操作手冊驅動的檢索增強生成（RAG）系統報告

作者：111502562

日期：2025-05-15

本報告分為四個章節，分別闡述文檔處理、提示工程、系統架構與互動設計，並引用相關學術與實踐文獻。

1. 文檔處理 - 為 RAG 系統準備操作手冊 (20%)

要使操作手冊（PDF、掃描影像、網頁）可用於 RAG，需完成以下步驟：

1. 提取及轉換多格式文本
    - **PDF to Markdown**：使用 PyMuPDF 提取文字並保留標題結構，轉為 Markdown。2
    - **OCR 影像處理**：對掃描或掃描影像版 PDF，採用 Tesseract OCR 或 PaddleOCR 進行文字辨識，再轉為純文本。6
    - **網頁抓取**：以 Selenium + BeautifulSoup 讀取 DOM，再過濾廣告與頁首頁尾，保留章節化內容。7
2. Chunking 策略以獲取最佳檢索粒度
    - **固定長度分塊**：每塊 500 字元、重疊 50 字元，兼顧上下文與檢索效率。4
    - **語義分段切分**：先依章節、標題切分，再對過長段落細分，確保一塊內容聚焦單一主題。
3. 向量資料庫構建
    - 使用 SentenceTransformer（如 all-MiniLM-L6-v2）將每個文本塊編碼為向量，捕捉語義關係。
    - 選擇 ChromaDB 做本地化持久化，支持快速相似度檢索，並於批次（50 塊）寫入以降低記憶體峰值。
4. Metadata 處理
    - **章節標籤**：保留檔名、章節名稱、chunk_id，回傳結果時可定位來源。

- **時間戳記**：記錄上傳與索引時間，用於資料版本管理與過期內容過濾。
- **自定欄位**：可附加作者、文件類型、語言等提升相關性排序。

2. 提示工程 - 為 RAG 瀏覽代理設計高效 Prompt (25%)

2.1 系統級提示

- 定義角色與行為約束：
- 設置語言、格式與禁止範圍，確保一致性與安全性。3

2.2 任務級提示

- 根據用戶目標動態拼接。
- 使用 Instruction Chaining 將多步驟拆分多輪，降低單次提示長度並提升模型集中度。5

2.3 檢索內容整合

- **In-Context Examples**：在 prompt 中嵌入典型 Q&A 範例，提供上下文比較參考。
- **Rescoring**：對多個檢索結果做二次排序，挑選 Top-K 內容插入提示，以減少冗餘與衝突。4

2.4 長文管理與幻覺防範

- 控制 prompt 總 Tokens，對超長手冊內容做摘要或抽樣引入。
- 設置 “如果無法找到答案，請回覆‘查無相關信息’”，避免模型憑空生成。
- 使用 Retrieval Feedback Loop，若生成與檢索不一致，自動觸發重檢索或補充查詢。

3. 系統架構 - 查詢驅動的 RAG 任務完成 (25%)

3.1 端到端架構

[用戶查詢]

↓ 嵌入 (Retriever)

[向量資料庫] - Top-K Chunks → [Prompt Engineering] → [Perplexity API] (Generator)

↓ ↓

[檢索結果後處理] ←--- Summarization/Rerank ---→ [可執行步驟輸出]

3.2 動態檢索觸發

- 代理生成中間意圖（Interim Intent），用於判斷何時呼叫 query_vector_db。
- 檢索語句包括當前動作上下文（如「正在填寫表單」），提升相關性。

3.3 後處理技術

- **Reranking**：結合 TF-IDF 關鍵詞匹配與向量相似度加權，提升 Top-K 精度。
- **Summarization**：對多塊文本做快速摘要，輸入 Generator 前減少提示長度與噪音。

3.4 結合生成

- 在 prompt 中嵌入檢索片段，並以「Step n of N」標示進度，指引用戶或代理專注當前步驟。
- 生成結果附帶來源與章節標籤，方便後續驗證與追溯。

4. 互動設計與增強建議

- **Step Tracker**：在回應中明確顯示當前步驟與總步驟數（如 **“Step 3 of 7”**），幫助代理保持任務流程感。
- **Instruction Cue**：在多步任務中，附加行動提示（如 **“接下來請專注於→ 輸入付款資訊”**），避免上下文跳脫。
- **Adaptive UI**：根據檢索結果的重要性動態調整提示字體大小或高亮，強調關鍵指令。
- **示例任務**：在多份手冊間同時檢索，完成「多設備安裝流程」等複雜任務，展示 RAG 系統跨文件整合能力。

結語

本系統通過結合先進的文檔處理、精細的提示工程與動態檢索生成架構，實現了基於操作手冊的智能瀏覽代理，顯著提升了任務完成的準確度與可操作性。