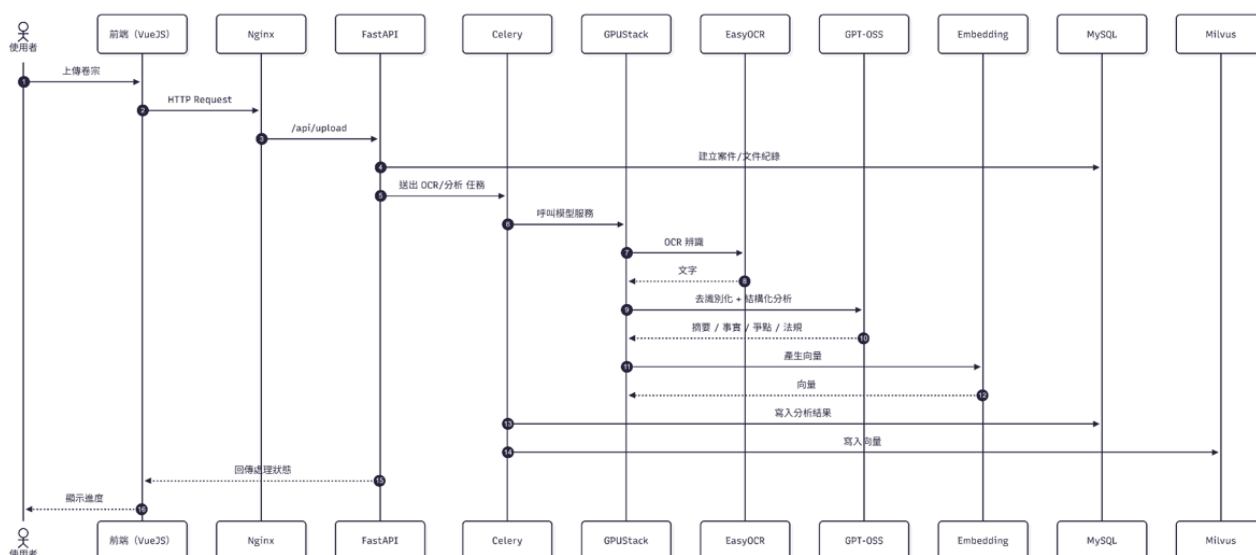


圖 4 卷宗上傳與多模態 AI 分析時序流程圖

2. 相似案例語意檢索流程

為提升法律見解之檢索效率，系統實作了基於 RAG 架構之語意比對技術：

- 查詢向量化：FastAPI 接收查詢關鍵字後，呼叫 Embedding 模型產出代表查詢意圖的高維度向量。
- 高維度空間檢索：利用 Milvus 向量資料庫執行 Top-K 相似度比對，鎖定最具關連性的歷史案件索引。

- 事實內容還原：根據檢索結果，自 MySQL 提取完整案例事實與法理邏輯，並由前端呈現檢索結果。

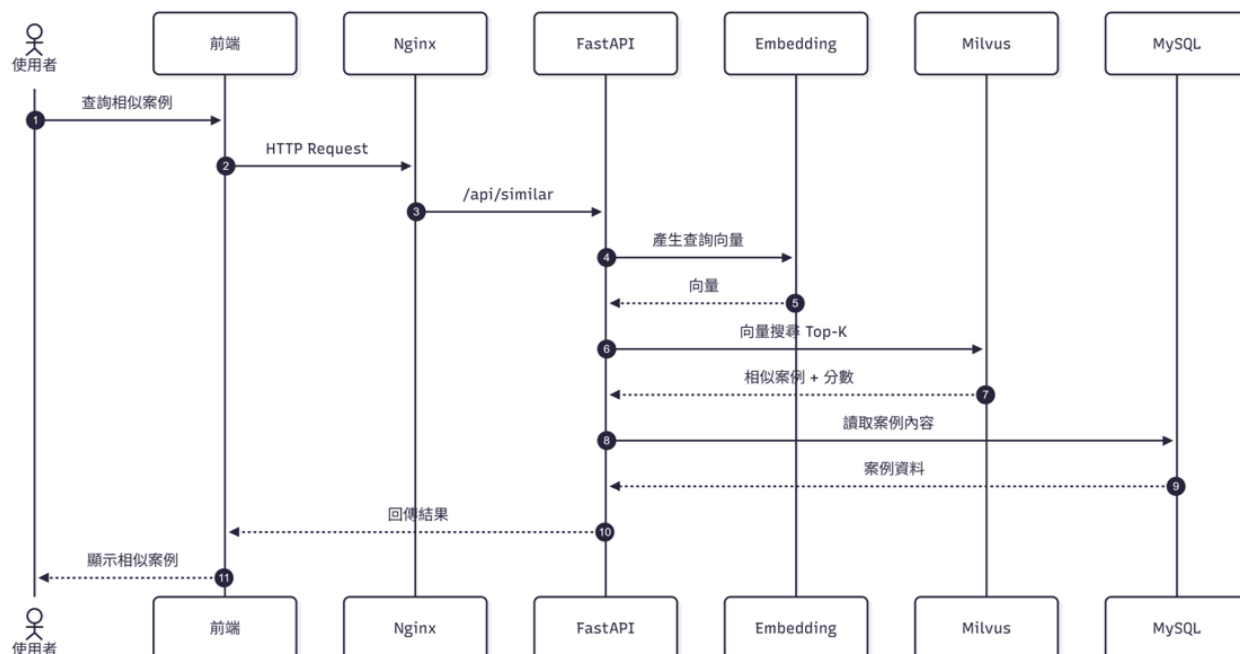


圖 5 相似案例語意向量檢索時序流程圖

3. 決定書草稿生成流程

在完成案例檢索後，系統可進一步協助使用者執行決定書之初稿撰寫：

- 生成任務調度：使用者選定參考案例後，由 FastAPI 發起草稿生成指令至 Celery。
- RAG 脈絡合成：GPUStack 調用 GPT-OSS 模型，並注入 RAG 檢索之法理依據作為脈絡。
- 結構化逐段生成：模型依據規範，逐段擬具主文、事實、理由、法律依據及決定等內容。
- 成果儲存與預覽：生成的草稿回存至 MySQL，並由後端通知前端介面呈現分析結果，供使用者進行後續匯出。

