



國立中山大學企業管理學系

碩士論文

Department of Business Management

National Sun Yat-sen University

Master's Thesis

**基於 LegalBERT 與知識圖譜的  
美國反托拉斯法案進入障礙文本檢索研究**

Research on Text Retrieval of Entry Barriers in U.S. Antitrust  
Laws Based on LegalBERT and Knowledge Graphs

研究生：唐嘉宏

Jia-Hong Tang

指導教授：佘健源

Prof. Chien-Yuan Sher

中華民國 115 年 1 月

January 2026

# 論文審定書

## 國立中山大學研究生學位論文審定書

本校企業管理學系企業管理碩士班

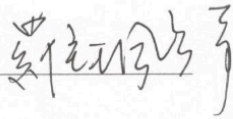
研究生唐嘉宏（學號：M124111043）所提論文

基於LegalBERT與知識圖譜的美國反托拉斯法案進入障礙文本檢索研究  
Research on Text Retrieval of Entry Barriers in U.S. Antitrust Laws Based on  
LegalBERT and Knowledge Graphs

於中華民國 114 年 7 月 3 日經本委員會審查並舉行口試，符合碩士學位論文標準。

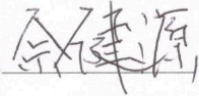
學位考試委員簽章：

召集人 羅珮綺

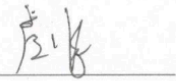


委 員

余健源



委 員 盧憶



委 員

\_\_\_\_\_

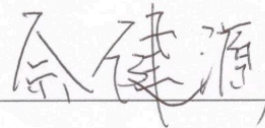
委 員

\_\_\_\_\_

委 員

\_\_\_\_\_

指導教授(余健源)



(簽名)

# 誌謝

本論文的完成，凝聚了眾多師長的教誨、同儕的支持與家人的陪伴。在這段充滿挑戰與成長的研究旅程中，每一位給予我啟發與幫助的人，都是我能堅持走到最後的重要力量。謹在此致以最誠摯的感謝與敬意。

首先，衷心感謝指導老師余健源教授，在整個研究過程中給予我細膩而深刻的指導。從研究主題的選定、資料架構的設計，到論文摘要撰寫與年份分布分析的邏輯推演，余老師總是耐心聆聽、精確點評，協助我在思緒紛雜時釐清方向。尤其在研究方法敘述上，老師不僅教導如何條理分明地呈現內容，更教會我以科學精神謹慎推論，這份啟發將成為我日後學術與人生道路上的珍貴資產。

同時，誠摯感謝盧憶老師在本研究中對法律專業知識的深入指導。盧老師以其嚴謹且精湛的法學素養，協助我在進入障礙的法律概念建構上更為精確，並針對各判例中的法理細節提供了寶貴意見。盧老師的指導不僅使我能從更高層次理解反托拉斯法的核心精神，也讓我體會到學術研究中對細節的堅持與尊重。也感謝盧憶老師從中山大學的法律資料庫下載關於反托拉斯法的相關資料集供本研究使用，此資料集來源為中山大學的 Nexis Uni® 法律資料庫，並且為線上公開授權可供下載。<sup>1</sup>

此外，感謝羅珮綺老師在 LLM 與知識圖譜建構領域上的協助。羅老師在微調策略到知識結構設計上提供專業建議，更鼓勵我從跨領域整合的角度思考，開拓了我對人工智慧與法律應用結合的想像。每當遇到技術瓶頸與架構難題時，羅老師提出具體而可行的方向，使本研究得以逐步落實。

在研究過程中，曾因資料預處理繁瑣、模型資源受限、語義推理設計困難而

---

<sup>1</sup>[https://nsysu.primo.exlibrisgroup.com/permalink/886NSYSU\\_INST/do7u9r/alma991008596139607977](https://nsysu.primo.exlibrisgroup.com/permalink/886NSYSU_INST/do7u9r/alma991008596139607977)

陷入焦慮與迷惘。所幸在師長們的鼓勵與指引下，我學會了如何一步步拆解問題、堅持嘗試與不斷優化，也深刻體會到「研究」並非一蹴而就，而是透過不懈累積與思辨反覆淬煉出的成果。

同時，也要感謝在研究期間給予我支持與鼓勵的同儕與朋友。與他們的交流與討論，讓我在思考問題時有了更多元的視角，也在面對壓力與挫折時，獲得了持續前行的力量。

最後，最深的感謝獻給我的家人。感謝家人們無條件的愛與支持，讓我能夠無後顧之憂地專注於研究。是他們的理解與陪伴，成就了今天的我。每一份成果，背後都有著家人默默守護的身影，這份恩情銘記於心。

本研究僅是探索智慧法律應用的一小步，未來仍有無數課題等待挑戰。感謝所有引導我走到這裡的人，我將帶著這份感激，繼續在學術與實務的道路上不斷努力前行。

[唐嘉宏] 謹於

國立中山大學

中華民國 115 年 1 月

## 摘要

本研究旨在探討如何結合語言模型如 LegalBERT 與知識圖譜技術來處理自然語言處理 (NLP) 任務，提升對美國反托拉斯法案中「進入障礙」相關段落的語義檢索能力。反托拉斯法案核心在於防止壟斷、維護市場競爭秩序，而進入障礙作為其關鍵議題，隨著經濟結構、政策立場與科技進展不斷演變。傳統基於關鍵詞的檢索方式難以捕捉法律語句中的語義變體與隱喻性描述，因此極需語義理解導向的檢索方法進行補強。在文章段落中常使用語言模型或是大型語言模型 (LLM) 來標示，這邊統一說明有生成式 AI 功能 (如 ChatGPT 和 Deepseek) 統稱為 LLM，而如 BERT、LegalBERT 等沒有生成式 AI 功能的模型統稱為語言模型。

本研究所使用之資料涵蓋 9,305 筆美國法律判例，資料來源橫跨地方法院、上訴法院與最高法院，內容多樣且結構不一。資料處理階段採用正則表達式 (Regular Expression) 與滑動窗口 (Sliding Window) 方法清理與切割段落，將長文本切分為可處理的語義單元 (Chunk)，每段最大 token 長度限制為 256，以配合法律語言模型輸入規格。接著，透過問答式模板對 LegalBERT 模型進行微調，以強化其對進入障礙語境的識別能力。此外，本研究構建結構化知識圖譜，整合法學教科書中對進入障礙的定義、同義詞、分類與語義關係，並轉換為 JSON 格式於推論階段導入模型，強化語義擴展與相似語句匹配效果。

驗證結果顯示，微調後的 LegalBERT 在 Data2 資料集 (聯邦上訴法院案例) 中表現最佳，其檢索準確性自 88% 提升至 92%，而在其他資料集中亦有明顯改善。尤其在 Data3 中，語義擴展後的進入障礙段落命中數由 7 筆提升至 19 筆，證實知識圖譜的語意補強功能。惟在 Supreme Court 資料集中，無論是否導入語義擴展，命中率皆為 0%，顯示模型在處理高抽象語體與語意結構複雜文本時仍面臨挑戰。為進一步驗證檢索片段的語義正確性，本研究亦結合地端部署的 LLM (如 GPT-

4o、phi4 等) 進行語義推論與結果校對，透過提示工程 (Prompt Engineering) 與指示微調 (Instruction Tuning) 有效判別片段語意是否實質涉及進入障礙議題。

本研究不僅建立了一套以法律語義為導向的文本檢索方法，也展現出知識圖譜與中小型語言模型結合於法律任務中的可行性與經濟性。未來可朝多語言模型擴展、自動化知識圖譜建構與多代理人 (Agent) 協作工作流 (Work Flow) 方向發展，並推廣至更多國際法律文本與規範應用場景中。

本研究之程式碼已公開於 GitHub<sup>2</sup>，其中包含原始資料、資料預處理的原始程式碼。另一份針對 LLM 應用之擴展，亦已獨立上傳至 GitHub<sup>3</sup>，可供後續研究參考與再利用。

**關鍵詞：**反托拉斯法案、LegalBERT、知識圖譜、進入障礙、語義擴展、法律檢索、大型語言模型

---

<sup>2</sup><https://github.com/chrisj890926/textmining>

<sup>3</sup>[https://github.com/chrisj890926/Textmining\\_LLM](https://github.com/chrisj890926/Textmining_LLM)

# Abstract

This study aims to explore how language models, such as LegalBERT, can be integrated with knowledge graph techniques to enhance semantic retrieval in the context of U.S. antitrust law, specifically focusing on identifying paragraphs related to "barriers to entry." The core objective of antitrust enforcement is to prevent monopolistic behavior and preserve market competition. Barriers to entry, a central theme in this domain, are constantly evolving due to shifts in economic structures, policy orientations, and technological advancements. Traditional keyword-based retrieval methods often struggle to capture the semantic variability and metaphorical expressions prevalent in legal texts. To address this limitation, this study employs semantic-aware approaches by leveraging both language models (e.g., BERT, LegalBERT) and generative LLMs (e.g., ChatGPT, DeepSeek). For clarity, models with generative capabilities are referred to as LLMs, while non-generative models like BERT are referred to as language models.

