基於 LegalBERT 與知識圖譜的

反托拉斯法案進入障礙判例檢索系統

唐嘉宏M124111043

錢竑邑M113020075

林健驊M113010071

摘要

本專案旨在利用微調後的 LegalBERT 模型，結合擴展的知識圖譜，檢索美國反托拉斯法案案例中與進入障礙（Entry Barriers）相關的段落。專案的核心目標是構建一個完整且高效的法律文本檢索 Pipeline，包括數據清理、模型微調、知識圖譜擴展與檢索匹配分析等關鍵步驟，以提升檢索的準確性和語意理解能力。

首先，針對涵蓋 1891 年至 2022 年的 9305 個判例進行數據清理，包括處理雙欄文本格式、移除噪音以及統一文本結構。接著，通過對 LegalBERT 模型進行微調，進一步提升模型對法律術語及複雜語意的理解能力，特別是在進入障礙相關概念（如 "Economies of Scale"、"Predatory Pricing" 等）的檢索中表現出顯著提升。

此外，知識圖譜的構建與擴展是本研究的重要環節。我們基於進入障礙的基本定義，構建了初始知識圖譜，並新增多個類別（如 "Brand Loyalty" 和 "Patent Protection"），以及與之相關的同義詞和變體，從而增強模型對多樣化表達的檢索能力。

最後，整合檢索結果與知識圖譜進行語意匹配，並深入分析不同數據集（Data1、Data2、Data3 及 Supreme Court）的檢索表現。研究發現，知識圖譜擴展顯著提升了模型對多數數據集的匹配準確性，但對 Supreme Court 案例的檢索仍存在挑戰，可能需要進一步的微調與圖譜優化。

本專案不僅展示了結合自然語言處理技術與知識圖譜在法律文本檢索中的潛力，還為未來在專業領域中構建高效檢索系統提供了實踐經驗和技術基礎。

介紹

1. LexisNexis數據

本研究使用 LexisNexis 平台的 9,305 個案件作為主要資料來源。LexisNexis 是一個全球知名的法律和商業資訊平台，提供豐富的案件資料庫及相關資源，涵蓋多國司法判例、法律分析及其他輔助資訊。在研究過程中，系統透過文本分析技術捕捉案件中與特定字詞和相關文字的關聯，從而深入分析判決內容及法律論點。

2. LegalBERT

LEGAL-BERT 是一組專為法律領域設計的 BERT 模型系列，旨在解決法律文本處理中的特定需求。由於通用的 BERT 模型主要基於維基百科和書籍語料進行預訓練，其在處理法律文本時存在性能瓶頸。為了提升法律自然語言處理（Legal NLP）的表現，LEGAL-BERT 提供了一個專門針對法律語料進行預訓練的模型，支援法律科技（Legal Tech）和計算法律學（Computational Law）的研究與應用。該模型系列包括 LEGAL-BERT-FP，這是在 BERT-BASE 模型基礎上進行進一步法律語料預訓練的版本，適用於需要快速適應法律語料的情境；LEGAL-BERT-SC，這是從零開始基於法律語料進行預訓練的版本，並使用新的子詞詞彙表以精確處理法律專業術語；以及 LEGAL-BERT-SMALL，一個小型化版本，儘管參數較少（35M），但仍能提供優異的性能，特別適合資源有限的研究者和應用場景。訓練語料的總大小為 12GB，涵蓋來自歐盟法規（EURLEX）、英國法規（LEGISLATION.GOV.UK）、歐洲人權法庭案例（ECHR）、美國法庭案例（CASE LAW ACCESS PROJECT）、美國合同（SEC-EDGAR），以及 LexisNexis 數據等多個資料來源。LexisNexis 數據的加入進一步豐富了模型的法律知識庫，特別是在處理美國法律案例和法律條文的表現上具有顯著提升。模型結構方面，LEGAL-BERT 使用與 BERT-BASE 相同的 12 層、768 隱藏單元和 12 個注意力頭，而 LEGAL-BERT-SMALL 則擁有 6 層、512 隱藏單元和 8 個注意力頭，提供資源需求更低的解決方案。