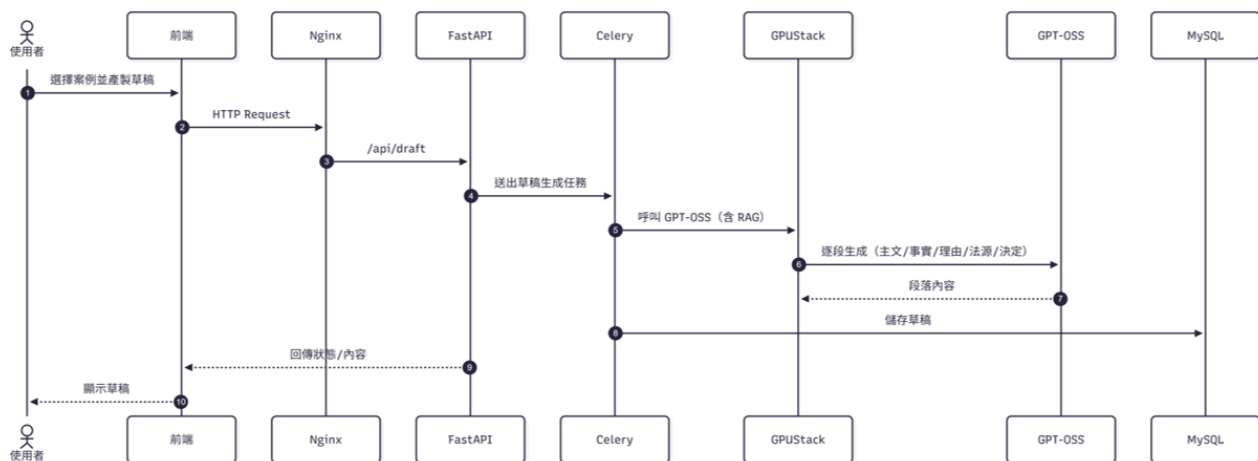


圖 6 決定書草稿生成與結構化產出時序流程圖

參、系統管理及操作說明文件

一、帳號註冊與登入驗證

本系統採用嚴格的身分准入機制，確保僅有經授權之特定機關人員可存取系統資源。

(一) 註冊與登入流程說明

系統設計以「安全性優先」為原則，具體操作流程如下：

1. 網域初步篩選：僅有使用 @csptc.gov.tw 之信箱方可提交註冊申請。提交申請後，系統將發放有效期為 10 分鐘的驗證連結至所申請之信箱。
2. 使用者資料設定：使用者須自行設定「使用者名稱」及「密碼」。為提升帳戶安全性，密碼將使用 Argon 2 來做加密儲存。
3. 帳號登入驗證：使用者通過身分驗證後，系統將發放有效期為 30 分鐘的加密 JWT Token，確保連線安全並引導至首頁。



The image shows a registration form titled "註冊帳號" (Register Account). It contains four input fields: "Email" with a placeholder "請輸入 Email", "用戶名" (Username) with a placeholder "請輸入用戶名 (至少3個字元)", "密碼" (Password) with a placeholder "請輸入密碼 (至少6個字元)", and "確認密碼" (Confirm Password) with a placeholder "請再次輸入密碼". Below these fields is a blue "註冊" (Register) button. At the bottom, there is a link "已有帳號？立即登入" (Already have an account? Log in immediately).

圖 7 系統帳號註冊介面圖

二、 卷宗上傳與文件檢視

負責處理原始卷宗的導入與自動化預處理程序，確保大量公務文件能有序且高效地轉化為系統可分析的數據。

（一） 卷宗上傳與佇列管理

1. 格式支援與上傳驗證：系統支援 PDF 格式上傳，單一檔案上限為 50 MB。上傳時，系統會同步進行初步的檔案損壞檢查，確保後續 OCR 識別之穩定性。此外，系統具備重複檢查機制，會自動比對檔名與檔案大小，若偵測到重複檔案將自動跳過並提醒，避免重複運算造成之資源浪費。
2. 異步處理機制：考量到大型卷宗解析耗時，系統採用異步處理執行。使用者完成上傳後，無須留在原頁面等待，由後端排程自動執行本文提取任務。

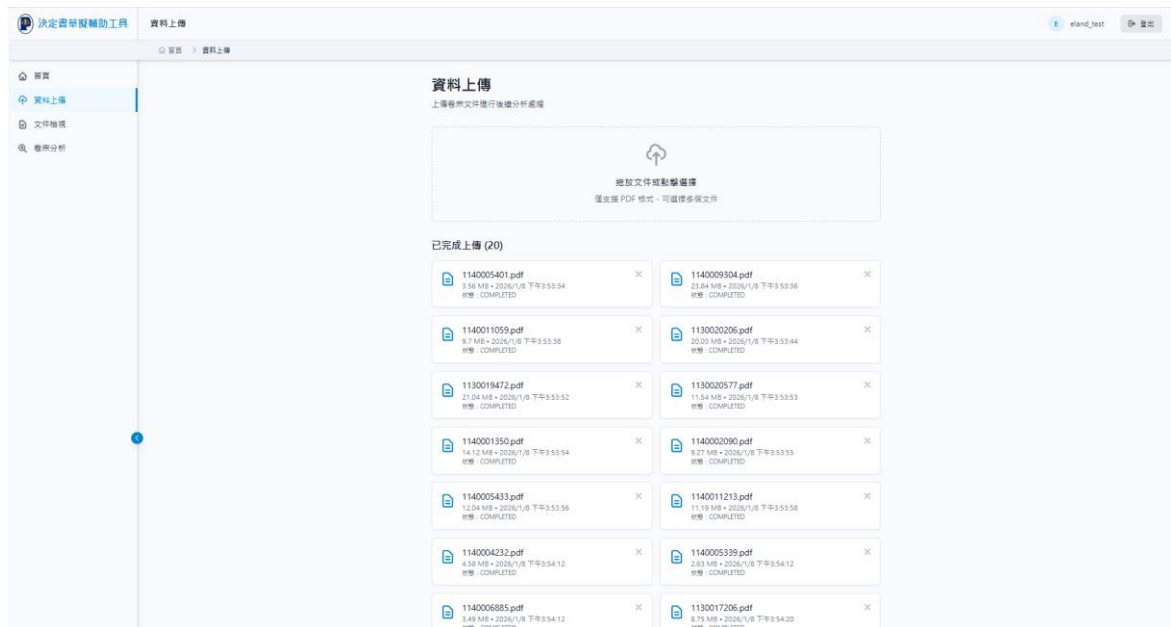
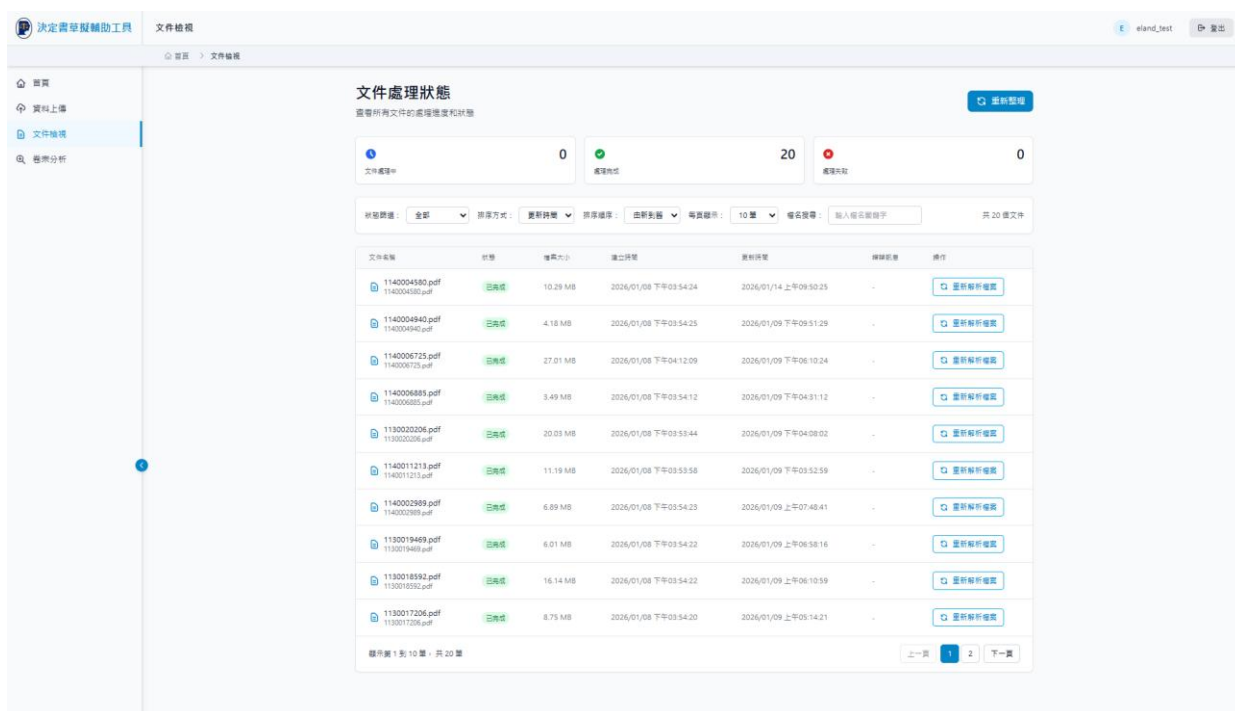