The dataset used in this research comprises 9,305 U.S. legal case documents spanning district courts, appellate courts, and the Supreme Court, encompassing diverse formats and structural patterns. During preprocessing, we employed regular expressions and a sliding window strategy to clean and segment paragraphs into semantic chunks suitable for model input, with a maximum token length of 256 tokens per chunk to align with LegalBERT input constraints. Subsequently, we fine-tuned the LegalBERT model using a question-answering template tailored to the semantic context of barriers to entry. Additionally, a structured knowledge graph was constructed by extracting definitions, synonyms, classifications, and semantic relationships from legal textbooks. The graph was transformed into JSON format and incorporated during inference to support semantic expansion and

similar-sentence matching.

Evaluation results demonstrate that the fine-tuned LegalBERT model achieved its best performance on the Data2 dataset (federal appellate court cases), with retrieval accuracy improving from 88% to 92%. Other datasets also showed significant gains. Notably, in the Data3 dataset, the number of matched "barriers to entry" paragraphs increased from 7 to 19 after semantic expansion via the knowledge graph, confirming its effectiveness. However, in the Supreme Court dataset, the hit rate remained 0% regardless of semantic expansion, indicating the model's difficulty in handling highly abstract or complex legal discourse. To further validate the semantic correctness of the retrieved fragments, we integrated local deployments of LLMs such as GPT-4o and phi4 for inference and verification. Using prompt engineering and instruction tuning, these LLMs were able to confirm whether each passage substantively discussed barriers to entry.

This study not only proposes a semantic-driven legal text retrieval framework but also highlights the feasibility and cost-effectiveness of integrating knowledge graphs with compact language models in legal applications. Future work may explore extensions toward multilingual model adaptation, automated knowledge graph construction, and multi-agent legal workflow orchestration. The proposed methodology also holds potential for broader applications across international legal documents and regulatory contexts.

All source code, including original data and preprocessing scripts, is publicly available on GitHub<sup>4</sup>. An additional repository dedicated to LLM-based extensions is also available<sup>5</sup> for further research and reproducibility.

---

<sup>4</sup><https://github.com/chrisj890926/textmining>

<sup>5</sup>[https://github.com/chrisj890926/Textmining\\_LLM](https://github.com/chrisj890926/Textmining_LLM)

Keywords: antitrust laws, LegalBERT, knowledge graph, entry barriers, text retrieval, LLM.



# 目 錄

論文審定書 .....	i
誌謝 .....	ii
摘要 .....	iv
Abstract .....	vi
目錄 .....	ix
圖次 .....	xi
表次 .....	xii
第一章 介紹 .....	1
1.1 研究背景 .....	1
1.2 研究目標 .....	3
1.3 章節結構 .....	3
第二章 文獻探討 .....	5
2.1 反托拉斯法的相關研究 .....	5
2.2 語言模型的相關研究 .....	7
2.2.1 為何選擇 LegalBERT .....	7
2.2.2 語言模型與知識圖譜的結合 .....	8
2.3 知識圖譜技術 .....	11
2.3.1 方法比較 .....	13
2.3.2 公式與原理 .....	14

2.4 研究缺口 .....	18
第三章 資料及研究方法 .....	20
3.1 研究資料 .....	20
3.2 研究架構 .....	21
3.3 數據預處理 .....	23
3.4 模型微調 .....	26
3.4.1 模型訓練原理補充：反向傳播與梯度下降 .....	27
3.4.2 模型訓練參數配置 .....	29
3.5 知識圖譜構建 .....	30
第四章 模型測試結果分析與優化 .....	36
4.1 模型測試概述 .....	36
4.2 驗證資料集文本是否正確切割 .....	37
4.3 性能評估與外部模型驗證 .....	38
4.4 語義擴展效果 .....	41
4.5 進入障礙相關判例年份分布 .....	45
第五章 結論 .....	48
5.1 研究貢獻 .....	48
5.2 研究局限性 .....	49
5.3 未來研究方向 .....	51
參考文獻 .....	55
附錄 .....	60
5.4 附錄資料 .....	60

## 圖 次

圖 2-1	RAG 流程架構圖 . . . . .	16
圖 3-1	研究架構圖 . . . . .	22
圖 3-2	滑動窗口處理方法示意圖 . . . . .	25
圖 3-3	知識圖譜示意圖 . . . . .	35
圖 4-1	語義擴展對檢索效果的提升 (Before vs. After Expansion) . . . . .	42
圖 4-2	進入障礙相關判例年份分布 (Opinion) . . . . .	45
圖 4-3	進入障礙相關判例年份分布 (Other Labels) . . . . .	45
圖 I	數據集示例圖 . . . . .	60

## 表 次

表 2-1	不同方法的比較 . . . . .	13
表 3-1	模型訓練參數設定 . . . . .	30
表 4-1	數據集統計信息 . . . . .	37
表 4-2	模型性能評估（加入與未加入知識圖譜對比） . . . . .	39
表 4-3	LegalBERT 模型在各資料集上的檢索效能（Precision / Recall / F1） . . . . .	44
表 4-4	法院類別分類統計 (Opinion) . . . . .	46
表 4-5	法院類別分類統計（Core Terms、HN、FN） . . . . .	46



# 第一章 介紹

## 1.1 研究背景

反托拉斯法案的主要目的是防止壟斷行為，確保市場中維持有效的競爭環境，進而保護消費者利益、促進創新與效率提升。傳統上，反托拉斯法的執行仰賴法律分析、經濟模型與實證數據。然而，隨著數據規模的快速增長和技術的飛躍進步，大數據與機器學習（Machine Learning）技術在反托拉斯實務中的應用日益廣泛，為執法機關提供新的分析工具與證據來源。

在人工智慧領域中，語言模型因其在 NLP 任務中的強大表現，成為近年來的研究焦點。這些模型不僅能進行語義理解與推論，亦能自動生成具語境意識的文字，在法律、醫療、金融等領域展現出強大的潛力。在反托拉斯分析中，語言模型將有助於處理大量判決書、申訴資料、企業報告等非結構化文本，進行資訊抽取與語意分析，提升法律文件處理效率與準確性。

特別值得關注的是檢索增強生成（Retrieval-Augmented Generation, RAG）技術的崛起。該架構結合了資訊檢索與文字生成，允許模型根據外部知識庫回應查詢，使其生成內容更加準確與有根據。此技術相對於純粹的生成式模型更具可解釋性與可控性，尤其適用於需要高度準確性的法律應用場景。

近期 Chen et al. (2024) 提出一種基準化方法，旨在系統性地評估語言模型在 RAG 任務中的表現。該研究建立一套涵蓋多種任務類型與語料情境的評估框架，有助於比較不同模型架構與檢索機制在法律文本處理上的成效，為本研究提供了方法論上的重要參考依據。

在反托拉斯法的實務分析中，法律文本中常含有大量抽象的概念、冗長複雜的句構以及專業術語，例如「潛在競爭者」、「市場進入障礙」、「支配地位的濫用」等。傳統的關鍵詞匹配或基於統計的方法，如 TF-IDF (Term Frequency—Inverse Document Frequency) 一種計算「某個詞在一篇文章中有多重要」的指標。或是 BM25 (Best Match 25) 在 TF-IDF 基礎上發展而來的演算法，被廣泛應用於現代的搜尋引擎系統，其設計目的在於更準確地計算「使用者查詢與某段文字內容的相似程度」。這兩種方式雖能處理部分關鍵詞檢索任務，但面對具有語意延伸、邏輯推論需求的文本分析時，往往無法掌握上下文間的語義關係與隱含意涵。

為此，本研究關注語言模型在法律文本處理上的應用潛力。LegalBERT 由 Chalkidis et al.(2020) 所提出，該模型基於預訓練的 BERT 架構進行訓練，並透過大規模法律文本（如判決書、法規條文、法律期刊等）進行微調，使其具備處理法律語言的專業語義能力。研究指出，LegalBERT 在多項法律 NLP 任務（如法律問答、案例分類、條文比對等）中表現優於通用語言模型，顯示其具備良好的領域適應性。

此外，單純依賴語言模型所產出的語義向量仍可能面臨語義模糊或概念遺漏等問題。為了進一步提升模型對於「進入障礙」相關法律文本的理解與檢索準確性，本研究引入知識圖譜進行語義擴展。透過結構化的概念定義與關係鏈結，知識圖譜可提供補充語意背景，使模型能更精確地辨識法律術語間的邏輯層次與相似概念（例如將「自然障礙」與「地理限制」進行語義對齊），進而提升檢索結果的完整性與一致性。

綜上所述，結合 LegalBERT 與知識圖譜的語意強化機制，不僅可補足傳統檢索方法的限制，也能為反托拉斯法相關案件中的進入障礙分析提供更有力的技術支撐，作為本研究設計與實驗方向的關鍵動機。