3. 知識圖譜

知識圖譜（KG）作為一種結構化的知識表示方法，已廣泛應用於搜尋引擎、推薦系統等領域。隨著知識需求的增長，知識圖譜的擴展（KGE）成為提升其覆蓋範圍和表達能力的關鍵。知識圖譜擴展主要通過知識補全、多模態擴展、自動化擴展、歸納式擴展及基於圖嵌入的方法來進行。這些方法能夠有效補充缺失的實體和關係，並通過融合文本、視覺等模態數據，提升

圖譜的語義表示能力。隨著技術的發展，知識圖譜擴展在智能搜索、推薦系統和語義推理等應用中發揮著越來越重要的作用。

4. 檢索增強生成(Retrieval-Augmented Generation，RAG)

檢索增強生成是結合了信息檢索和生成模型的創新方法，旨在解決大型語言模型（LLMs）在處理事實性幻覺、知識過時以及缺乏專業知識等挑戰中的不足。傳統的LLM在生成文本時，會依賴訓練數據中已經包含的知識，但這些知識可能是過時的或不完整，導致生成的內容存在事實性錯誤或不符合領域專業要求。為了解決這些問題，RAG通過檢索外部知識庫中的相關信息，並將這些信息與生成模型結合，使生成的文本更具準確性和時效性。

在RAG框架中，檢索模塊負責從大型文檔庫或知識庫中搜尋與輸入問題或任務相關的信息，然後生成模塊根據檢索到的資料來補充或強化生成的內容。這樣的策略不僅改善了模型的事實性準確度，還能夠提供更多的專業知識，特別是在需要深度領域理解的場景中。

方法

1. 數據收集

本研究使用 LexisNexis 平台的 9,305 個案件作為主要資料來源。LexisNexis 是一個全球知名的法律和商業資訊平台，提供豐富的案件資料庫及相關資源，涵蓋多國司法判例、法律分析及其他輔助資訊。在研究過程中，系統透過文本分析技術捕捉案件中與特定字詞和相關文字的關聯，從而深入分析判決內容及法律論點。

為了有效組織和檢索案件資料，本研究採用了結構化的分類方法。首先，依據法院類型進行分類，將案件按法院層級（如最高法院、上訴法院、地方法院等）分組，以識別不同層級司法機構的裁決特點。其次，按照地區進行劃分，涵蓋國家、州或地方法院的司法管轄範圍，以適應地域性法律研究的需求。此外，基於案件的裁決年份進行時間排序，有助於分析法律發展趨勢和歷史變遷。同時，對案件中的律師與法官進行記錄和分類，將其參與角色及裁決模式納入考量，為法律專業人員的相關研究提供參考。最後，案件中的法庭意見（如多數意見、反對意見、協同意見）也被詳細標記，為深入理解案件背後的法律論證提供支撐。

這一分類體系的實現，結合了結構化元數據與文本分析技術，利用自然語言處理（NLP）從判決書中提取關鍵資訊，並通過規則和機器學習演算法進行自動分類，最終在用戶界面中以多維篩選和檢索功能呈現。此分類方法不僅實現了快速且精確的資料定位，還滿足了法律研究

者多樣化的需求，為後續法律趨勢分析和相關研究奠定了堅實基礎。

本研究使用四組數據集（Data1、Data2、Data3 及 Supreme Court），涵蓋 1891 年至 2022 年的 9305 個判例作為研究對象（如表一所示）。研究流程包括數據預處理、模型微調、知識圖譜構建與擴展以及結果匹配與分析。首先，針對數據進行清理與格式化，處理文本噪音及雙欄結構等特性。接著，利用微調後的 **LegalBERT** 模型對文本進行語意理解和預測，進一步提升對專業術語與語境的適應能力。此外，構建基於進入障礙的初始知識圖譜，並進行同義詞與類別擴展，以增加檢索的多樣性與準確性。最後，透過模型生成的結果與知識圖譜進行匹配與分析，深入探討不同數據集在檢索與生成中的表現差異，並評估模型在專業場景中的應用潛力。

數據集	文件	文件標籤	年份範圍	判例數量 (單一PDF)	共計
Data1	a-c	a1-c10	1891-01-01 ~ 1991-10-01	b2, c2：93個判例， 其餘100個判例	2986個判例
Data2	e-g	e1-g10	1998-09-14 ~ 2009-04-23	f2, g2：99個判例， 其餘100個判例	2998個判例
Data3	h-j	h1-j10	2009-04-23 ~ 2017-07-05	h2：98個判例， i2：99個判例， 其餘100個判例	3302個判例
	k	K1-k4	2021-08-31 ~ 2022-12-31	k1：5個判例， 其餘100個判例	
Supreme Court	d	d1-d19	1891-01-01 ~ 2022-12-31	1 個判例	19個判例
總計	a-k	A1-k4	1891-01-01 ~ 2022-12-31		共9305個判例