圖 8 法律卷宗資料上傳介面圖

（二）處理狀態即時監控

為提供直觀的作業回饋，系統設計了狀態監控介面，將處理流程分為以下狀態：

1. 文件處理中：系統正在進行 OCR 文字識別、語意結構化分析及向量特徵提取。
2. 處理完成：代表文本已成功轉化為結構化數據並存入資料庫。此時系統顯示「已完成」標籤，使用者則可正式開始 AI 卷宗彙整、爭點判斷與相似案件查詢等核心流程。
3. 處理失敗：如遇分析處理異常，系統將標註「處理失敗」。為提升操作便利性，系統內建自動容錯機制，使用者無需刪除重新上傳，僅需點擊「重新解析檔案」即可再次啟動處理程序，確保作業連續性。



文件名稱	狀態	檔案大小	進出時間	解析時間	後續狀態	操作
1140004580.pdf 1140004580.pdf	已完成	10.39 MB	2026/01/08 下午03:54:24	2026/01/14 上午09:50:25	...	重新解析檔案
1140004940.pdf 1140004940.pdf	已完成	4.18 MB	2026/01/08 下午03:54:25	2026/01/09 下午09:51:29	...	重新解析檔案
1140006725.pdf 1140006725.pdf	已完成	27.01 MB	2026/01/08 下午04:12:09	2026/01/09 下午08:10:24	...	重新解析檔案
1140006885.pdf 1140006885.pdf	已完成	3.49 MB	2026/01/08 下午03:54:12	2026/01/09 下午04:31:12	...	重新解析檔案
1130020206.pdf 1130020206.pdf	已完成	20.03 MB	2026/01/08 下午03:53:44	2026/01/09 下午04:08:02	...	重新解析檔案
1140011213.pdf 1140011213.pdf	已完成	11.19 MB	2026/01/08 下午03:53:58	2026/01/09 下午03:52:59	...	重新解析檔案
1140002989.pdf 1140002989.pdf	已完成	6.89 MB	2026/01/08 下午03:54:23	2026/01/09 上午07:48:41	...	重新解析檔案
1130019469.pdf 1130019469.pdf	已完成	6.01 MB	2026/01/08 下午03:54:22	2026/01/09 上午08:58:16	...	重新解析檔案
1130018592.pdf 1130018592.pdf	已完成	16.14 MB	2026/01/08 下午03:54:22	2026/01/09 上午08:10:59	...	重新解析檔案
1130017206.pdf 1130017206.pdf	已完成	8.75 MB	2026/01/08 下午03:54:20	2026/01/09 上午05:14:21	...	重新解析檔案

圖 9 文件處理狀態監控介面圖

三、 AI 卷宗分析與爭點判斷

為系統之運算核心，運用大型語言模型（LLM）針對解析完成之本文進行深度語意挖掘，協助使用者從海量資料中快速釐清案件全貌。

（一） 文本摘要與事實提取

利用 LLM 的強大提取能力，將非結構化的卷宗轉化為高度結構化的案件背景資料，為後續的爭點分析奠定基礎，具體分析維度包含：

1. 核心實體資訊提取：系統自動辨識並梳理案件之基本屬性，確保資料的一致性與可檢索性。透過提取案件類型、申訴人資訊等實體，為後續的自動化統計奠定基礎。
2. 文本脈絡化處理：針對繁雜的文件內容，系統將執行深度語意壓縮及還原，產出以下結構化文本：

- 案件摘要：自動歸納全案重點，供承辦人員快速掌握全貌。
- 案件事實：精確還原事件發生經過，移除冗餘修辭，保留相關事實。

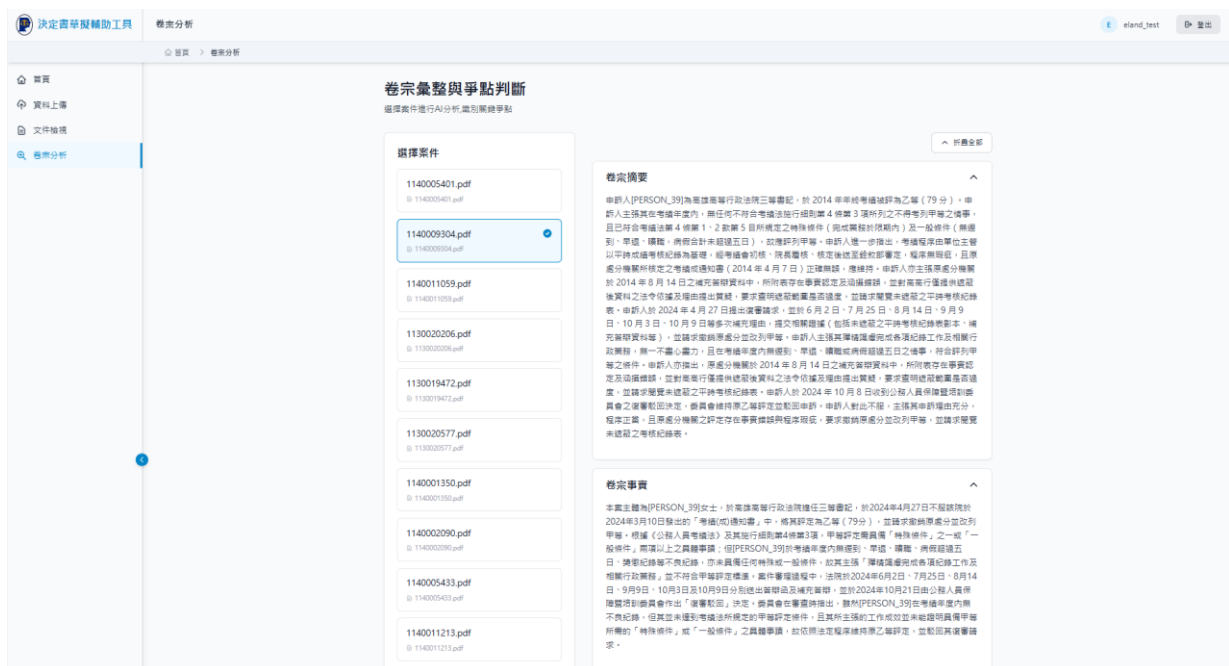


圖 10 卷宗彙整與 AI 分析結果介面圖

（二） 智慧爭點偵測與法理分析

結合 LLM 深度語意理解後與後端法規資料庫，將案件分析從單純的「事實歸納」提升至「法律推論」層次，具體包含以下流程：

1. 多層次爭點識別與自動化標籤技術：系統針對卷宗內容進行全方位的法律衝突偵測，並將關鍵特徵轉化為視覺化標籤，以便使用者快速檢索與分類：

- 分類標籤自動化生成：系統透過 AI 語意識別，自動於介面生成對應之標籤組，包含「懲處種類」（如：申誡、記過）、「爭議點類別」（如：程序違法、事實認定爭議）及「行為類型」（如：言語霸凌、精神壓迫）等。
- 爭點描述與分析：除了標籤化識別，系統亦提供精確的爭點描述與初步 AI 分析建議，作為後續撰寫理由段落之參考。