## 1.2 研究目標

本研究旨在結合法律語言模型與知識圖譜，提升對反托拉斯法案中「進入障礙」相關概念的文本檢索與理解能力，具體目標如下：

1. 微調 LegalBERT 模型，以提升其在反托拉斯法語境中對「進入障礙」相關語句的語義辨識與檢索準確性。本研究將選取含有進入障礙概念之法律判決書與相關文獻，構建訓練與測試資料集，並透過監督式學習進行模型微調，強化其領域適應能力。
2. 構建結構化的法律知識圖譜，將進入障礙相關子類型（如自然障礙、策略性障礙、人工障礙等）及其定義、關聯語句與延伸語彙納入語義網路中。藉此，模型在處理輸入文本時，能透過知識查詢或語義擴展理解相關上下文與概念層次，補強純語言模型語意模糊或片段理解的問題。
3. 設計一套完整的法律文本檢索流程，包括資料收集與標註、文本預處理（如段落切分、詞彙正規化）、模型訓練與微調、知識圖譜整合流程，以及最終效果評估（包括準確率、與案例分析等）。該流程可作為未來應用於其他法律主題（如市場支配地位、價格操縱等）的基礎模組，具備良好之可擴展性與再利用性。

## 1.3 章節結構

本論文共分為五章，各章內容安排如下，旨在系統性說明本研究的背景脈絡、方法設計與實驗成果，逐步建構出一套結合法律語言模型與知識圖譜的法律文本檢索架構：

- 第一章：緒論。本章說明研究背景、動機與研究目標，並闡明本研究的重要性與可行性，為後續章節奠定問題意識與研究方向。
- 第二章：文獻回顧。回顧語言模型（如 BERT、LegalBERT）與 RAG 之發展脈絡，並探討知識圖譜於法律語義建構中的應用案例，從中找出本研究的切入點與研究缺口。
- 第三章：研究方法。詳細說明資料蒐集與預處理流程、LegalBERT 微調策略、知識圖譜建構邏輯與整合方式，並說明整體系統架構與技術選型依據。
- 第四章：模型優化與驗證結果。透過驗證微調後的模型與結合知識圖譜後系統的效能表現，採用準確率評估指標與實例分析進行比較，分析模型對進入障礙文本的語義掌握與檢索能力。
- 第五章：結論與未來展望。統整本研究之成果與貢獻，檢視實驗結果與原研究目標之契合度，並提出未來可行的擴展方向與應用建議。

## 第二章 文獻探討

### 2.1 反托拉斯法的相關研究

反托拉斯法主要目的在於維持市場公平競爭，防止企業藉由壟斷行為或不正当交易手段限制競爭者或消費者選擇。其中，經濟分析為評估競爭行為合法性與合理性的核心工具之一，廣泛應用於市場界定、市場勢力測量、價格策略評估與進入障礙分析等層面。

Baker (2024) 的研究系統性地探討了經濟學家在反托拉斯分析中所扮演的角色，強調理論模型與實證資料在執法判斷中的影響力。特別是在「進入障礙」的判斷方面，Baker (2024) 指出，若市場存在顯著的結構性或策略性障礙，即使目前競爭看似充分，未來仍可能因缺乏潛在競爭者而出現市場失靈。本研究即借重其所提出的理論架構，將進入障礙視為影響市場競爭強度的關鍵變數，並據此作為法律文本分析的焦點。

許多反托拉斯法的研究中，對「進入障礙」有明確的理論分類與實證分析。Bain (1956) 認為自然障礙如高額的資本需求、技術密集程度與規模經濟，會使新進入者難以與現有企業競爭。Motta (2004) 則從政策與法規面出發，指出專利制度、進口限制與政府發照制度亦可能形成法律上的進入壁壘，進一步扭曲市場進入機會。Carlton and Perloff (2015) 則將焦點擴展至策略性障礙，舉凡掠奪性定價、品牌忠誠、獨占性合約，皆屬於企業主動設下、目的在於排除潛在競爭者的障礙。

然而，在實務中這些障礙往往不會直接以標籤式的語言出現，而是透過隱

喻、間接描述、上下文推論等方式呈現於法院判決或監管報告之中。法律語言本身具備高度結構化與語義模糊並存的特性，這使得傳統檢索系統與 NLP 工具難以有效擷取其中蘊含的語意線索。Chalkidis et al.(2020) 指出，法律文本的語句通常冗長、上下文依賴性高，且許多關鍵資訊並非明示而是潛藏於語意之中。Zhong et al.(2020) 亦發現，在法律摘要或判決重點擷取任務中，模型若缺乏跨句推論與語義推廣能力，常難以抓住判決主旨。近年，Yao et al.(2019) 提出結合知識圖譜與語言模型的方法，用以補足模型對專有概念與隱性語義的識別力，提升語義補全與推論精準度。

反托拉斯法的執行與分析，長期以來皆倚賴產業經濟學（Industrial Organization, IO）中所發展的理論基礎與計量工具，藉以釐清企業行為對市場結構與消費者福利的實質影響。Bain（1956）所提出的市場結構理論，揭示企業規模、資本門檻與產品差異化等因素對競爭程度的關鍵作用，奠定了後續反托拉斯分析的理論根基。進一步地，IO 理論中的「結構—行為—績效（Structure-Conduct-Performance, SCP）」架構，提供了一套分析市場力量與策略行為如何影響最終市場成果的系統性方法，並被廣泛應用於競爭政策的制定與法律案件的經濟分析中。Carlton and Perloff (2015) 則在此架構之上，補充了更多動態競爭、資訊不對稱與策略互動的模型，使現代反托拉斯實務得以更精緻地理解企業壟斷與進入障礙等問題。

Cao (2022) 在其研究中，深入探討了 IO 概念在反托拉斯法律文本中的實際應用，特別關注於法院在判決過程中如何評估市場結構（如市場集中度、進入障礙）與企業行為（如價格協議、捆綁銷售、掠奪性定價）之間的因果關係。在 Cao (2022) 中強調，許多法律實務案例中雖未直接引用 IO 理論模型，但其推論邏輯與判斷標準多與 IO 分析方法一致，顯示法律實踐與經濟理論之間具有高度互補性。因此藉由在 Cao (2022) 中提到的分析脈絡：強調法院在反托拉斯案件中進行法律

推理時，雖不一定直接引用 IO 模型，但其判斷邏輯往往隱含了對市場結構、企業行為與競爭結果三者之間因果鏈的評估思維。此種推理模式本質上延續了 IO 理論的 SCP 架構的邏輯脈絡，亦即先界定市場與產業結構，再評估企業是否採取具有排除效果的策略行為，最終推導是否造成對市場競爭或消費者福利的不利影響。於資料處理階段特別標註反托拉斯案件文本中涉及市場結構、企業策略行為與競爭結果的相關語句，並藉此設計語義擴展策略與模型訓練標籤。透過結合 IO 分析邏輯與語言模型的語言理解能力，有助於提升模型對法律語義的掌握精度，進而提高檢索結果的準確性與可解釋性。

## 2.2 語言模型的相關研究

近年來，語言模型在 NLP 領域中取得突破性進展，廣泛應用於機器翻譯、文本生成、情感分析及問答系統等任務。其中，法律文本作為一種高度專業且結構複雜的語言類型，亦逐漸成為語言模型應用與優化的關鍵場域。

### 2.2.1 為何選擇 LegalBERT

法律文本具有高度專業性與語言結構複雜的特性，常包含抽象術語、長句結構、交叉參照與形式化邏輯，其語意理解難度遠高於一般自然語言。傳統的關鍵詞匹配或語句相似度計算方法，難以處理如「構成壟斷」與「具有市場支配力」這類語義相近但文字表達差異大的法條或意見書描述。

為有效處理此類文本，本研究採用 LegalBERT 模型作為語義理解的核心工具。LegalBERT 是基於 BERT 架構所發展出的領域特化預訓練語言模型，專為法律文本設計。該模型由 Chalkidis et al.(2020) 提出，LegalBERT 為一種基於 BERT 架構的預訓練語言模型，透過大規模法律文本（涵蓋歐洲法院判決、法律期刊、

法條等) 進行語言建模訓練。相較於傳統以關鍵字或統計為基礎的檢索方法，語言模型透過上下文編碼 (Encoding) 與多層語義抽象，能更準確地掌握詞彙在特定語境下的意義，並處理跨句子甚至跨段落的邏輯推論。LegalBERT 模型藉由雙向編碼的特性，能同時考慮目標詞彙的左右語境，對於法律文本中常見的多義詞(如“charge”、“bar”、“consideration”等) 有較強的語意辨識能力。舉例而言，詞語「entry」在法律語境中可指涉「進入市場」、「入場條件」或「報關文件」，語言模型可透過上下文學習其具體含義，避免誤判。

根據在 Chalkidis et al.(2020) 中的驗證結果，LegalBERT 在處理法律術語、多義詞彙、長句結構與邏輯判斷時展現出更高的準確性與穩定性。其已廣泛應用於法律文書分類、案件比對、法條檢索與問答系統等任務，在多項 benchmark 任務中表現優異。