表一：LexisNexis數據

2. 資料清理

本研究基於 LexisNexis 平台的 9,305 份案件資料，進行了嚴謹的數據清理與結構化處理，旨在提取精確的 Opinion 段落以供後續分析。由於不同法院的 PDF 結構特性差異，我們採用了一套系統化的資料清理流程，結合高效文本提取技術與上下文維持策略，確保處理後數據的準確性與完整性。

首先，利用 PyMuPDF 工具提取 PDF 文件中的文本區塊與黑線（black line），以準確識別判例中的段落結構和層次。在此過程中，對無效資訊（如頁眉、頁腳及註腳 Footnotes）進行過濾和移除，確保數據的純淨性。

對於 Supreme Court 案例，因其 PDF 為雙頁並排文本結構，我們設計了專門的文本解析策略。通過判斷文本排版特性並進行分欄處理，成功提取了左、右兩欄的文本內容。這一適配的處理方式有效避免了內容混淆或缺失，保障了 Supreme Court 案例文本的完整性與準確

性。

針對較長段落的文本裁切，我們採用了 Loading Window 策略。設定最大 Token 長度為 256，步長為 128，以實現文本的分段裁切，同時保持裁切後段落上下文的連續性，避免關鍵資訊丟失。這一策略特別適用於語義分析場景，確保模型能夠有效理解長文本中的語境。

清理完成後，所有處理過的文本數據被保存為 Processed_Opinion.csv，該檔案包含三個主要欄位：Case Index（案件索引）、Page（頁碼）及 Paragraph（段落文本）。這些結構化數據為後續的多維度分析（如法院類型、地區、年份等）奠定了堅實基礎。

通過採用上述方法，本研究有效應對了不同法院文本結構的多樣性與處理挑戰，結合文本解析技術和上下文維持策略，為法律文本的高效提取與深度分析提供了可靠支持。這一處理流程不僅提高了數據的整潔度與可用性，還在一定程度上為其他法律研究提供了通用的技術框架。