2. 法律元件之智慧化關聯：系統透過對語意分析與後端法律資料庫比對，自動勾稽案件關聯法源，建立專業法律分析基礎：

- 關鍵法源自動檢索：系統根據卷宗與爭點內容，從現有法律資料庫中自動篩選出本案適用的法律規章（如：公務人員保障法、考績法等）並以條列清單呈現。
- 法規清單之精確化：透過 AI 自動過濾，彙整出與案情具備高度關聯性之法律名稱，協助使用者快速確認法律框架，減少手動搜尋相關法典之負擔。

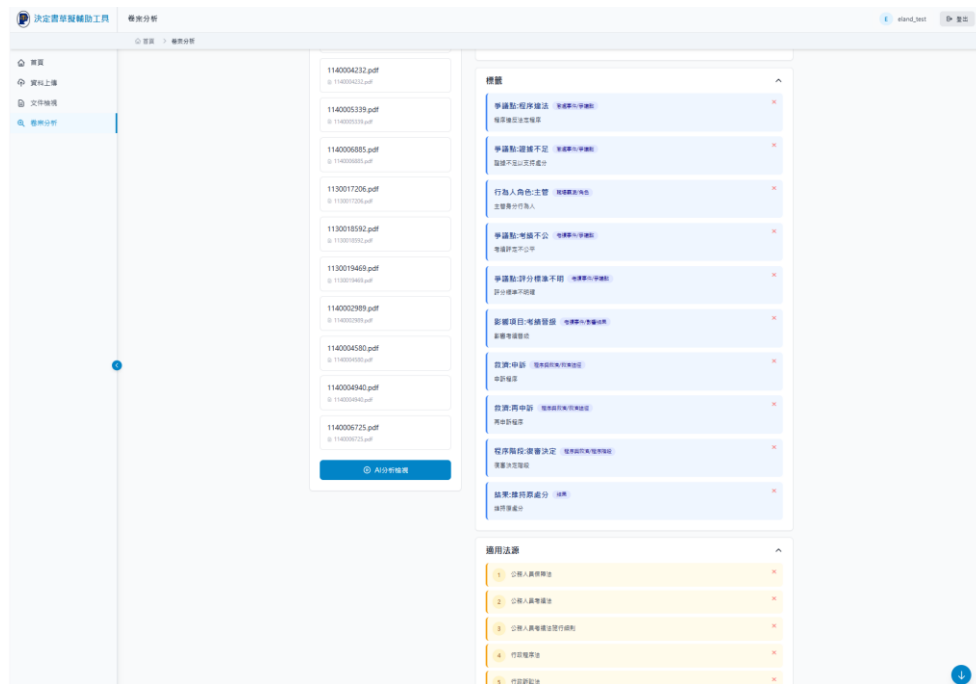


圖 11 AI 自動化法律標籤提取與適用法源分析介面圖

(三) 人機協作之動態調修機制

考量法律實務判斷之嚴謹性，系統設計了高度靈活的編輯環境，確保每一份分析報告都能在 AI 的處理效率與使用者的專業判斷間取得平衡：

1. 爭點內容之增修與刪除：使用者能針對 AI 生成的「爭點描述」或「分析建議」進行修訂或刪除以排除偏誤。針對案情複雜或涉及最新法理見解之特殊情境，系統賦予使用者完整主導權，可視需求於分析介面中直接「新增」爭點，彌補資料庫之侷限性。
2. 標籤管理與智慧去冗餘：針對系統自動標記之各類爭點與類別標籤，使用者若認為不具關聯性，可點擊「X」圖示即時刪除，以確保後續案件之辨識精確度。
3. 適用法源之精煉與校正：針對系統自動檢索出之適用法律清單，介面提供即時刪除機制。使用者若認為特定法規與本案情節不符，可點擊「X」圖示予以剔除，確保最終生成的決定書草