然而，語言模型仍存在語意模糊、事實一致性不足等問題，特別是在需要引用明確知識或法條時容易產生虛構現象。所以結合結構化的外部知識資源如知識圖譜，成為強化語言模型語意準確性與可解釋性的有效手段。在本研究中，我們將 LegalBERT 作為基礎模型，進一步微調其對於「進入障礙」相關語句的辨識能力。透過結合人工標註資料與知識圖譜語義擴展，本模型能有效解析法律文本中與進入障礙概念相關的潛在語義資訊，並應用於特定主題檢索任務中。

## 2.2.2 語言模型與知識圖譜的結合

雖然語言模型(如 BERT) 具備強大的語義理解與現今 LLM 的自然語言生成能力(如 ChatGPT)，但在面對知識密集或概念複雜的專業任務時，仍存在若干限制，例如語意模糊、事實錯誤以及知識涵蓋不足等問題。這些限制在處理法律文本等高度專業化領域時尤其明顯，原因在於法律語句中常包含抽象概念、特定術

語與上下文依賴強的語義結構。

爲了解決上述問題，近年研究逐漸朝向結合語言模型與知識圖譜的方向發展。這類方法的核心精神在於：語言模型負責處理語言結構與語義關係，而知識圖譜則扮演知識支持與語境補強的角色。Yao et al. (2019) 提出，知識圖譜能以「實體—關係—實體」(entity—relation—entity) 的形式建構語義網絡。簡單來說，「實體」可以理解爲圖譜中的關鍵詞或重要名詞，例如一家公司、一個法律概念或一條法規；「關係」則是實體之間的連結方式，例如「屬於」、「違反」、「受到保護」等。舉例來說：

「掠奪性定價」—— 是 → 「策略性障礙」；

「策略性障礙」—— 屬於 → 「進入障礙」；

「反托拉斯法第 2 條」—— 規範 → 「壟斷行爲」。

這種結構就像在建立一張「語意地圖」，每個實體之間都透過清楚的关系連接起來。這不僅能幫助語言模型理解不同專有名詞之間的邏輯與層級，也能在語義推論過程中提供上下文線索與協助消除語意歧義。例如，在處理「進入障礙」這類概念時，知識圖譜可進一步拆解爲自然障礙、法律障礙與策略性障礙，並附帶其定義、相關條文與實際案例，有助於語言模型理解其完整語意與應用脈絡。這種結構不僅能幫助模型建立專有名詞之間的意義連結，也能在語義推論過程中提供上下文與消除歧義的能力。

此外，Yao et al. (2019) 中也證明，結合語義相似度與知識圖譜補全，可以有效擴展語義覆蓋範圍，特別是在處理模糊查詢或需要跨概念推論 (conceptual bridging) 的任務中表現尤佳。另一方面，在 Zhang et al. (2024) 和 Chen et al. (2024) 中探討了語言模型與知識圖譜結合後所形成的 RAG 架構，該架構將外部知識檢

與 NLP 結合，在模型訓練資料量不足或查詢過於寬鬆的情況下，能顯著提升資訊搜索與語句生成的正確性。

本研究參考了上述方式，將知識圖譜作為語義擴展與模型輔助的核心機制，用以強化模型在處理進入障礙相關概念時的準確性與可解釋性。

進入障礙屬於高度語意抽象且易被各種隱喻描述的概念，例如「進入成本高」、「難以複製商業模式」、「法規設限」等，常以不明顯的關鍵字出現在判決文本中，或是整段判例在描述進入障礙卻沒提到任何進入障礙的關鍵字。透過結合 LegalBERT 模型與進入障礙主題導向之知識圖譜，能進一步引導模型捕捉潛在語意與上下文邏輯，強化語義擴展與段落匹配的準確性。因此，語言模型與知識圖譜的結合不僅為語義檢索任務提供了技術創新方向，也為本研究中法律語義推論與主題聚焦檢索提供了有效支撐架構。

以下為本研究引用之具體應用案例：

- **語義鄰居檢索技術**：此技術可透過拓展知識圖譜中節點間的語義關聯(如同義詞和延伸詞)，使語言模型能根據語義近似詞彙與概念，判定潛在相關文本。在知識圖譜節點(如“策略性障礙”、“價格戰略”)與查詢關鍵詞(如“掠奪性定價”)之間建立語義連結，可在知識不完整或語言表達多樣的情境下，生成更具涵蓋性的檢索結果。Yao et al. (2019) 提出結合語義相似度與知識圖譜節點連接擴展的語義檢索方法，證實其對於模糊查詢與抽象概念的詞語具有良好效果。
- **多模態知識圖譜補全**：透過融合文本、結構與上下文特徵，補強圖譜中缺漏的語意連結，原先節點跟節點之間只有一種關係，但可以透過構建跨節點的方式來達到抽象的跨語境理解，這種方式會讓關係不只有一種，因為進入障

礙涵蓋多種主題，可以使得跨主題之間也會有連接關係。本研究導入該技術於進入障礙相關法律概念的建構知識圖譜過程中，使 LegalBERT 模型在處理例如“法律障礙”、“自然限制”與“規範性授權”這些術語時，能準確掌握語義層級與相對應的語句，提高檢索任務中的語意對齊精確度。Zhang et al. (2024) 則指出，多模態知識補全能強化結構化語意在複雜概念推理中的應用表現。

## 2.3 知識圖譜技術

知識圖譜是一種以節點（實體）與邊（關係）構成的結構化語義表示方式，適合用於知識組織與機器可讀的概念關聯建構。所謂「實體」，可以想像成是知識中的一個「名詞」，像是某家公司、法律條文或法律概念（例如「策略性障礙」、「壟斷行為」、「進入障礙」）；而「邊」，則代表實體之間的語意連結關係，例如「是 □□ 的一種」、「受到 □□ 規範」、「與 □□ 有關」等。

這些實體與關係通常會以「三元組 (Triple)」的形式表達，也就是「主詞—謂詞—受詞 (subject—predicate—object)」。

舉個例子：

「掠奪性定價」——（是一種）→「策略性障礙」

「策略性障礙」——（屬於）→「進入障礙」

「進入障礙」——（定義於）→「反托拉斯法教科書」

這樣的表示方式就像是在畫一張語意關係圖，每個節點（名詞）之間都有標註好的邊（語意關係），讓語言模型也能像人一樣理解各個概念之間的關係。

Hogan et al. (2021) 指出，知識圖譜正是藉由這種三元組表示方式，在跨領域

語義連結中發揮強大效用，尤其適用於 NLP 領域中的語義推理、實體對齊、上下文補強與查詢展開等任務。其優勢在於可視化、可組合與可更新，使其能在語言模型之外提供穩定、邏輯一致且結構清楚的知識支撐，有助於強化語言模型在處理專業文本時的理解與解釋能力。值得進一步說明的是，Yao et al. (2019) 所提出的「實體—關係—實體結構」，與 Hogan et al. (2021) 所提的三元組，本質上是同一種邏輯架構，只是命名與應用語境略有不同。

在語言學中，通常使用「主詞—謂詞—受詞」來描述句子的結構邏輯，重視語句如何被機器理解與儲存。而在知識圖譜的應用中，如在 Yao et al. (2019) 中提出的 KG-BERT 模型，則偏好使用「實體—關係—實體」的命名，強調的是兩個知識實體（如法律概念）之間的語義關係。兩種結構在邏輯上是等價的，差別在於一種偏向語言結構分析，一種偏向知識表示與模型訓練（知識圖譜與 KG-BERT）。本研究因關注於語意推論與語義擴展的應用，因此採用 Yao et al. (2019) 提出的實體導向表示方式，有助於結合語言模型與圖譜知識進行語意補強與段落片段比對。

在法律領域中，知識圖譜被廣泛用於建構法條條文間的交互引用、案例與適用法條、法律概念的語義網絡。Chalkidis et al. (2020) 中指出，法律知識圖譜可以提升法條檢索與推理效率，並支援語言模型在多層級語義上建立規範邏輯。例如，將「壟斷行為」與其下的「搭售」、「價格歧視」等具體行為標註為下位節點，並連接相關法條與判例，即可形成一組語義清晰、可供檢索的知識結構。

本研究依循此邏輯，針對「進入障礙」主題建構主題導向知識圖譜，節點包含「自然障礙」、「人工障礙」、「策略性障礙」等核心子概念，並串接至對應的法律定義、判例片段與模型訓練文本。此圖譜不僅支援模型進行語義擴展，也可作為 RAG 任務中的檢索基礎，強化模型的泛化性與可解釋性。

### 2.3.1 方法比較

如表 2-1 所示，現有法律文本檢索與理解方法在準確性與運算效率之間存在一定的權衡，不同方法各具優勢與限制。

表 2-1: 不同方法的比較

方法	準確性	效率
傳統 TF-IDF	中等	高
LegalBERT	高	中等
知識圖譜結合	高	中

傳統 TF-IDF 方法為基於統計的詞項權重計算技術，具備運算快速、實作簡易的優點，適合用於結構單一或關鍵字明確的檢索任務。Salton and Buckley (1988) 指出，TF-IDF 在資訊檢索系統中能有效評估詞語的重要性，成為基本模型的常見選擇。然而，在面對法律文本中語義抽象、句型多變與詞彙多義的情況下，其準確性容易受限，無法有效處理語境依賴與概念延伸的語句。