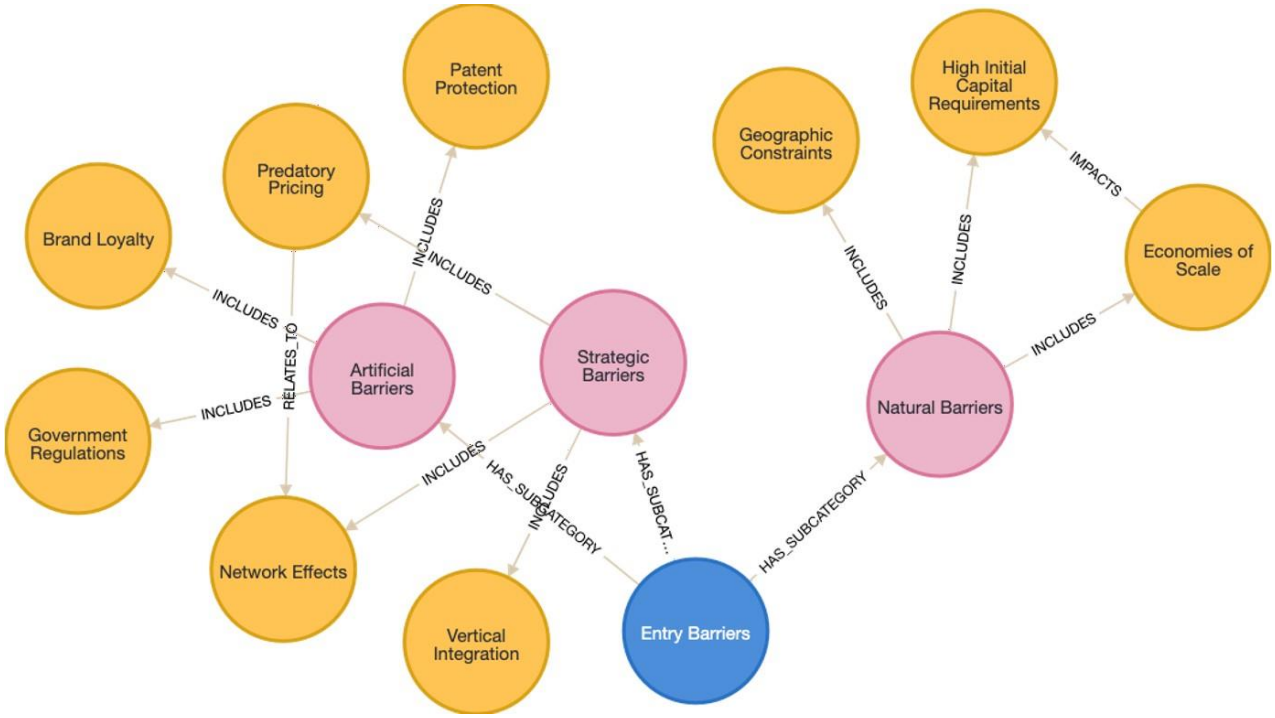
3. LegalBERT 模型微調

本研究旨在微調 LegalBERT 模型，以提升其在法律文本中辨識與進入障礙相關段落的能力。進入障礙指的是在行業或市場進入過程中所面臨的法律與競爭限制。為達成此目標，我們設計了兩個具體問題來指導模型識別相關段落，分別是："What is the judge's opinion regarding entry barriers in case {Case Index}?" 和 "What is the prosecutor's argument regarding entry barriers in case {Case Index}?" 由於 GPU 資源有限，我們從 Data2 資料集中選取了約 1000 個案例進行微調，這些案例包含了與進入障礙相關的法律段落。資料集中的進入障礙相關段落比例較高，使得模型能夠專注於此類法律文本特徵。微調過程中，我們基於 LegalBERT（BERT-Base 架構），並在法律語料庫上進行預訓練。每對問題與相關段落進行編碼，並標註正確答案的開始與結束位置，以幫助模型學習定位答案。微調過程使用 AdamW 優化器，學習率為 2e-5，批次大小為 16，訓練 epochs 設為 3，並設置 Weight decay 為 0.01，以幫助正則化模型、防止過擬合。微調後，我們利用 Hugging Face Trainer API 進行訓練與評估，並測試模型在識別進入障礙相關段落方面的準確性。

4. 知識圖譜擴展

在本研究中旨在通過結合原始知識圖譜並擴展進入障礙的相關定義，提升模型在語意模糊情境下的檢索成功率。我們首先基於進入障礙的基本定義，如“Entry Barriers”和

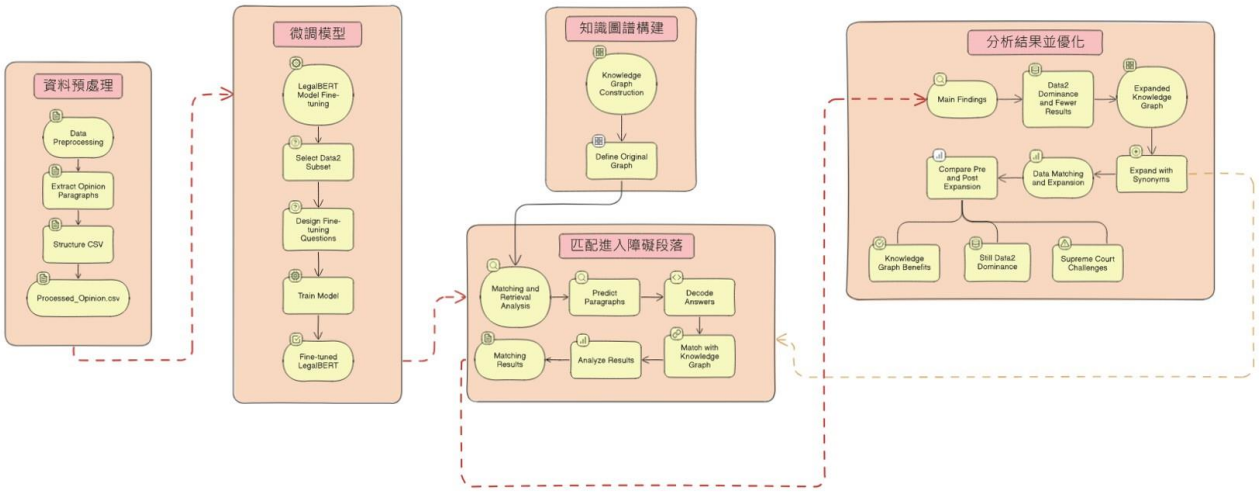
“Economies of Scale”等，構建了包含13個檢索類別的原始知識圖譜。接著，我們進行知識圖譜擴展，新增同義詞（如“Barriers”和“Entry Costs”）以增強模型對多樣表達的理解。進一步地，我們對進入障礙進行了更詳細的定義與類型擴展，包括“Brand Loyalty”和“Patent Protection”等新類型，從而更全面地捕捉進入障礙的各種形式。通過這些擴展，我們能夠提升模型在語意模糊情況下的檢索成功率，確保模型能夠更準確地識別和匹配相關信息。如圖一所示，這是我們所構建的知識圖譜結構圖，展示了各類進入障礙的關聯及其擴展定義。該圖表現了從原始知識圖譜到經過擴展後的多層次結構，並明確顯示了每個類別之間的語意聯繫，進一步說明了如何在語意模糊的情況下實現精確的檢索。