稿僅引用精確且適當的法源基礎。

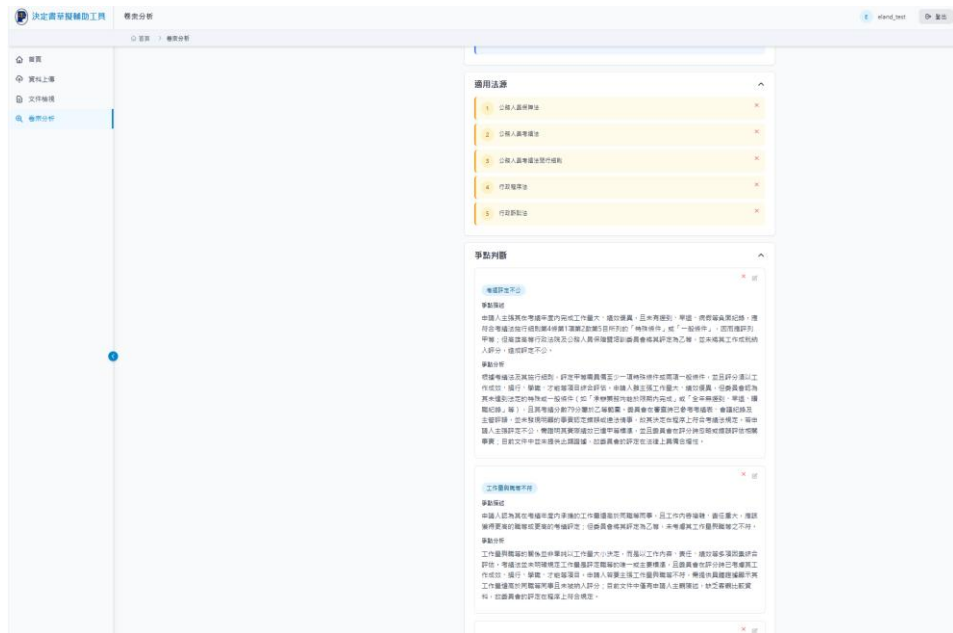


圖 12 AI 自動化爭點判斷與深度分析建議介面

四、 相似決定書查詢與結果統計

旨在建立「以索引案」的智慧檢索機制，協助使用者透過歷史案件之實務見解，確保行政處分之穩定性與一致性。

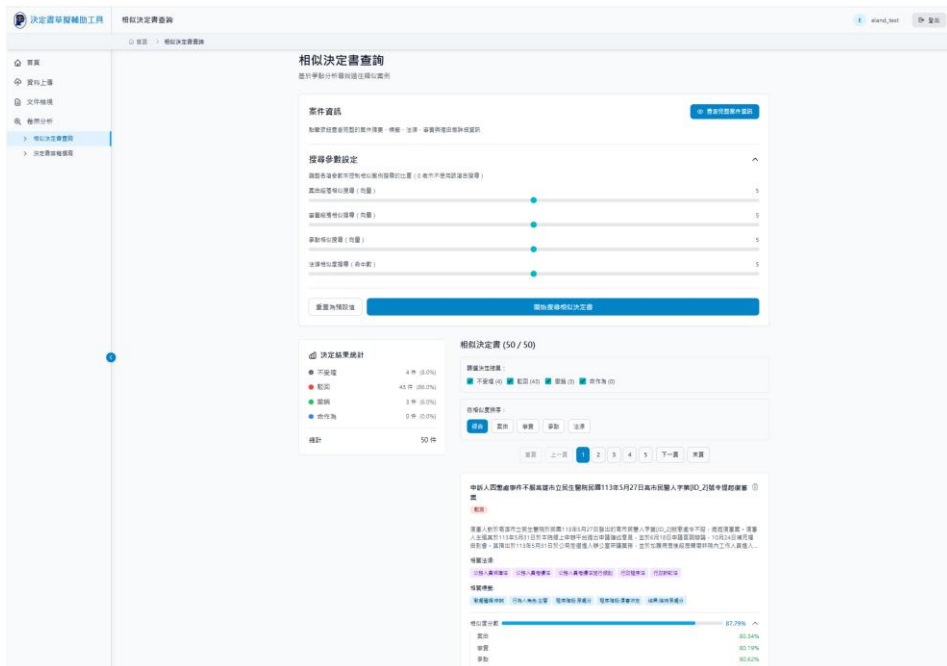


圖 13 相似案例檢索與多維度權重調整介面圖

（一） 向量語意檢索

捨棄傳統關鍵字檢索，改採先進之語意比對技術，精準找出法理邏輯最為貼近之歷史案件：

1. 向量資料庫運用：採用 Milvus 高性能向量資料庫，將當前案件之自動貼標結果、爭點、事實及理由轉化為高維度向量特徵。
2. 相似度量化評分：系統自動計算語意向量間的空間距離，回傳相似度評分，並依照關聯度由高至低呈現相關歷史案件。

（二） 搜尋參數設定與進階檢索功能

為了提升檢索結果的精確度與實務適用性，系統提供高度自訂義的搜尋設定介面，賦予使用者對檢索邏輯的主導權：

1. 多維度權重調整：透過「案由、事實、爭點、法源」比重滑桿（0~10 級），由使用者依據案情重點定義各項指標對搜尋結果的影響力，實現動態加權檢索。
2. 決定結果進階過濾：提供即時勾選功能，使用者可針對搜尋結果進行二次篩選，依據「不受理、駁回、撤銷、命作為」等決定類型過濾搜尋範圍，精準定位特定之實務前案。
3. 多元排序切換機制：檢索結果支援自訂義排序條件，可依據「綜合評分」或特定維度（如：案由、事實、爭點、法源）之相似度由高至低重新排序，協助使用者從不同法律維度審視參考案例。

（三） 判決趨勢之視覺化統計

為協助使用者快速掌握過往實務傾向，系統將搜尋結果進行量化分類與統計，並以色彩管理加強辨識度：

表 4 相似案例判決結果分類與視覺化標示說明表

判決結果	標示色彩	實務說明
不受理	灰色	代表案件因程序不符或不具備受理條件而結案
駁回	紅色	代表原處分經審查後維持，申訴人之請求被駁回
撤銷	綠色	代表原處分因程序或實體違誤而被撤銷
命作為	藍色	代表命令相關機關應針對該訴求採取特定行政作為

系統會針對上述結果自動計算案件總數與百分佔比，透過視覺化圖表協助使用者快速判讀該類案件在實務上的主流見解。

（四） 智慧容錯與人工過濾機制

系統內建嚴謹的品質控管功能，確保參考資料之純淨度：

1. 空結果友善引導：若未偵測到相似前案，系統將引導使用者調整權重參數或採取手動新增案例之替代方案。
2. 人機協作過濾：使用者可針對搜尋結果進行品質審核。若發現 AI 推薦之案例與本案實務關聯度不高，可點擊「垃圾桶」圖示予以刪除，以確保後續草稿生成引用之參考品質。

五、 五段式決定書草稿生成與匯出

為系統之最終產出單元，旨在將前述之「AI 卷宗分析」、「爭點判斷」與「相似案例見解」進行結構化整合，自動生成符合行政體例之五段式決定書草稿，大幅節省承辦人員初稿撰擬之時間。

（一） 決定書結構化自動填充

系統將 AI 分析後的結構化數據與 RAG 檢索結果進行深度整合，自動產製決定書草稿。其生成內容涵蓋以下核心段落：

1. 案由段 (Summary)：依據案件摘要與當事人屬性，將原始描述轉化為正式書面語句。本段落旨在以標準化方式呈現案件主軸，確保後續事實與理由段落能建立在一致的敘事基礎上。
2. 事實段 (Facts)：根據卷宗所提取之關鍵事實，自動編排時間、地點及行為經過。系統盡量貼近實際卷宗用語與公文格式，將紛雜資訊轉化為邏輯連貫且條理清晰之事實背景。
3. 理由段 (Reasoning)：系統透過 RAG 機制綜合相似歷史案件之判斷邏輯與適用法規，針對各項爭點進行法律涵攝。透過條列式論證，逐項說明救濟要件之符合性，確保生成結果具備可追溯之實務基礎。
4. 主文段 (Main Decision)：依據預期判決結果，自動產出標準化之主文句型，確保表述符合法律慣用格式與法定用語。
5. 結論段 (Conclusion)：系統彙整理由段所引用之法規與判讀結果，產出結論性語句，將法律依據與最終處分整合為一致的邏輯閉環，確保文書結構之嚴謹與完整。