相較之下，LegalBERT 透過深層語義編碼，能理解上下文間的語意關係與句法結構，顯著提升在專業語境下的語句匹配與片段檢索能力。Chalkidis et al. (2020) 所提出的 LegalBERT 模型即專為法律文本進行預訓練，展現出優於通用語言模型的檢索與分類能力。然而，語言模型運算成本相對較高，尤其在處理長文檔或大量查詢時，容易造成效能瓶頸。

為此，本研究進一步採用結合知識圖譜之方法，將結構化知識圖譜轉化為 JSON 格式加入語言模型的處理流程中，使模型在處理如「進入障礙」等抽象概念時，能藉由知識圖譜提供額外語義支援與語詞連結，進一步提高語意推理準確性。Yao et al. (2019) 和 Zhang et al. (2024) 分別提出結合語言模型與知識圖譜的補全與擴展機制，在處理模糊查詢與結構推理任務上，顯著改善檢索涵蓋與語義

對齊能力。雖然知識圖譜建構與整合需要額外預處理步驟，但在執行階段透過語義擴展與檢索範圍縮小，可在準確性與效率間取得較佳平衡。知識圖譜結合語言模型的方式不僅保有深度理解能力，亦透過結構化知識引導檢索方向，為本研究在反托拉斯法「進入障礙」文本檢索任務中提供最適解。

### 2.3.2 公式與原理

為了進一步提升法律語義檢索的準確性與語意涵蓋能力，近年研究開始導入結合語義向量與知識圖譜的 RAG。Yao et al. (2019) 與 Zhang et al. (2024) 提出的語義擴展與圖譜補全方法，提供了整合外部知識以強化語句表示的可行框架，特別適用於法律文本中概念結構複雜、語言表達多樣的應用場景。

本研究所採用的核心公式如公式 2.1 以及 2.2 所示，可將檢索任務形式化為「在所有候選文件中，選出與查詢語意最相近的那一筆」的問題：

$$\hat{n} = \arg \max_{n \in \{1, \dots, M\}} \text{sim}(X_q, X_n) \quad (2.1)$$

$$\text{sim}(X_q, X_n) = \frac{X_q \cdot X_n}{\|X_q\| \|X_n\|} \quad (2.2)$$

在公式 2.1 和 2.2 中，我們希望找出語意上「與查詢最接近的段落」。將每一筆資料都轉換成一個語義向量  $X_n$ ，再透過向量比對方式計算每一筆與查詢語句的相似度，最後挑出分數最高的那筆資料  $\hat{n}$ 。這個邏輯可以視為：從全部資料中，找出跟使用者想問的問題「最像的」那段文字。其中， $\hat{n}$  為最相關段落的索引； $M$  表示候選段落總數； $X_q$  為查詢語義向量， $X_n$  為第  $n$  個候選段落的語義向量； $\text{sim}(\cdot, \cdot)$  為語義相似度函數，本文使用餘弦相似度 (cosine similarity)，如公式 2.2 所示。

這種方式常見於語義搜尋與問答系統，計算方式主要依賴餘弦相似度來衡量向量的接近程度。(Mikolov et al., 2013)。

雖然本研究參考了 RAG 架構的設計理念，但需特別說明，本系統除了最後的 LLM 驗證以外並不包含語言生成任務，因此不屬於 RAG 原始定義下的完整架構 (Lewis et al., 2020)。本研究所採用之方法，主要強調結合知識圖譜進行查詢語義補強與語義重排序，因此更嚴謹的分類應屬於「知識圖譜輔助的語義檢索 (KG-augmented retrieval)」或「基於知識的語意重排序 (Reranking)」。

然而，考量本方法仍繼承了 RAG 架構「整合語言模型與外部知識」的核心精神，本文於圖與流程敘述中仍採用 RAG 作為整體系統設計的簡稱，並於本節中明確說明其應用範圍與差異性。

對應的 RAG 原理如架構圖 2-1 所示。該方式在 LegalBERT 編碼基礎上，透過結合知識圖譜節點的語義信息，對查詢向量進行擴展，進一步提升模型對法律語境下抽象詞彙（如進入障礙）之涵蓋與判讀能力。

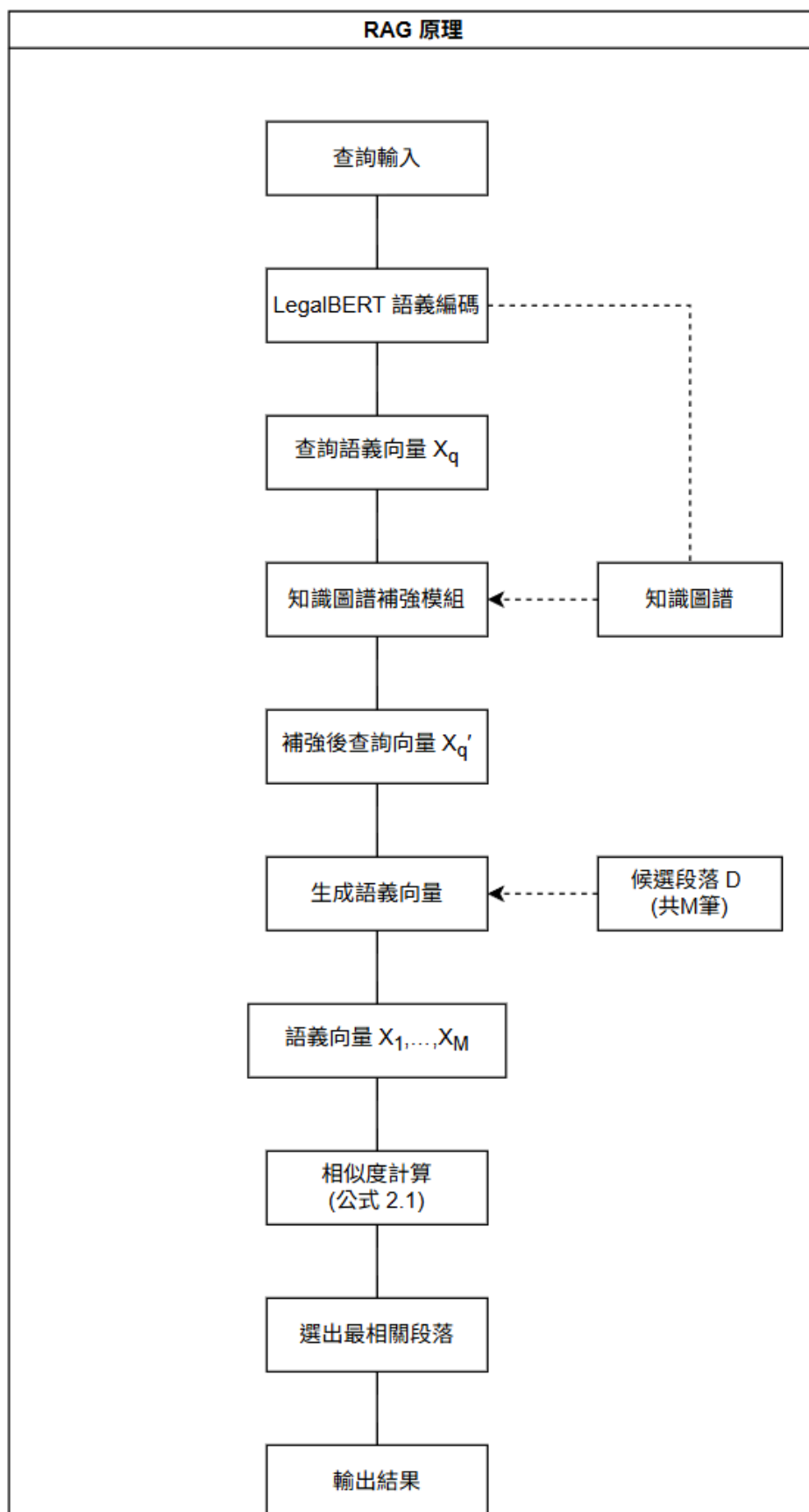


圖 2-1: RAG 流程架構圖

上述的 RAG，其目的是在進行文本檢索時，同時結合語言模型與知識圖譜來提升模型對查詢語意的理解與比對能力（Lewis et al., 2020；Yao et al., 2019）。詳細檢索流程如下：

1. 初始化模型：首先載入預訓練語言模型（LegalBERT）與知識圖譜中各節點的語義向量（Zhang et al., 2024）。
2. 語義編碼：針對使用者的查詢語句以及資料庫中每一筆候選文本，使用 LegalBERT 將其轉換為語義向量。這些向量代表了每段文字在語意空間中的位置與意涵（Mikolov et al., 2013）。
3. 語義補強：利用知識圖譜中的節點資訊對查詢向量進行擴展，例如：若查詢內容為「進入障礙」，圖譜可補充與其相關的子概念（如「自然障礙」、「策略性障礙」），讓模型理解更多語義上的可能性（Yao et al., 2019）。
4. 相似度計算：將補強後的查詢向量與每一筆資料進行語義相似度計算（使用餘弦相似度），衡量兩者在語義空間中的接近程度（Mikolov et al., 2013）。
5. 選出最相關文本：從所有候選資料中選出與查詢最相似者，即為本次檢索的結果。

