國立台北大學資訊工程學系專題報告

結合圖形擷取增強生成技術與小型語言模型生成法律起訴狀
Integrate graph RAG and SLMs for Drafting Legal Complaints

專題組員：黃星揚、陳彥霖、沈劭丞、丘觀仁

一、摘要

現今，隨著人工智慧技術的快速發展，其應用領域日益拓展，涵蓋法律、醫療及教育等多個領域。對普通民眾而言，法律領域的專業門檻通常較高，導 致其在撰寫具法律效力的民事起訴狀時面臨困難。過去已有研究嘗試透過商用大型語言模型（LLM）與提示詞技術來處理特定法律任務，然而這些方法都是 引用非本國的法律，對台灣法律文本的支持較為有限，且伴隨一定的隱私風險、難以解釋生成的問題。為解決上述問題，本計畫提出一套專注於生成中文 民事起訴狀的端對端系統。該系統結合小型語言模型（SLM）與圖形擷取增強生成技術(Graph RAG)，旨在提供一個高效、可靠且符合隱私需求的解決方案。此外，本研究將以司法院法學資料檢索系統與法官學院教材為依據，構建出中文民事起訴狀數據集，進一步補足現有資源的不足。

二、簡介

民事起訴狀撰寫具高度專業門檻，對缺乏法律知識的民眾而言是一大挑戰。鑑於台灣律師人力資源相對短缺，法律資源分配不均，本研究著眼於開發一套能協助民眾撰寫合法起訴狀的系統。透過自然語言處理（NLP）與檢索增強生成技術的結合，本系統旨在提供一個具可靠性與隱私保障的法律文本生成解決方案，以回應民眾對高品質法律協助的迫切需求。

三、專題進行方式

本專題計畫的核心目標是利用小型語言模型，如 Project TAME 所開發之 llama-3-taiwan:8b-instruct 和 Google 的 gemma3:27b，為交通事故人身傷害的案例提供訴訟書狀的自動生成。圖 3-1 為本計畫的系統架構，應用不同的因子（如請求權人傷害程度、職業、年齡、性別業等），輸入指令進語言模型，因而生成出相對應的訴訟狀。因此本計畫有助於法律專業人士更迅速、精確地提供必要文件，進而提升整體司法效率。以下為詳細的研究方法和步驟：

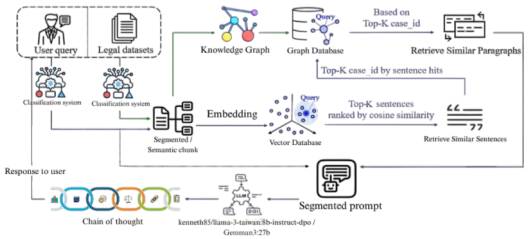

圖 3-1. 整體架構

1. 使用者輸入

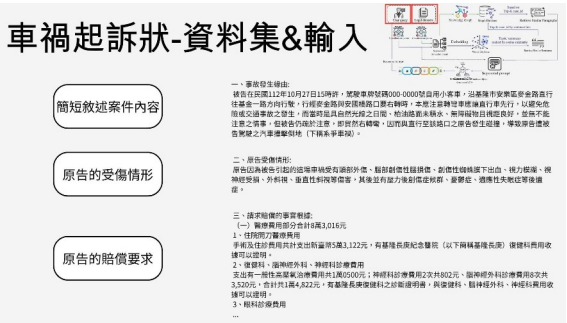

圖 3-2. 使用者輸入範例

圖 3-2 為使用者輸入之範例，使用者需提供三個關鍵資訊作為起訴狀生成之依據，分別為：案件發生經過、原告之受傷情形，以及具體的賠償請求內容（如醫療費、慰撫金等）。透過結構化填寫方式降低法律專業門檻，使非專業人士亦能有效輸入符合法律文書撰寫邏輯的資訊。

2. 資料庫

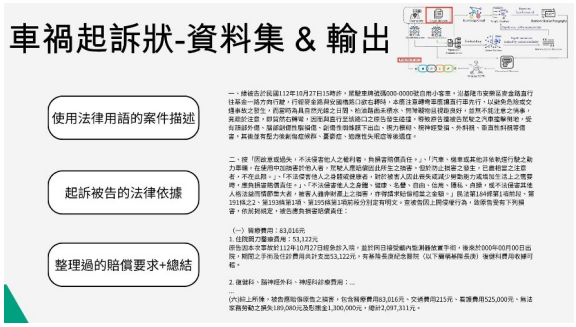

圖 3-3. 資料庫中的起訴書範例

系統蒐集並處理約 3,000 筆如圖 3-2 和圖 3-3 所示的民事案件資料，並以如圖 3-4 所示之知識圖譜（Knowledge Graph）形式進行結構化儲存。