圖一：知識圖譜結構圖

5. 加入 RAG 預測與驗證

本研究的目標是結合檢索增強生成（RAG）預測與驗證，利用滑動窗口切割的段落進行推理，並通過知識圖譜匹配來提升檢索準確性。具體流程如下：首先，使用微調後的 LegalBERT 模型對段落進行預測，生成對應的答案。隨後，通過知識圖譜中的類別及其定義對生成的答案進行匹配，篩選出與進入障礙相關的結果。這一過程不僅能夠提高答案的準確性，還能有效地利用知識圖譜的結構化信息來增強檢索結果的語義準確性。通過這一方法，我們能夠在處理法律領域相關問題時，提升模型的推理能力與檢索精度。



圖二：機器學習整體流程圖

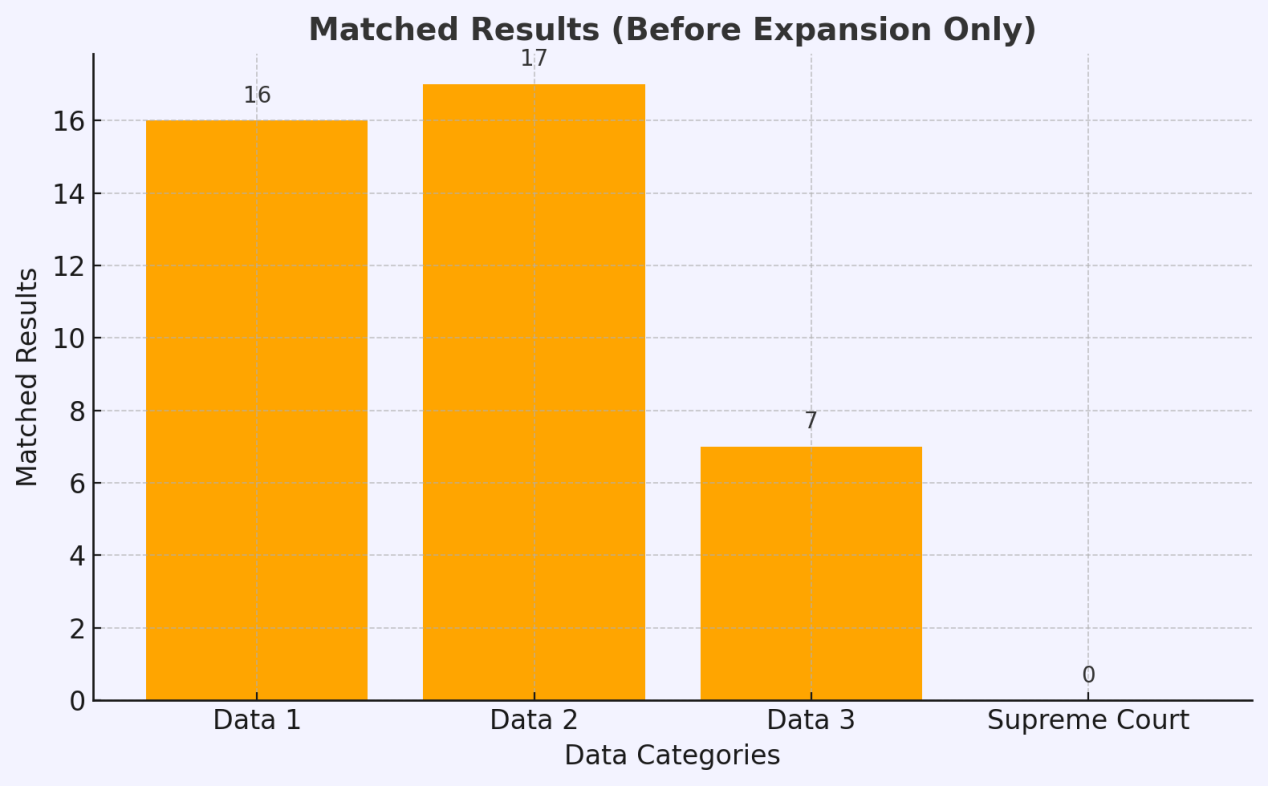
如圖二所示，這是本研究的機器學習流程圖，展示了從段落切割、預測、匹配到最終結果篩選的完整過程。該圖表明了如何通過微調的LegalBERT模型進行初步推理，並利用知識圖譜結構進行結果的優化和驗證。