這種結合 RAG 的架構不僅強化了語意理解，也提升了模型對法律概念（如進入障礙）之間潛在關聯的掌握能力，使其更適用於專業領域的語義檢索任務（Chalkidis et al., 2020；Yao et al., 2019）。這個方式具有兩項核心優勢：首先，透過 LegalBERT 建立查詢與文件的語義向量，能保留法律語境中的結構與語意；其次，知識圖譜提供查詢語意的擴展與補強，協助模型更精確地檢索到如「策略性障礙」或「自然進入壁壘」等概念相關段落。根據驗證結果顯示，該方法在準確率與召回率等評估指標上，皆優於僅使用語言模型的基準檢索策略。

## 2.4 研究缺口

回顧現有技術與研究觀察，仍面臨下列挑戰與待突破的研究問題：

- **超長文本的語義模糊問題**：法律判例往往篇幅冗長，且單一段落中可能同時提及多個主題或交叉引用其他法條。由於目前多數語言模型（如 LegalBERT）受到最大 token 長度（512 tokens）限制，需先進行文本切割，但這樣容易造成語義中斷與上下文遺失。如何保留語句連貫性與上下文邏輯，是處理法律長文本的一大瓶頸（Devlin et al., 2019；Chalkidis et al., 2020）。
- **知識圖譜的建構效率與涵蓋性限制**：現階段的知識圖譜建構仍大量仰賴人工定義與關鍵詞，無法涵蓋所有進入障礙的概念與整體表達。Yao et al. (2019) 指出，若缺乏語義連結與詞彙變化的設計，系統容易誤判段落是否與查詢主題相關。因此本研究特別強調從概念層次理解整段語意，而不僅僅依賴單一關鍵詞比對，所以擴展了相關詞彙的同義詞與延伸關鍵詞。
- **多模態數據的融合能力不足**：許多判例文件中會包含表格、標示圖、頁碼或頁眉等非結構化資訊。傳統 NLP 系統主要處理純文字輸入，難以融合圖像與文字之間的語義。Zhang et al. (2024) 指出，多模態語義融合是提升法律語境理解與文件解析能力的關鍵方向。

針對上述問題，本研究朝以下方向進行技術深化與應用拓展：

- 結合 **滑動窗口策略**與 **跨段落語義推理**，來處理超長文本上下文斷裂的問題，透過滑動窗口的切割方式能夠保留重要法律語義，並且透過編碼的方式讓模型記憶文本的開始和結束位置，這樣就不會導致切割後的文本上下文不連貫（Beltagy et al., 2020）；

- 導入 **知識圖譜自動化建構流程**，以 Neo4j 結合程式碼撰寫，將進入障礙相關主題與關鍵詞還有同義詞建構為語義節點網絡，再將這些關係輸出為 JSON 格式供語言模型和 LLM 使用，依照這種方式就不需要人工手動去連接節點建立多種複雜關係 (如跨主題之間的節點關係) (Zhao et al., 2022)；
- 探索 **多模態融合架構**，以本地端串接 Ollama API 實作 GPT-4o、llama3.2:latest 等先進 LLM，能夠處理判例中所有的圖像、表格與文字資料，且進行跨模態語義理解，通過先進的 LLM 對文件整體結構的理解，驗證加入知識圖譜後的進入障礙檢索效果 (Zhang et al., 2024)；
- 建構 **互動式法律 AI 檢索系統**，以 Chatbot AI 進行 NLP 對話介面結合語意檢索流程，協助使用者快速定位法律條文、案例重點與腳註依據，提升法律專業知識查詢的效率與決策品質 (Lewis et al., 2020；Liao et al., 2022)。

## 第三章 資料及研究方法

### 3.1 研究資料

本研究所使用的資料集特別感謝盧憶老師從中山大學的 Nexis Uni® 法律資料庫中下載，<sup>6</sup>共計 9,305 份美國法律判例，涵蓋多層級法院與多元議題，作為本研究語言模型訓練、RAG 任務、以及 LLM 驗證的基礎法律文本，詳細數據表格可參考附錄表I。主要數據來源可分為以下四組：

- **Data1**：收錄聯邦地方法院（U.S. District Courts）之部分判決，包含民事與反托拉斯相關案件，涵蓋時間區間約為 1990—2010 年；
- **Data2**：選取上訴法院（U.S. Courts of Appeals）之案例，特別聚焦於經濟與市場競爭議題，為本研究進入障礙文本的重要來源；
- **Data3**：涵蓋混合型資料來源，包括州級法院、行政機構判決與特定公開法律資料庫彙整案件；
- **Supreme Court**：整理自美國最高法院公開資料庫之歷年判例，具高度法律權威性，適合作為語義基準與檢索評估參考依據。

每筆判例資料中皆包含案件編號、法院名稱、裁判日期、參與法官與正文內容。其中，研究特別聚焦於每筆資料的 “**Opinion**” 段落，此段為法官對案件事實、爭點與裁判理由的詳盡敘述，涵蓋法律推論與概念定義，具有高度語意密度

---

<sup>6</sup>[https://nsysu.primo.exlibrisgroup.com/permalink/886NSYSU\\_INST/do7u9r/alma991008596139607977](https://nsysu.primo.exlibrisgroup.com/permalink/886NSYSU_INST/do7u9r/alma991008596139607977)

與分析價值。本資料集為本研究語義檢索系統設計與法律語言模型微調的核心資源，為進一步分析「進入障礙」語意片段、訓練 LegalBERT 模型與構建知識圖譜提供了關鍵基礎。

## 3.2 研究架構

本研究的整體設計架構如圖 3-1 所示，整體可分為五個主要階段，分別是：(1) 數據預處理（第 3.3 節）；(2) 模型微調（第 3.4 節）；(3) 知識圖譜構建（第 3.5 節）；(4) 匹配進入障礙文本（第四章）(5) 結果分析與優化（第四章）。

每一個階段都是依據法律文本的特性與分析需求所量身打造，並針對語意理解與運算效率進行調整，確保整個流程在資料處理與模型推論上保持語意一致與執行效率。簡單來說，我們的研究流程從處理法律判決資料開始，先進行「資料清洗與切割」，讓模型可以正確閱讀與處理冗長複雜的法律文字。接著進行模型的「微調」，透過已有的法律語言模型（LegalBERT），在我們的任務上做進一步訓練。第三階段，我們建立「知識圖譜」，這是一種幫助模型理解概念與關係的知識庫，裡面整理了與「進入障礙」相關的定義、同義詞與分類。最後一個階段是「結果分析與優化」，我們根據模型的表現進行調整與優化，讓模型能更準確地抓出法律文本中與進入障礙相關的段落。整個架構中，我們善用了多項工具與技術，包括 Python 程式語言、知識圖譜資料庫 neo4j、HuggingFace 提供的模型函式庫、以及 LegalBERT。經過每一階段的處理與優化後，最終我們將產出的模型與外部 LLM（如 GPT-4o 與本地端部署的 Ollama 模型）進行交叉驗證，以確認其準確性與實用性。接下來的章節會詳細介紹每一個階段的設計邏輯、實作方式與遇到的挑戰，並說明我們是如何一步一步處理並優化這些問題的。