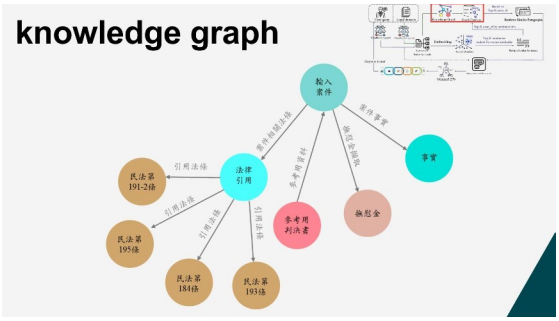

圖 3-4. 民事案件在知識圖譜中保存的格式

透過節點與邊的設計，精確表示案件中的法律實體（如法條、事實、損害項目等）與其關聯性。同時，將部分關鍵節點轉換為語意嵌入向量（Semantic Embeddings），儲存至向量資料庫中，以支援高效的相似案例檢索與語意查詢，有效提升系統於生成前階段的資訊召回精準度與上下文關聯性。

3. 資料預處理

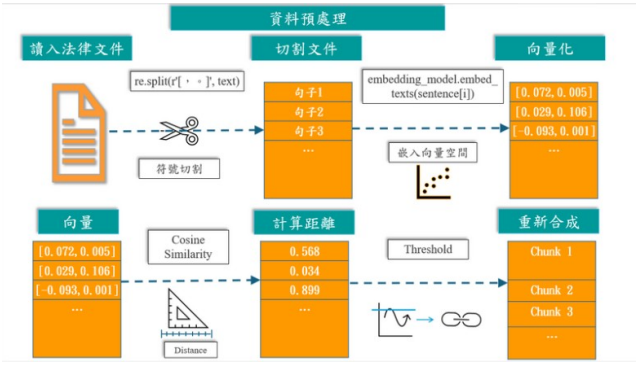

圖 3-5. 資料預處理流程

本研究採用基於語義相似度的智能文本分塊技術對 3000 筆法律資料進行預處理。首先，系統將每筆資料按照中文標點符號（逗號、句號）進行句子切分，並為每個句子建立索引(index)。接著，利用嵌入模型(embedding model)將所有句子轉換為向量(vector)表示，通過計算相鄰句子間的

餘弦相似度(cosine-similarity)來衡量語義連貫性。系統根據設定的百分比閾值(threshold)識別語義轉折點，當相鄰句子的相似度低於閾值時，視為潛在的分塊邊界。

由於分塊是動態決定的，在分塊過程中，系統同時考慮語義連貫性和文本長度限制，確保每個文本塊在限定的字數之間，避免產生過短或過長的片段。這種動態分塊策略的優勢在於：（1）保持語義完整性，避免將相關內容強制分離；（2）確保文本塊大小適中，提高後續向量檢索的效率；（3）針對法律資料的特殊結構進行優化，能夠有效區分案件事實、受傷情形、賠償請求等不同段落。

經過分塊處理的法律資料隨後進行向量化，儲存至向量資料庫中，為後續的案件相似度比對和檢索提供高品質的資料基礎，確保系統能夠精確找到與使用者輸入最相關的案件資料。

4. 分類器

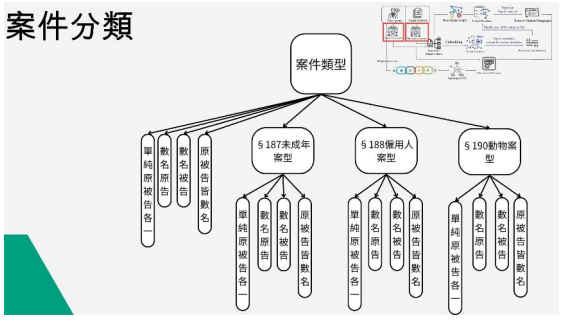

圖 3-6. 案件分類器

為提升檢索階段的精準度與效率，本系統設計一套如圖 3-6 所示的案件分類器，針對資料庫中已結構化的案例與使用者輸入資訊進行分類處理。分類依據包含原、被告人數（如單一或多數當事人）以及被告性質（如是否為未成年人、僱用人等特殊法律主體），共有十六種類別。此分類結果作為檢索器的前置篩選條件之一，有助於縮小檢索範圍並強化語意比對準確性。

5. 檢索器

檢索模組負責比對使用者輸入的案件事實與原告受傷情形，從資料庫中以餘弦相似度計算的方式找出語意上最相近的 K 筆案件。系統將參考這些相似案件中所引用之法條內容，做為後續生成起訴狀的法律依據，以確保生成文本所援引之法條具備合理性、正確性與法律依附性。此一設計不僅強化了法條引用的準確度，也提升了整體法律文書生成的專業性與信賴度。

6. 思維鏈與起訴狀生成模組

本系統採用分段式生成機制，將起訴狀依結構劃分為三大段落：案件描述、法律引用、賠償金額，並分別套用不同技術以確保生成結果的準確性與合規性。

6-1. 案件描述段落

利用思維鏈（Chain of Thought, CoT）技術，針對使用者輸入之事發經過進行逐步推理與邏輯展開，確保關鍵細節（如事發時間、地點、行為責任等）在生成文本中完整保留且符合事實描述邏輯。