結果

1. 比較匹配結果分布：

在本研究中，我們分析了知識圖譜擴展對模型檢索能力的影響，並對不同數據集的匹配結果進行了評估。結果顯示，Data2 在所有數據集中匹配數量最多，且微調對其影響顯著，成為主要的匹配來源。這表明微調後的模型在處理該數據集時表現最佳，能夠有效識別和匹配相關信息。相較之下，Data1 和 Data3 的匹配數量低於Data2，但在知識圖譜擴展後，這些數據集的匹配類型和數量均有所增加，顯示出知識圖譜的擴展有效提升了這些數據集的檢索能力。

然而，Supreme Court 數據集未能成功匹配到任何案例，可能是由於該數據集的結構或語意與現有知識圖譜中的定義不符，這揭示了知識圖譜在處理某些特定數據集時的局限性。為此，未來的研究應進一步探討如何調整知識圖譜，或對模型進行優化，以提升對這類數據集的匹配能力。總體而言，擴展知識圖譜顯著增強了模型對大多數數據集的檢索能力，特別是在語義理解和多樣表達的匹配上，但也暴露了如Supreme Court等數據集的結構性問題，這需要進一步的研究和改進。匹配結果分布如圖三所示，清晰展示了各數據集在知識圖譜擴展後的匹配情況。