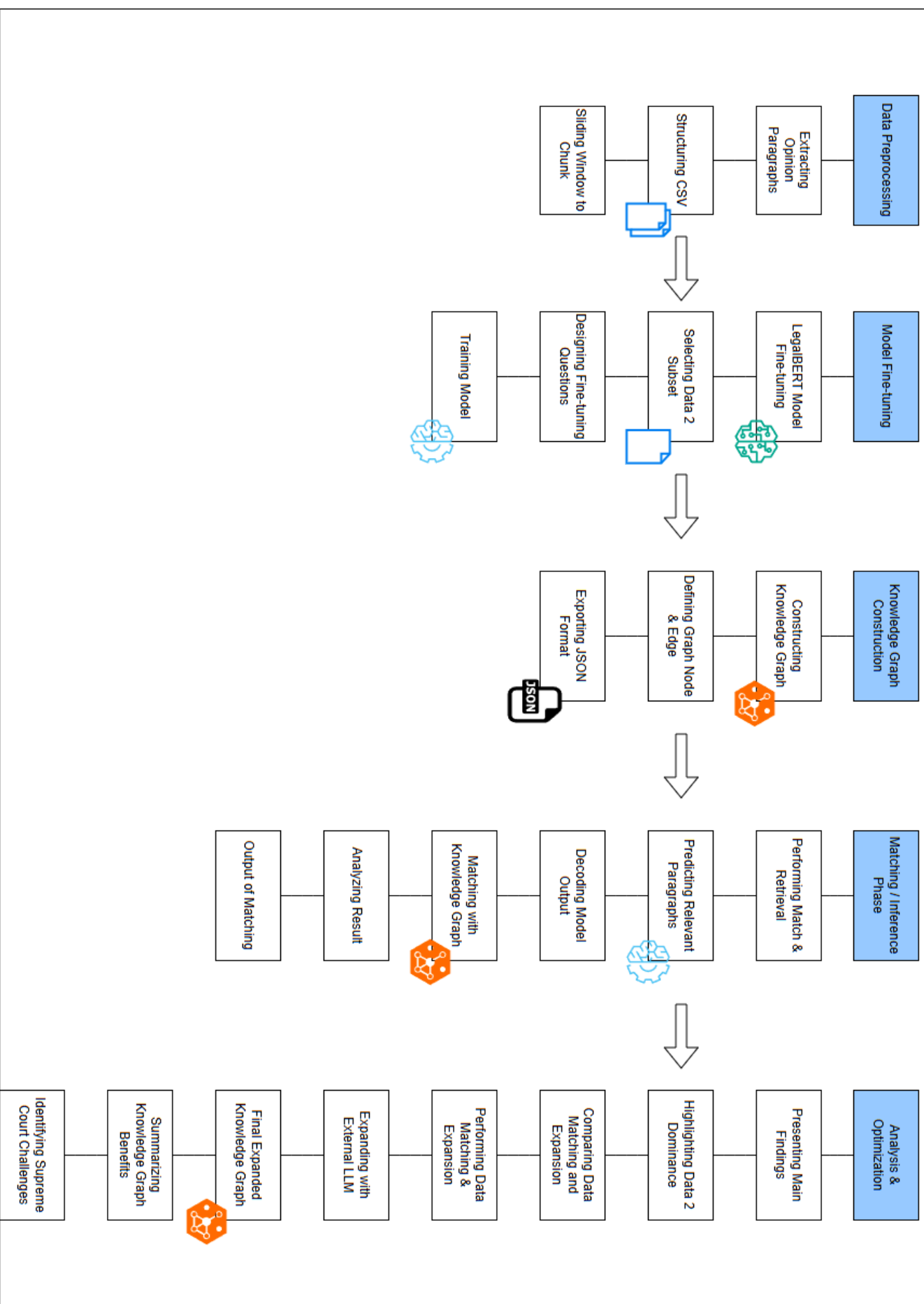


圖 3-1: 研究架構圖

### 3.3 數據預處理

在處理法律判例的原始 PDF 文件時，原始資料常以連續頁面呈現，缺乏結構化段落標記。因此，本研究透過 Python 套件 PyMuPDF (fitz) 與正則表達式，配合視覺線條與文字位置特徵，準確分割出如 Opinion、Footnotes、Headnotes、Core Terms 等法律段落，提供後續語言模型更乾淨且語義聚焦的文本輸入。以下要先介紹甚麼是正則表達式，正則表達式是一種可讓電腦比對與擷取文字樣式的語法規則，就如同人類閱讀時可藉由標題或格式辨識段落主題。在實作中，我們對每一頁的文字區塊進行比對，範例如下：

程式碼 1: 抓取 Core Terms 的正則表達式

```
matches = re.finditer(
    r'Core Terms\s+([\s\S]+?)(?=\n(?:[A-Z] [\^\\s]*(?:\s[A-Z] [\^\\s]*)*[:\n] |
    Opinion by:))',
    page_text,
    re.MULTILINE
)
```

這段程式碼的目的是抓取”Core Terms”標題開頭到下一個段落標題（如”Opinion by:”）之間的所有文字內容，`[\s\S]+?` 表示匹配多行任意文字，`(?=...)` 則用來判斷段落終止條件但不包含該標題。

除了文字比對，對於 Footnotes、LexisNexis Headnotes 等區塊，我們額外結合線條資訊進行偵測。例如 PDF 中每頁腳註下方會出現一條黑色線條，搭配如下條件判斷即可擷取其下方內容：

程式碼 2: 抓取頁面中腳註區塊的判斷邏輯

```
if item['type'] == 's' and item['color'] == (0.0, 0.0, 0.0) \
    and width_range[0] <= item['width'] <= width_range[1]:
    line_start = item['items'][0][1]
    if x_range[0] <= line_start.x <= x_range[1]:
        # 代表此線條為黑色且落在指定區域
```

進一步，我們也使用正則表達式抓取如 Opinion by、Judges、Counsel 等摘要資訊段落，採用多種組合模式搭配通用條件，也就是滿足所有可能情況的排列組合，例如在判例中會出現只有 Counsel 或是只有 Opinion by 等情況，才以這種方式以防漏掉某一個判例的段落內容：

程式碼 3: 抓取 Opinion by 與 Judges 等段落的 pattern

```
pattern = r'(Counsel:.*?)(?=Counsel:|Judges:|Opinion by:|Opinion|$)' \
          r'|(Judges:.*?)(?=Counsel:|Judges:|Opinion by:|Opinion|$)' \
          r'|(Opinion by:.*?)(?=Counsel:|Judges:|Opinion by:|Opinion|$)'
```

這些模式均為「非貪婪式比對」，意思是當程式在比對文字時，會「盡快」在符合條件時就停下來，只抓取最短必要的段落內容。舉例來說，若一個段落開始於 Opinion by:，而之後還有另一個 Opinion by: 出現，若使用一般的「貪婪式比對」可能會一次抓到兩個段落中間所有的文字；但改用非貪婪比對（在正則表達式中加入？符號），則能保證每次只抓到最近一次的結尾，避免跨段錯誤。這在法律判例的資料處理上特別重要，因為標題格式不總是統一、內文結構複雜，若一次擷取過多內容會導致段落分類錯誤或語義混淆。因此，本研究所有段落擷取策略皆採用非貪婪式設計，確保抽取結果更為準確且語義集中。

在資料處理過程中，針對判例中的意見書“Opinion”、腳註“Footnotes”、法律要點“HN Labels”、“Headnotes”、核心詞彙“CoreTerms”、“LHN Labels”、“LexisNexis® Headnotes”等段落使用 Python 撰寫正則表達式的相關程式碼來切分各個目標段落，匯入 CSV 並透過標註類別以 Table 的形式呈現(如 Case Index、PDF 檔案名、Labels 編號等)，將這些 CSV 進行標準化清洗，包括去除頁碼頁眉、不明代號與多餘空格等，並以 a1,a2...,b1,b2...,k4 的英文數字排列方式存放，以便針對對應的 PDF 內容進行檢查驗證。

此外，每筆資料皆標註其所屬資料集與案件編號，以利後續模型訓練、推論與檢索任務中的資料追蹤與對應。資料處理流程可重現，並已實作為模組化程式碼，<sup>7</sup>確保研究結果具可驗證性與延展性。

在法律文本的預處理階段，經過正則表達式進行初步文本清洗後，還不能直接丟入模型訓練，因為這些清理完的 CSV 因為法律判例文本過長，受到算力與模型限制無法直接進行訓練，為因應語言模型的輸入長度限制，本研究進一步採用滑動窗口策略進行段落切分。具體設定如下：每一輸入片段的最大 Token 長度為 256，滑動步長設為 128，確保相鄰片段之間有 50% 重疊，避免因語句分割而造成語境中斷。此策略可兼顧上下文完整性與模型計算資源效率。(Beltagy et al., 2020)

圖 3-2 示意了滑動窗口處理方法的整體流程與重疊關係。

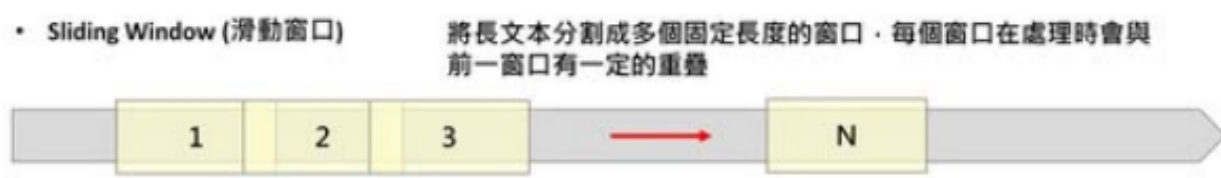


圖 3-2: 滑動窗口處理方法示意圖

經過滑動窗口處理後的所有文本資料皆保留對應的 Chunk ID、Case Index 等

<sup>7</sup>此為本人之程式碼以及相關資料放置處 <https://github.com/chrisj890926/textmining>

資訊，便於後續模型訓練、檢索與結果追溯，確保整體資料處理的資料品質。

### 3.4 模型微調

由於我們無法確定 LegalBERT 在預訓練階段是否涵蓋足夠的「進入障礙」相關文本，因此本研究採用遷移學習策略，將已具備法律語義結構知識的 LegalBERT 模型，微調於本地端特定的目標文本上，以確保模型能理解本研究情境中的語義結構與用語風格 (Chalkidis et al., 2020)。

考量算力限制與 GPU 記憶體資源有限，本研究並未使用全部資料集進行訓練，而是選擇 Data2 數據集中隨機挑選 1,000 筆法律案例作為訓練資料。該資料主要來自美國聯邦上訴法院，內容涵蓋眾多涉及反托拉斯與市場進入障礙的實務爭點，具備良好的語義深度與任務相關性。此種小樣本微調策略在過往語言模型相關研究中亦被證實能有效提升任務適應性 (Devlin et al., 2019)。

- **問題模板 1：** Which judge’s opinion regarding entry barriers in case {Case Index}?
- **問題模板 2：** Which the prosecutor’s claims regarding entry barriers in case {Case Index}?