圖 14 AI 自動生成之決定書段落草稿介面圖

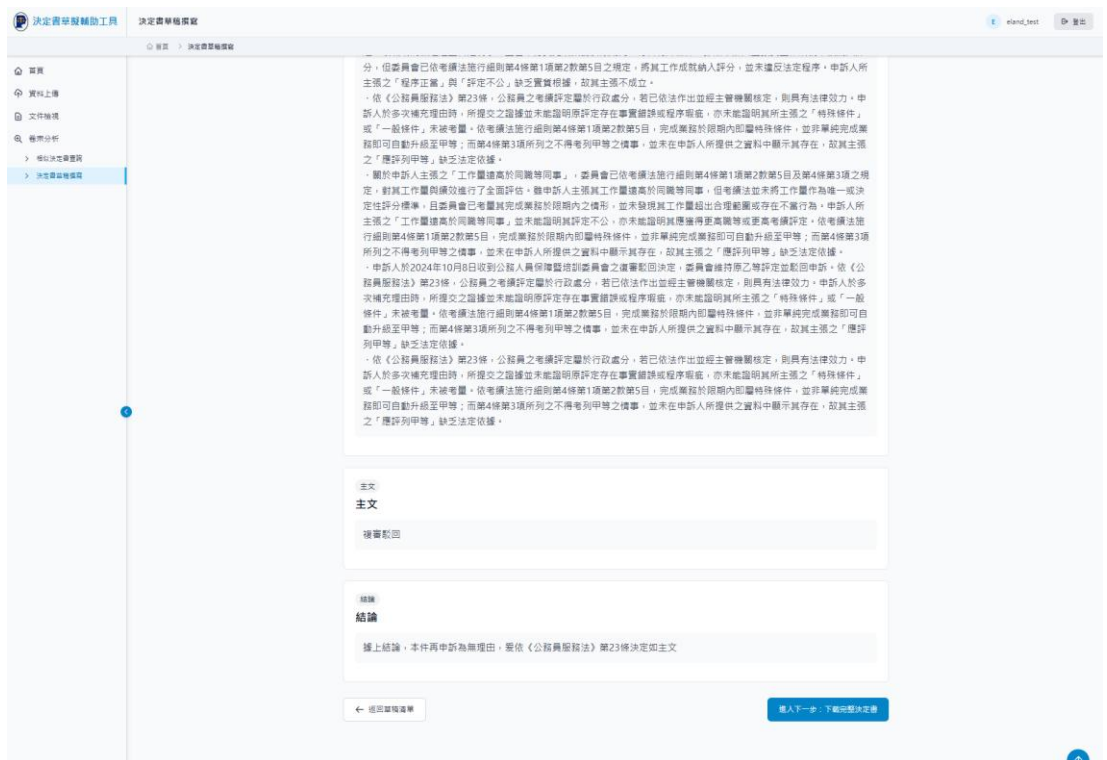


圖 15 決定書草稿之主文與結論生成結果介面圖

(二) 判決結果切換機制

系統具備靈活的「多立場動態論證」能力，以因應法律實務中可能存在之不同見解。當使用者於介面選取不同的判決結果時，

系統將依據選定立場，從 Milvus 檢索庫中重新提取對應之同向判例作為脈絡。透過 AI 的動態重寫技術，系統會自動調整「理由段」之論證方向與敘事邏輯，並同步校正「主文」與「結論」段落之句型，確保整份決定書草稿在不同法律觀點下，依然維持嚴謹的前後呼應與邏輯一致性。

（三） 文書自動化匯出

在完成草稿內容的審閱後，系統提供高效的文書轉換服務，確保從數位分析到實務產出的流程能無縫銜接。使用者可透過一鍵匯出功能，將系統生成的結構化內容產出為標準之 Word 格式檔案。此機制完整保留了決定書的五段式架構，將 AI 生成之初步成果轉化為可編輯之實體文件，讓使用者能隨後在熟悉的文書處理環境中，依據專業判斷進行最終的文字潤飾與細節微調。透過此種設計，系統在維持使用者既有作業習慣的同時，大幅提升了公文製作的整體效率。

六、 系統可同時支援使用者人數與文件處理時間說明

（一） 系統可同時支援使用者數量

本系統採用任務佇列（Task Queue）之非同步處理架構，使用者操作與後端文件解析流程彼此分離。使用者完成卷宗上傳後，系統會將文件解析任務提交至後端任務佇列，由系統依據可用運算資源進行排程處理。

因此，系統整體效能主要取決於後端硬體資源配置（如 CPU、GPU、記憶體與處理節點數量），而非單純受同時在線使用者數量限制。即使多位使用者同時上傳卷宗文件，系統仍可透過佇列機制依序或並行處理任務，不會直接影響使用者端之操作流程。

（二） 卷宗文件處理時間說明

卷宗文件自完成上傳至可供檢視（本文提取完成）之處理時間，主要取決於文件解析流程中的 OCR（光學文字辨識）處理時間。

由於卷宗文件多為掃描影像或影像型 PDF，系統需逐頁進行文字辨識與版面解析，因此整體處理時間與需進行 OCR 辨識之頁數具有高度關聯性；一般而言，卷宗頁數越多、影像品質越低或版面結構越複雜（如包含表格或手寫內容），所需處理時間亦相對增加。

（三） OCR 處理效能

依據目前測試環境之實測結果，在未使用 GPU 加速之情況下，單一卷宗之 OCR 辨識平均處理時間約為 6 小時，最大處理時間約為 18 小時以上。該處理時間主要受卷宗頁數及影像品質影響。

（四） AI 分析處理效能

在 OCR 文字提取完成後，系統將進一步執行 AI 語意分析流程（包含案件摘要、事實提取、爭點分析及法源關聯等）。目前測試環境使用 1.5 張 NVIDIA L40S GPU 進行任務處理，其平均分析時間約為 2 至 5 小時。

（五） 整體文件處理時間

綜合 OCR 與 AI 分析兩階段流程，在目前測試環境下，卷宗文件自完成上傳至本文提取及分析完成之整體平均處理時間約為 8 至 16 小時，實際時間仍會依卷宗頁數、影像品質與系統資源配置而有所差異。

（六） GPU 加速之效能推估

目前測試環境之 OCR 辨識流程尚未使用 GPU 進行推論加速。未來若於正式系統建置階段導入支援 GPU 推論之 OCR 架構並搭配 GPU 運算資源（如 NVIDIA L40S 或 H100），在相同文件條件下，OCR 辨識處理時間預期可顯著縮短。

依據目前測試結果與 GPU 推論能力推估，OCR 平均處理時間可望縮短至約 1 至 3 小時，最大處理時間亦可望降低至約 3 至 8 小時。惟實際效能仍將受文件頁數、影像品質、OCR 模型架構及系統資源配置等因素影響，相關數值將於正式系統部署及效能測試後進一步確認與優化。

七、 重新解析機制與錯誤回饋說明

當使用者於系統介面點擊「重新解析檔案」時，系統會將該文件重新建立解析任務並重新加入任務佇列（Task Queue）進行排程處理。此過程與初次解析流程相同，因此會再次使用 OCR 解析及 AI 分析等相關運算資源。

由於本系統採用非同步任務佇列架構，重新解析任務僅會被加入佇列並依序或依可用資源進行處理，不會中斷或影響其他已在執行中的任務。當系統同時存在多個處理任務時，系統將依據運算資源狀況進行排程，因此對其他使用者之操作影響有限，主要僅可能反映於任務等待時間上。

若解析程序於處理過程中發生異常，系統將於處理狀態欄位標示為「處理失敗」，並提供重新解析功能供使用者再次啟動處理程序。此外，系統亦會於後端記錄相關處理狀態與錯誤訊息，以利系統管理者進行問題排查與後續系統優化。於未來正式系統實作階段，亦可依實際需求進一步強化錯誤訊息顯示與處理紀錄機制，以協助使用者更快速判斷問題來源與後續處理方式。

肆、 測試結果與可行性分析

一、 測試目標與測試原則

本專案測試以「概念驗證 (PoC)」為核心，目的在於驗證系統是否能依設計流程穩定運作，並於行政申訴案件處理情境中，提供具實務參考價值之 AI 輔助分析結果。

測試重點不在於評估 AI 模型之最終正確率，而在於確認系統各階段功能是否可正常執行，並確保 AI 產出內容具備可介入、可修正與可回溯之特性。

二、 系統功能與流程驗證成果

本案已完成以下核心流程之功能驗證：

(一) 卷宗文件上傳與處理流程

驗證系統可成功接收 PDF 格式文件，並依序完成 OCR 文字擷取、去識別化、語意分析與向量化處理，處理狀態可即時回饋供使用者查詢。

(二) 案件分析與結構化成果產出

驗證系統能將非結構化卷宗內容，轉換為案件摘要、事實描述、爭點分析與法律依據建議等結構化資訊，並於前端介面中清楚呈現，供使用者檢視與判斷。

三、 AI 分析輔助成果說明

於測試過程中，系統能依案件內容，自動整理案件重點摘要，並列示需進一步判斷之關鍵事項，作為承辦人員審議時之輔助資訊。

同時，系統可依案件分析結果，自動檢索相似歷史案例，提供相似度指標與案例內容摘要，協助使用者快速理解案件可能之判斷方向與實務背景。上述 AI 產出結果僅作為輔助參考，並未直接形成任何具決定效力之結論。

