pdf-md — 智慧 PDF 轉 Markdown 工具

將您的 PDF 文件轉換為乾淨、結構化的 Markdown，以深度學習 AI 驅動 — 打造 AI 知識庫與 RAG 應用的關鍵第一步。

為什麼 AI 需要高品質的文件資料？

現代 AI 助理與企業智慧客服，核心技術是**檢索增強生成（RAG，Retrieval-Augmented Generation）** — AI 在回答問題之前，先從您自己的文件知識庫中搜尋相關資訊，再據此生成答案。因此，知識庫的資料品質，直接決定了 AI 每一次回答的準確性。

問題在於：**企業的核心知識大多儲存在 PDF 中** — 報告、技術手冊、研究論文、財務報表 — 而 PDF 對機器極不友善。傳統的 PDF 文字擷取方式會破壞表格結構、丟失公式、錯亂欄位順序、失去標題階層。當知識庫中充斥著破碎、混亂的資料，AI 的回答也將跟著出錯。

pdf-md 解決這個問題。 它以多個深度學習模型理解每一頁的視覺版面，將文字、表格、公式、圖表等每一個元素，轉換為結構完整、語意清晰的 Markdown — 讓向量資料庫得以精準、完整地建立索引，讓 AI 找到正確的答案。

---

pdf-md 是什麼？

pdf-md 是一套 AI 驅動的文件轉換服務，能讀取任何 PDF 或圖片檔案，輸出包含標題、表格、數學公式、圖表與圖片的結構化 Markdown — 所有內容都忠實還原並格式化，直接供 AI 系統使用。

RAG 為什麼需要這類資訊？

RAG 流程分為兩個階段：**建立索引**（事前準備）與**檢索**（查詢時）。pdf-md 的輸出品質在兩個階段都至關重要。

建立索引時 — 更好的分塊，更準確的向量

向量資料庫會將文件切割為語意片段（chunk），再對每個片段進行向量嵌入（embedding）。如果片段內容殘缺不全 — 表格列被切斷、標題與正文分離、公式變成亂碼 — 嵌入向量就會產生偏差，AI 在需要時找不到正確的片段。

pdf-md 產出的 Markdown 能確保：

- **標題結構完整保留**（`#`、`##`、`###`），讓分塊邊界對齊文件的邏輯章節，而非任意字元數
- **表格以完整結構呈現**（HTML `<table>`），同一張表的資料列永遠不會被切斷
- **公式以 LaTeX 表示**（`$$E = mc^2$$`），而非亂碼的 Unicode 字元
- **圖片說明文字與圖片保持對應**，上下文不會流失

結果：您的嵌入模型看到的是連貫、有意義的語意片段，儲存的向量能精準代表每個章節的含義。

檢索時 — AI 找到並引用正確的答案

使用者提問時，RAG 系統取回最相關的片段，作為語言模型的參考上下文。如果這些片段包含：

- 結構完整的**表格** → AI 能跨欄讀取，回答「亞洲區第三季營收是多少？」
- 以 LaTeX 表示的**公式** → AI 能理解方程式，而非看到 `ï€r2` 這樣的亂碼
- **圖表轉換後的數值表格** → AI 能引用圖表中的具體數字
- 保留完整的**章節標題** → AI 能理解段落隸屬於文件的哪個部分

文件解析品質差，AI 就找不到答案，甚至產生幻覺。pdf-md 的精準解析，讓 AI 檢索到真實資訊並正確引用。

核心功能

版面感知的精準擷取

pdf-md 不是單純傾印原始文字。它使用頂尖的版面偵測模型（**PP-DocLayout**），識別每一頁的每一個區域 — 標題、正文、說明文字、表格、圖表、公式、圖片 — 再分派給對應的專用 AI 模型處理。輸出的 Markdown 忠實還原文件的邏輯結構，正是 RAG 分塊器所需要的格式。

完整的表格理解

表格承載了企業文件中最高密度的資訊 — 財務數字、產品規格、比較矩陣。傳統 PDF 擷取工具往往讓欄位錯亂、合併格消失。pdf-md 逐一偵測每個儲存格，正確處理合併格（`rowspan` / `colspan`），輸出乾淨的 HTML `<table>`。RAG 系統取回表格片段時，AI 拿到的是完整的結構化資料，而非一串混亂的數字。

您還可以透過 Excel 定義**自訂表格範本**，對固定格式的報表強制套用一致的欄位標籤，讓 AI 讀取到更具預測性的表格內容。

數學公式辨識

以純文字擷取的公式 — `E=mc2`、`∫f(x)dx` — 對語言模型而言毫無意義。pdf-md 使用 **PP-FormulaNet** 辨識行內與獨立公式，輸出標準 LaTeX（`$...$` / `$$...$$`）。現代大型語言模型原生理解 LaTeX，AI 因此能理解並解釋從文件中取回的方程式。

圖表轉數值表格

PDF 中的圖表是不透明的圖片 — RAG 系統索引原始 PDF 時只能拿到圖片檔，無法查詢其中的數據。**PP-Chart2Table** 解析長條圖、折線圖等視覺圖表，將底層數據轉為 Markdown 表格。當使用者問「銷售趨勢圖顯示什麼？」，AI 能直接引用具體數字作答。

高品質 OCR 文字辨識

由 **PP-OCRv5 Server** 驅動，pdf-md 對印刷與掃描文字達到接近人工的辨識準確率，支援多語言。OCR 每辨識錯一個字，就是 RAG 系統永遠無法檢索到的一個字。可調整的信心門檻讓您針對不同文件類型微調品質與召回率的平衡。