圖三：匹配結果圖

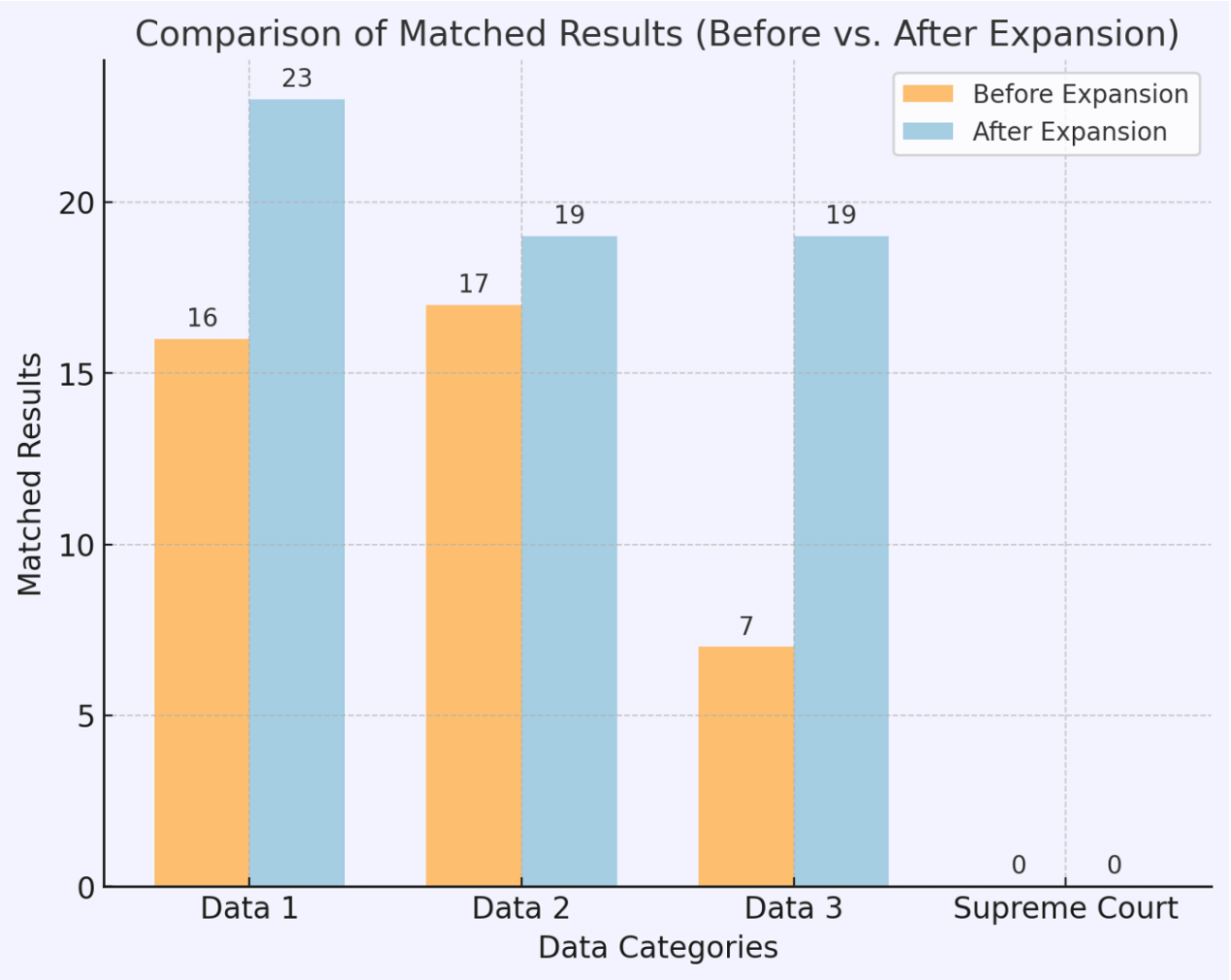
2. 擴展的結果分布：

如圖四的條形圖所示，本研究分析了擴展知識圖譜對檢索能力的影響。圖中土黃色條形代表擴展前的匹配結果，淺藍色條形則代表擴展後的結果。分析顯示，擴展前，匹配結果主要集中於少數類別，例如 "Entry Barriers"，類別覆蓋範圍相對有限。

在擴展後，匹配數量整體提升了約 20%，並新增了如 "Brand Loyalty" 和 "Patent Protection" 等新的類別，顯示出知識圖譜擴展對於增強模型多樣性匹配能力的顯著作用。同時，擴展後 Data1 和 Data3 的匹配數量和多樣性均有所增加，進一步證實了擴展知識圖譜的有效性。

然而，對於 Supreme Court 數據集，擴展後依然無法檢索到結果，這可能是由於該數據集的語意或結構與知識圖譜的定義不匹配，反映了模型在某些特定領域上的局限性。總體而言，擴展知識圖譜顯著增強了模型的檢索性能，尤其是對多樣類別的匹配能力，但也指出了部分領域需要進一步改進的方向。

擴展知識圖譜影響：匹配數量提升 20%，新增檢索類如"Brand Loyalty"的匹配效果顯著。



圖四：擴展前和擴展後匹配結果圖

討論

1. 針對 Data2 的微調效應：

本研究發現微調顯著增強了模型對該數據集的理解能力，特別是對術語如 "Entry Barriers" 和 "Predatory Pricing" 的敏感度顯著提升。這表明，微調後的模型在處理段落結構和語意關聯方面更加精確，能更好地捕捉進入障礙相關的關鍵信息。

此外，知識圖譜的擴展進一步展示了微調的基礎作用，模型在擴展後能有效檢索出更多與進入障礙相關的判例，顯示微調與知識圖譜擴展相輔相成的效益。然而，對於 Supreme Court 案例，模型未能成功檢索到相關結果，這可能是由於該數據集的語意或結構與現有知識圖譜的定義不匹配。這一挑戰表明，針對特定領域的數據集，可能需要專門進行微調，或對知識圖譜