四、 人工介入與修正機制驗證

本案亦驗證系統於關鍵分析節點提供人工介入之可行性，包括：

1. 承辦人員可檢視並調整系統萃取之爭點內容。
2. 可補充或修正案件重點與判斷說明。

測試結果顯示，系統能有效整合 AI 推論與人工輸入內容，並於後續草稿生成階段反映使用者之判斷意見。

五、 系統穩定性與例外處理測試

針對實務使用情境中可能發生之例外狀況，本案進行基本測試驗證，包含：

1. 檔案格式不符或檔案損毀時之錯誤提示。
2. OCR 或 AI 分析失敗時之狀態標示與重新處理機制。
3. 長時間處理任務下，前端介面仍可即時回應並顯示處理進度。

透過非同步任務與狀態管理設計，系統於測試情境下可維持穩定運作，並提供明確之處理回饋。

六、 測試成果與可行性分析總結

整體測試結果顯示，本系統已成功驗證以 AI 技術輔助行政申訴案件分析與決定書草稿產製之可行性。系統於設計上保留充足之人工介入空間，使 AI 成為輔助工具而非取代專業判斷，符合本案 PoC 之預期目標。

相關測試成果可作為後續評估正式導入、功能擴充或流程優化之重要參考依據。

詳細相關內容可參考附錄一之二、測試結果。

七、 測試資料來源與資料安全控管機制說明

本案概念驗證（PoC）測試過程中，系統主要以卷宗文件作為分析輸入資料，並於驗證階段搭配既有決定書內容進行比對，以評估系統產製之案件摘要、爭點整理及決定書草稿與實際決定書之差異情形。相關比對結果再由研究團隊以人工方式進行複查，以確認 AI 輔助分析之合理性與實務參考價值。

考量案件資料可能涉及敏感資訊，本系統於測試環境中採用封閉環境地端部署之大型語言模型進行推論運算，所有資料均於內部環境處理，未透過外部雲端服務或第三方 API，以降低資料外洩或資安風險。

此外，測試資料於匯入系統前亦進行去識別化處理，針對可能涉及之個人姓名或可識別資訊進行適當遮蔽，在不影響案件內容分析之情況下，降低個資識別之可能性，以確保測試資料之安全性與客觀性。

八、 AI 分析結果差異說明

於測試過程中，部分案例之 AI 分析結果與本會既有決定書內容在表述完整性及用語精確度上仍存在差異，例如未完全涵蓋復審人主張內容，或於爭點整理及案件事實描述上之敘事順序與本會既有撰擬方式未盡一致。上述情形主要反映大型語言模型於法律文本分析時，仍可能受到語料結構、案件複雜度及上下文理解能力等因素影響。

本案係以概念驗證（PoC）為主要目的，重點在於驗證系統流程之可行性，包括卷宗文件解析、語意分析、案件結構化整理及決定書草稿生成等功能是否能順利運作。因此，本次測試並未針對特定法律領域語料進行模型微調（Fine-tuning）或專門化調校，而是以通用模型能力驗證整體系統架構之可行性。

未來若正式導入系統，仍可透過多項技術方式進一步提升分析品質，例如建立專門法律語料庫、導入模型微調（Fine-tuning）、優化提示詞設計（Prompt Engineering）或強化檢索式生成（RAG）之資料來源等，以改善案件事實描述之完整性、爭點整理之精確度及法律用語一致性。

因此，上述案例所反映之差異，主要屬於 AI 模型表現之優化議題，並不影響本案於系統流程與技術架構層面之可行性驗證結果。

九、 測試案件處理紀錄補充說明

依審查意見，本案原規劃提供約 30 件測試案件之處理紀錄。後續於保障法制處另提供本會審議之保障事件案件資料（僅含卷證資料部分），並請本案依該資料重新執行系統流程以產生實際執行結果。

考量系統 OCR 與 AI 分析流程需一定運算時間，本次係依新提供之案件資料重新執行測試流程，並彙整 27 件案件之處理紀錄，包含檔案上傳時間、解析啟動與完成時間等資訊。相關紀錄已整理於附錄四：執行效能紀錄及結果，供檢視系統運作流程與處理效能。

伍、結論與建議

考量行政申訴案件審議流程具高度專業性與判斷彈性，本系統未來發展方向建議以「分段分析、分流程輔助、人機協作」為核心設計原則，避免將決策過程過度自動化，而是透過 AI 提供多層次輔助資訊，保留承辦人員於各階段進行判斷、修正與補充之空間。

（一）導入分段式、流程化之案件分析機制

未來系統建議依實際工作流程，將案件處理拆解為多個分析階段，並於各階段產出對應之輔助結果，而非一次性生成完整分析或草稿內容。

以新案卷宗為例，系統可於卷宗讀入後，先行產出以下初步成果，供使用者檢視與判斷：

- 案件重點摘要與內容整理：由系統整理案件背景、主要事實與初步爭點摘要，作為承辦人員快速理解案件全貌之參考。
- 需判斷重點之列示：針對案件中可能涉及判斷之關鍵事項（如程序爭議、事實認定差異、適法性問題等），以條列方式呈現，協助承辦人員聚焦審議重點。

透過此一分段式設計，可降低承辦人員一次面對大量 AI 產出內容之負擔，並使 AI 分析更貼近實務審議流程。

（二）強化人工介入與判斷輸入之設計

未來系統建議在各分析階段，提供承辦人員可明確介入之操作介面，包含但不限於：

- 對系統列示之爭點進行確認、補充或刪修。

- 針對特定爭點輸入承辦人員之初步判斷結果或審議方向。
- 於關鍵判斷節點提供多種選項，或由使用者自行輸入判斷說明。

上述人工輸入結果可作為後續 AI 分析與草稿生成的重要參考依據，使系統產出內容不僅來自模型推論，亦能反映實務人員之專業判斷。

（三） 擴充判例參考來源與人工指定機制

在相似案例參考方面，除既有以語意向量方式自動搜尋相似判例外，未來建議擴充以下功能：

- 人工指定參考判例：允許承辦人員自行選擇或指定特定歷史案件作為本案之參考案例，補足自動檢索可能未能涵蓋之實務情境。
- 混合式參考機制：將系統自動檢索之相似案例與使用者指定案例併同提供給 LLM 作為分析與生成之依據，提升產出內容之實務貼近度。

此一設計有助於將實務經驗有效納入 AI 輔助流程中，而非僅依賴模型自動比對結果。

（四） 決定書草稿之人機協作生成模式

未來系統建議採取「AI 初稿＋人工調整」之協作模式進行決定書草稿產製。系統可整合以下資訊來源進行草稿生成：

- LLM 對卷宗內容之分析結果
- 自動與人工指定之參考判例
- 承辦人員於前述各階段所輸入之判斷意見

透過整合 AI 推論與人工輸入內容，產出具結構性之決定書草稿，供使用者進一步修改、調整用語與補充說明。最終內容仍由承辦人員負責確認與定稿，以確保法律適切性與實務一致性。