浮水印移除

浮水印覆蓋在文字上，會汙染 OCR 結果 — 讓重要詞彙在向量資料庫中變成無法讀取的雜訊。啟用內建的浮水印移除功能，在 OCR 執行前預先清理每一頁，確保知識庫中的文字準確完整。

專為機器消費設計的輸出結構

pdf-md 產出的 Markdown 針對機器消費而非單純人類閱讀進行優化：

- 文件階層以 Markdown 標題表達 → RAG 分塊器在有意義的邊界切割
- 表格輸出為 HTML 而非 Markdown 管線語法 → 複雜的合併格版面在分塊後依然完整
- 公式以 LaTeX 區塊呈現 → 可被大型語言模型識別與解析
- 圖片儲存為獨立檔案並附帶描述性檔名 → 多模態 RAG 流程可分別建立索引

無需命令列 — 完整 Web 介面

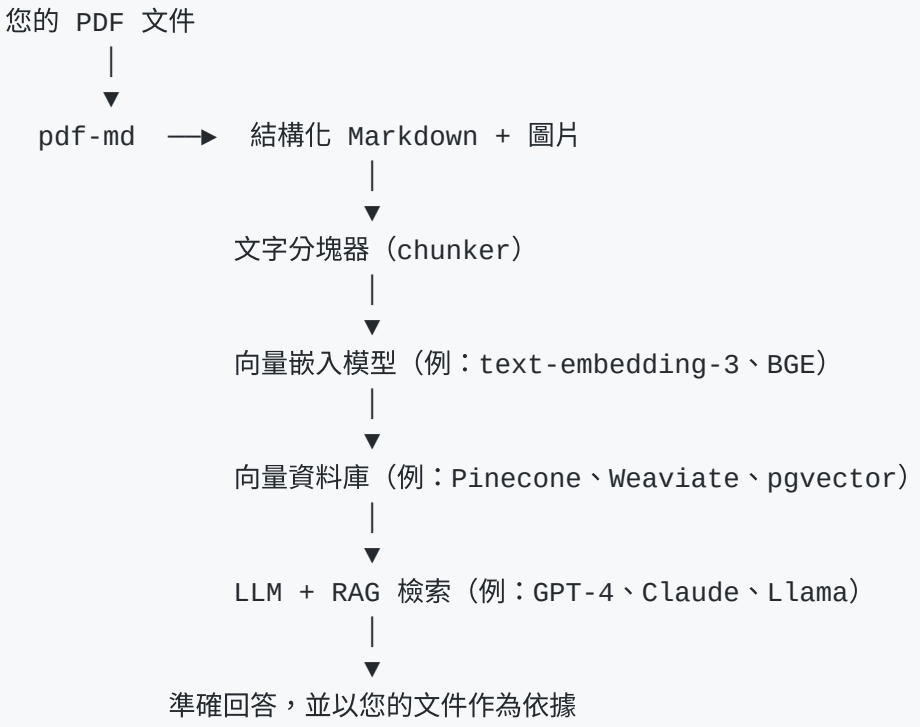
透過瀏覽器上傳 PDF、即時監控處理進度、下載 Markdown 結果 — 全程在 Web 介面完成，無需伺服器權限或終端機操作。

使用流程

1. **上傳** 透過 Web 介面上傳您的 PDF 或圖片。

2. **設定** 選擇性調整參數：頁碼範圍、DPI、浮水印移除、公式解析等。
3. **送出** 任務進入佇列，由 GPU 加速的深度學習模型在背景處理。
4. **下載** 取得包含 Markdown 檔案與所有擷取圖片的 ZIP，直接載入您的向量資料庫或知識庫。

RAG 資料流程



處理模式

模式	最適合	行為
非同步（建議）	大型 PDF（10 頁以上）	上傳後離開 — 在任務列表查看進度，完成後下載
同步	小型 PDF（10 頁以下）	在瀏覽器等待，處理完成後自動觸發下載

功能一覽

功能	RAG 價值	技術細節
版面偵測	正確的分塊邊界	標題、正文、說明文字、表格、圖表、公式、圖片
OCR 引擎	精準文字供嵌入使用	PP-OCRv5 Server，多語言，高準確率
表格解析	結構化資料可被檢索	儲存格偵測、合併格處理、乾淨 HTML 輸出
公式辨識	LLM 可讀取的方程式	行內與區塊 LaTeX，由 PP-FormulaNet 驅動
圖表解析	圖表數據可被查詢	圖表 → 結構化 Markdown 表格，由 PP-Chart2Table 驅動
自訂範本	一致的欄位標籤	Excel 驅動的表格範本，適用固定格式報表
浮水印移除	乾淨文字供建立索引	灰階閾值預處理
輸出格式	直接接入 RAG 流程	Markdown + 圖片 ZIP
處理方式	因應企業規模	非同步佇列（背景）或同步（立即下載）

適合的使用者

- **AI / 資料團隊** — 基於企業文件庫建立 RAG 聊天機器人或智慧搜尋系統
- **企業數位化團隊** — 將 PDF 文件庫轉換為 AI 可消費的知識庫
- **技術文件人員** — 將參考 PDF 轉換為可編輯、可索引的 Markdown 文件
- **研究人員** — 大規模從學術論文中擷取表格與公式供 AI 分析

立即開始

開啟 Web 介面，將您的第一份 PDF 拖入上傳區，點擊**開始處理**。您的結構化、RAG 就緒的 Markdown 幾分鐘內即可下載。

進階設定 — 自訂 DPI、頁碼範圍、公式 DPI 上採樣、表格範本 — 請參閱應用程式內的設定面板或 API 文件。