進一步擴充，以提升檢索性能。

總體而言，微調與知識圖譜擴展共同提升了模型在多數數據集上的表現，但同時也暴露出某些特定領域的局限性，這為後續的改進提供了方向。

2. 主要發現

本研究的主要發現集中在微調後模型的表現及知識圖譜擴展的影響上。首先，Data2 顯示了明顯的主導效應，微調後的 LegalBERT 在該數據集中表現突出，其對段落結構和語意的適配性顯著高於其他數據集，匹配成功率也遠遠超過其他數據集，表明微調能有效增強模型對特定數據集的處理能力。

其次，知識圖譜的擴展對於匹配類型的多樣性帶來了明顯的提升，特別是在 Data1 和 Data3 上，新增的類別有效擴展了模型的檢索範圍與能力。這說明，結合微調與知識圖譜擴展可以顯著改善模型在多數數據集上的檢索表現。

然而，Supreme Court 數據集仍未能成功匹配到相關結果，可能是由於其語意或結構特性與現有知識圖譜定義不符。這一挑戰表明，針對特定領域數據集，可能需要進一步擴展知識圖譜，或調整檢索策略以提升適配性。

總體來看，本研究強調了微調和知識圖譜擴展在提升檢索準確性和多樣性方面的關鍵作用，同時也指出了未來針對特定領域改進的潛在方向。

未來方向

為進一步提升檢索性能與處理能力，本研究提出以下改進方向。首先，針對 Supreme Court 數據集，進行專門微調以應對其雙欄結構與特殊語言模式，同時擴展知識圖譜的類別與定義，以更好地適配該數據集的語意需求。其次，在算力優化方面，引入高記憶體 GPU，支持更大規模的數據集處理與更複雜的模型訓練。第三，探索使用更強大的法律專用模型，例如 Lawformer 或 GPT-Legal，充分發揮其在法律語意理解方面的優勢，提升模型的專業表現。第四，將檢索與分類需求整合於微調過程中，採用多任務學習架構，並設計更聚焦的主題 Prompt，例如將通用 Prompt 「What’s the opinion from judges」 優化為 「What’s the opinion regarding entry barriers」，以提高檢索的效率與精準度。最後，進一步整合檢索增強生成（RAG）與知識圖譜的語意匹配，構建更高效的檢索與生成框架，實現更加準確且語意敏感的結果生成。這些改進方向將有助於提升模型在專業場景中的應用價值，並為檢索與生成技術的

未來發展提供新方法與新視角。

個人貢獻說明

唐嘉宏M124111043：45%
工作分配：主要負責撰寫程式碼、資料收集、資料清理、模型訓練、補充資訊和報告
錢竑邑M113020075：27.5%
工作分配：主要負責PPT製作
林健驊M113010071：27.5%
工作分配：主要負責書面報告製作
以上這些工作分配項目都是我們一起討論來完成這份專案，過程也許非常繁瑣複雜但我們學到互相合作分工和一起討論這個研究專案的能力以及知識。

參考文獻

1. Baker, J. B. (2024). How economists influence antitrust: the contributions of Tim Bresnahan, Janusz Ordover, Steve Salop, and Bobby Willig. *Journal of Antitrust Enforcement*, jnae049. https://academic.oup.com/antitrust/advance-article/doi/10.1093/jaenfo/jnae049/7900903

2. Cao, S. (2022). Economic Analysis in Antitrust Law: An Automated Approach Applied to US Appellate Courts. *Stan. Computational Antitrust*, *2*, 155. https://heinonline.org/HOL/LandingPage?handle=hein.journals/stfdcmp2&div=9&id=&page=

3. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*. https://arxiv.org/abs/2010.02559

4. Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence, https://ojs.aaai.org/index.php/AAAI/article/view/29728

5. Zhao, Y., Zhang, Y., Zhou, B., Qian, X., Song, K., & Cai, X. (2024). Contrast then Memorize: Semantic Neighbor Retrieval-Enhanced Inductive Multimodal Knowledge Graph Completion. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, https://dl.acm.org/doi/abs/10.1145/3626772.3657838

6. Zhou, R., Yang, Y., Wen, M., Wen, Y., Wang, W., Xi, C., Xu, G., Yu, Y., & Zhang, W. (2024).

TRAD: Enhancing LLM Agents with Step-Wise Thought Retrieval and Aligned Decision. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, https://dl.acm.org/doi/abs/10.1145/3626772.3657788