（五） 建立可回饋之案例累積與知識擴充機制

於案件完成並確認最終決定書內容後，未來系統建議將該案件及其相關分析結果，納入系統之參考資料庫，作為後續案件分析與檢索之基礎。

透過持續累積經人工確認之案例資料，可逐步提升系統於實務場景中的輔助價值，並形成正向循環之知識成長機制，而非一次性分析工具。

（六） 發展方向總結

整體而言，系統未來發展不以完全自動化決策為目標，而是以「可介入、可修正、可回溯」之輔助型 AI 系統為核心方向，透過分段分析與人機協作設計，使 AI 成為承辦人員的輔助工具，而非取代專業判斷之角色。

(七) 未來系統實作所需系統架構圖及軟體規格

本案為概念驗證性質，未實際導入正式營運環境，惟依本次 PoC 驗證成果，已可初步歸納未來正式系統建置所需之系統架構方向與軟硬體設備配置原則，作為後續規劃參考。

1. 系統整體架構說明

本系統採用分層式與模組化之服務架構設計，作為未來實際導入時之參考架構。整體架構可區分為「前端 UI 層」、「應用服務層」、「資料層」及「AI 模型層」，各層級職責明確，並透過標準化介面進行串接，以確保系統具備良好之可維護性、可擴充性及部署彈性。

系統流程由使用者透過前端介面進行操作，經由入口服務轉交應用服務層處理，核心分析與運算任務則透過非同步機制派送至 AI 模型層執行，分析結果再回存至資料層供後續查詢與應用。整體設計以模組解耦為原則，各元件均可依實際部署環境與既有資訊基礎設施進行替換，不限定特定廠牌或單一技術實作。

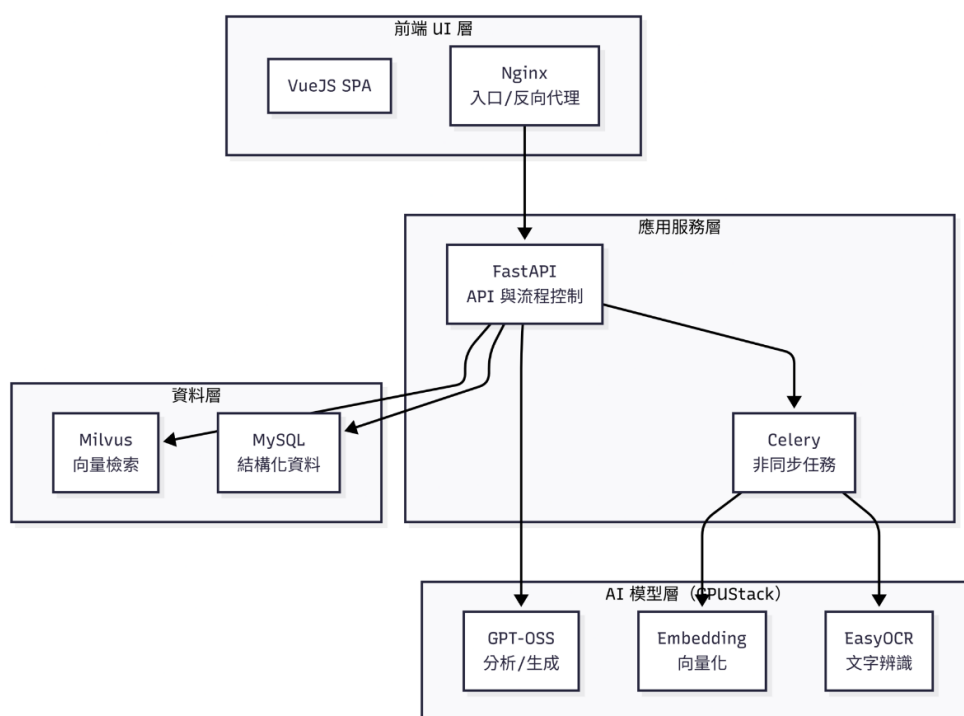


圖 16 系統整合服務組件與數據流向圖

本系統架構圖屬概念性與參考性設計，主要用以說明各功能模組之角色分工與資料流向，實際導入時可依機關資安政策、硬體資源及預算條件進行調整。

2. 軟體架構與開放性規格說明

本系統軟體架構以開放原始碼與通用框架為設計原則，避免依賴單一廠商之專屬技術，確保後續導入與維運具備彈性。各核心功能模組均可採用具相同性能與功能之替代方案進行實作，例如：

- 前端服務：可採用主流 Web 前端框架與標準瀏覽器技術
- 後端 API 服務：可採用支援 RESTful API 通用應用框架
- 非同步任務處理：可採用支援背景任務排程之通用機制
- 關聯式資料庫：可採用主流關聯式資料庫系統
- 向量資料庫：可採用支援高維度語意檢索之向量搜尋系統
- AI 分析與模型服務：可依實際需求選用適合之語言模型、向量化模型及 OCR 辨識引擎

上述組件均屬開放性與可替代設計，實際導入時可配合機關既有系統環境進行選型。

(八) 未來系統實作所需硬體設備規格與參考價格

依據本案概念驗證 (PoC) 測試結果，未來正式導入系統需建置具備 AI 推論與文件解析能力之運算環境，以支援卷宗 OCR 解析、語意分析、向量化處理、相似案例檢索及決定書草稿生成等功能。

整體系統架構建議以 GPU 運算伺服器作為核心運算節點，負責大型語言模型推論、文件 OCR 處理與 AI 分析任務，並搭配資料庫與儲存設備，以確保系統於多任務處理情境下仍能維持穩定運作。

考量本案測試環境之實際運作結果與 AI 推論運算需求，未來系統建置時建議之 GPU 配置規模如下：

■ 最低建議配置

建議配置 1 張 H100 GPU 搭配 1 張 L40S GPU，或 3 張 L40S GPU。此配置可支援基本之卷宗 OCR 文件解析、AI 語意分析與相似案例檢索任務，適用於使用人數較少或案件處理量較低之環境。

■ 建議配置（較佳運算效能）

建議配置 2 張 H100 GPU 以上之 AI 運算伺服器，可提供較高之 AI 推論與文件解析處理能力，並可在多案件同時處理情境下維持較佳之系統效能與處理效率，較適合正式系統導入及長期運作之環境。

為利未來系統建置之預算規劃，實際設備規格與採購數量，仍須依未來系統部署規模與使用需求進行評估與規劃。相關設備之詳細規格與價格參考資料，請參見附錄三：硬體價格參考表。

■ 硬體設備建議規格（參考型號，非指定）

為支援未來系統於實務環境中進行卷宗解析、AI 分析與語意檢索等運算需求，以下列出建議之硬體設備規格與同級參考型號，供實作規劃時評估使用：

表 5 建議 GPU 硬體規格參考表

設備類型	參考廠牌型號（非指定）	規格說明
GPU	NVIDIA L40S	48GB VRAM 適合中大型模型推論
	NVIDIA H200	高頻寬記憶體（HBM3e），適合 大型語言模型推論與高吞吐運 算
	NVIDIA H100	高階推論與多模型併發

表 6 建議伺服器設備參考表

設備類型	參考廠牌型號（非指定）	規格說明
Server	Dell PowerEdge R760XA	2U 雙處理器 AI 伺服器，支援 GPU 擴充
	HPE ProLiant DL380 Gen11	2U 雙處理器企業級伺服器，支 援 GPU 擴充
	Lenovo ThinkSystem SR650 V3	2U 雙處理器高效能運算伺服 器，支援 AI / GPU 工作負載