6-2. 法律條文引用段落

首先統整系統檢索出的相關法條，統計其在相似案件中出現的頻率，僅保留出現次數超過門檻（threshold）的條文作為候選清單（記為 $Laws_1$）。再從使用者輸入文本

中進行法條關鍵字比對，擷取出另一組候選條文（記為 Laws₂）。系統將 Laws₁ 與 Laws₂ 進行交叉比對，若結果不一致，則交由語言模型判斷是否採用特定法條，以提升法條引用之精準性與正當性。

6-3. 賠償金額段落

利用自然語言處理（NLP）技術擷取使用者輸入中的各項賠償金額（如醫療費、看護費、慰撫金等），再由後端邏輯計算模組進行總額加總，確保生成起訴狀中賠償金額之匯總結果準確無誤，並符合使用者所輸入之原始意圖。

四、主要成果與評估

圖 4-1. 使用者輸入介面

圖 4-1 為本系統之使用者輸入介面。使用者需在此處填寫案件事發經過、原告受傷情形，以及具體賠償請求金額，作為起訴狀生成之基礎資料。系統亦支援以 .txt 檔案匯入輸入內容，若有多筆案件需同時處理，則可透過上傳 .xlsx 檔案方式批次操作，並指定特定欄位作為輸入來源。

使用者可進一步選擇欲使用的 RAG 模式與語言模型，並設定從資料庫中檢索相似案件的數量，以輔助提升生成文本的法律合理性與準確度。完成上述設定後，點擊「生成起訴狀」按鈕，系統將自動啟動整合檢索與生成流程，輸出符合格式與法律邏輯的民事起訴書。

圖 4-2.

圖 4-2 為系統在啟動起訴狀生成流程後的初步結果畫面。畫面左側顯示從資料庫中檢索到的最相似案件，以供使用者參考其事實描述與法律適用；畫面右側則展示未經思維鏈（Chain of Thought）處理所產生的快速版本起訴狀，作為對照樣本，用以凸顯思維鏈技術對生成品質的實質影響。

圖 4-3.

圖 4-3 呈現系統採用分段生成策略時的中間處理過程，依序展示三個主要組成段落：事實陳述、法律條文引用與賠償金額計算，並搭配各段對應的 Chain of Thought 步驟說明，以視覺化方式說明該技術如何強化邏輯推理與生成準確性。

圖 4-4.

圖 4-4 為最終生成完成的民事起訴狀畫面，提供完整文書預覽，並支援將生成結果匯出為 PDF 檔案，方便使用者下載存檔或提交使用。

這個專題透過Knowledge Graph RAG和chain of thought製作出了能夠穩定產出固定格式起訴書的系統，並且在事實描述和賠償金額的細節部分和核心的法律引用方面都有不錯的正確率，顯示語言模型在法律領域的使用價值

五、結語與展望

本專題展示了以小型語言模型配合 Graph RAG 技術應用於法律文書生成的可行性與實用性。未來可望擴展至其他法律文件（如答辯狀、和解書等）之自動生成，並強化使用者介面設計，提升系統可用性。此外，亦可進一步優化圖形資料庫的更新機制，對應法規修訂與新增判例，使系統持續保持高準確率與實務適用性。

六、銘謝

感謝指導教授的專業指引，也感謝柏宏學長幫助我們建程式的環境、給我們參考資料和實作細節上的指導，以及莉庭學姊提供清楚又美觀的流程圖和示意圖。

七、參考文獻

[1] Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, Rivindu Perera “Better Call GPT, Comparing Large Language Models Against Lawyers” arXiv:2401.16212, January 2024

[2] Jinzhe Tan, Hannes Westermann, Karim Benyekhlef “ChatGPT as an artificial lawyer?” Workshop on Artificial Intelligence for Access to Justice (AI4AJ 2023) Vol3435 June, 2023.

[3] Andrew M. Perlman, “The Implications of ChatGPT for Legal Services and Society” Suffolk University Law School Research Paper No. 22-14, December , 2022

[4] X Wu, R Duan, J Ni “Unveiling security, privacy, and ethical concerns of ChatGPT” in Journal of Information and Intelligence, March 2024

[5] OpenAI privacy policy “https://openai.com/policies/row-privacy-policy/”, November 2024

[6] DLB Eliot “The need for explainable AI (XAI) is especially crucial in the law” Available at SSRN 3975778, December 2021

[7] L Chen, G Varoquaux “What is the role of small models in the llm era: A survey” in arXiv:2409.06857, September 2024

[8] P Barceló, M Monet, J Pérez, B Subercaseaux “Model interpretability through the lens of computational complexity” in Advances in neural information processing systems, 2020

[9] A Louis, G van Dijck, G Spanakis “Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models” in AAAI Conference on Artificial Intelligence, February 2024