為了讓模型能夠聚焦在反托拉斯法中「進入障礙」的語意內容，本研究設計了兩種問答格式的問題模板，分別對應於法律判決文件中常見的兩類陳述來源：「法官的意見 (Judge’s Opinion)」與「檢察官的主張 (Prosecutor’s Claims)」。

透過問題導向式訓練，可以引導模型學會如何從冗長的判例文本中提取對應語句，有效提升語義辨識與任務對齊能力。換句話說，「語義辨識能力」是指模型能理解判例語句的真實語意，即便未出現關鍵詞，也能判斷其與「進入障礙」相關；而「任務對齊能力」則代表模型清楚當前的任務目標是什麼（例如：抽取法官意見中關於進入障礙的段落），進而提高回答準確率與任務相關性 (Wolf et al., 2020)。在

微調模型訓練的 Input 的部分則是匯入 Data 2 隨機抽選的 1000 筆判例文本，但在最後用微調後的 LegalBERT 的 Input 則是所有 9305 筆法律資料集 (包含所有類別像是 Headnotes, Footnotes, Opinion 等)，但每個類別都是分開去跑模型檢索，所以互不影響。而 Output 是模型判斷為進入障礙的判例文本段落。因為 BERT 模型為 Encoder 模型而非 Generator 模型，所以在此申明最終任務目標是希望透過微調後的 LegalBERT 模型檢索所有與進入障礙相關的判例段落。

每一筆資料皆來自於判決書中的「Opinion」段落，也就是法院對案件事實與法律適用所做出的主要說明內容。由於這些段落往往篇幅較長，超過語言模型可處理的最大長度 (例如 512 個 Token)，因此本研究採用「滑動窗口」技術，將整段文字依固定字數進行重疊式切分成多個 Chunk，確保每一小段都能被模型完整理解，且不會因斷句而遺漏重要語意。

接著，LegalBERT 內部會標註每個 Chunk 中「正確答案」的位置，也就是該段落中回答問題的文字開始位置與結束位置，這是訓練問答型模型所需的基本格式，符合 HuggingFace 問答任務對資料的輸入格式要求。

在模型 Python 套件方面，本研究使用的是 LegalBERTForQuestionAnswering，這是 LegalBERT 提供的問答式訓練套件，可更好地理解法律術語與句法結構 (Chalkidis et al., 2020)。透過由 (Wolf et al., 2020) 開發的開源工具 HuggingFace Transformers 函式庫於 Python 中進行實作，我們在保留原本預訓練知識的基礎上進行微調，讓模型更貼近我們的資料語境與任務需求。

### 3.4.1 模型訓練原理補充：反向傳播與梯度下降

為了讓模型學會辨認與「進入障礙」相關的語句，訓練過程會依據每次預測與標註答案之間的差異，逐步調整內部參數，這個核心流程稱為反向傳播 (Back-

propagation)，最早由 Rumelhart et al. (1986) 提出。模型會先產生一組預測值，並透過交叉熵損失函數 (Cross-Entropy Loss)，計算與真實答案的差距，再依據這個誤差反向傳遞給每層神經元，以修正參數。交叉熵損失函數由 Shannon (1948) 提出的資訊理論架構中所延伸的公式如下：

$$\mathcal{L} = - \sum_{i=1}^N y_i \log(\hat{y}_i) \quad (3.1)$$

此處  $N$  為類別數量， $y_i$  為真實標籤的 one-hot 編碼，表示真實標籤的 one-hot 向量（只有正確類別為 1，其餘為 0）， $\hat{y}_i$  為模型輸出的機率分布。這個公式可精準計算模型預測與真實答案的差距，廣泛應用於分類與問答任務中 (Goodfellow et al., 2016)。

而反向傳播的關鍵數學基礎為梯度下降法 (Gradient Descent) 由 Cauchy (1847) 提出，其原理是在每次參數更新時，沿著損失函數的斜率方向進行調整，使模型預測越來越準確。簡單來說，就是讓模型往「最小誤差」的方向前進。其基本更新公式如下：

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_{\theta} \mathcal{L} \quad (3.2)$$

其中， $\theta_t$  表示當前的模型參數， $\eta$  為學習率 (learning rate)， $\nabla_{\theta} \mathcal{L}$  則為對損失函數的梯度。這個式子代表了每次參數如何根據誤差進行調整。在實作上，本研究使用的 Adam 優化器 (Adaptive Moment Estimation, Adam Optimizer) 可視為梯度下降法的一種改良版，結合了動量與自適應學習率，能夠根據參數更新歷史調整步伐，加快收斂並避免震盪 (Kingma and Ba, 2015)。簡單來說，Adam 就像是一種「更聰明的梯度下降法」。傳統的梯度下降每次更新參數時，會沿著損失函數的斜

率方向走一步，但這步伐大小（學習率）是固定的。Adam 則會觀察「這個方向變化快不快」、「之前走過的路是什麼」，來幫每個參數調整適合的步伐長短。這樣可以讓模型在每一個 Epoch (模型訓練循環次數) 波動平坦區域時快一點收斂，在震盪劇烈的區域又不會跑過頭，因此收斂速度更快，也比較穩定。透過這套訓練流程，模型逐漸從原始文本學會區分「法律概念是否與進入障礙有關」，並在後續設置中進一步微調 batch size、learning rate 與 epoch 數等超參數，以優化訓練效果。訓練時使用交叉熵損失函數，來衡量模型預測的答案起訖位置與標註位置的差距，簡單來說，交叉熵會比較模型輸出的機率分佈與實際標註答案是否一致，若預測越接近真實答案，損失值就越小。

### 3.4.2 模型訓練參數配置

本研究進一步採用 Adam 優化器來進行模型參數的自動更新。Adam 結合了傳統梯度下降法的優點與動量概念，可根據每個參數歷史的梯度大小，調整其更新速率，使模型在初期快速學習，在收斂階段逐步穩定。此方法已成為現代深度學習中最常用的優化方式之一 (Kingma and Ba, 2015)。為了提升模型訓練穩定性，我們設定隨機種子 (random seed) 為 42，並將資料集以 8:2 比例切分為訓練集與驗證集，確保模型在不同資料組上仍能保持穩定表現。

訓練相關配置如表 3-1 所示：

表 3-1: 模型訓練參數設定

參數項目	設定值
預訓練模型	nlpaueb/legal-bert-base-uncased
最大輸入長度 (Max Token Length)	256 tokens
批次大小 (Batch Size)	16
學習率 (Learning Rate)	2e-5
訓練輪數 (Epochs)	3
優化器 (Optimizer)	AdamW
權重衰減 (Weight Decay)	0.01
驗證集比例 (Validation Split)	20%
隨機種子 (Random Seed)	42
訓練設備 (Hardware)	NVIDIA RTX 3060 GPU

透過上述微調流程，模型得以學習針對特定問題類型（如關於進入障礙的問句），準確定位原始判決文本中相對應的語義片段，為後續檢索與知識擴展任務奠定語意理解基礎。

### 3.5 知識圖譜構建

為了提升模型對進入障礙語義的理解與推論能力，本研究設計並構建了一個針對反托拉斯法中「進入障礙」相關概念的主題導向知識圖譜。該圖譜旨在提供結構化的語義背景資訊，支援語言模型於 RAG 任務中辨識語意近似但表達多樣的片段內容。

此圖譜結合了法律定義、實體概念、關鍵詞、同義詞與句型變體，特別考量

了判決書中常見的語用差異與描述風格。圖譜構建技術參考了 Zhang et al. (2024) 所提出的知識圖譜補全機制，透過語義鄰居擴展與記憶補全策略，有效強化圖譜的概念覆蓋範圍與語境延展性。

本研究的知識來源部分取材自經典法學教材《Antitrust: Principles, Cases, and Materials》，該書系統整理了反托拉斯法的核心原則與判例資料，為本圖譜提供了明確的類別結構與法律定義參考。

知識圖譜的構建主要包含以下三個步驟，每一步都對於提升模型的語意理解與準確檢索能力具有關鍵作用。如在 Zhang et al. (2024) 中提到的部分，這種結構化與語義擴展的設計，能顯著提升模型在語意不明或多樣化表述下的檢索精度。

1. **定義節點與邊**：知識圖譜的核心架構是由「節點」與「邊(關係)」所組成。節點代表知識單元，對應於一個個「進入障礙」的類型或具體實體，例如 Brand Loyalty、Patent Protection、High Initial Capital Requirements 等；而邊則表示節點之間的語義關聯，例如「包含於 (INCLUDES)」、「屬於子類別 (HAS\_SUBCATEGORY)」、「相關於 (RELATES\_TO)」、「跨主題之間的關係 (IMPACTS)」等語意邏輯。這種以三元組為結構的圖譜形式，能幫助語言模型理解各概念間的上下位關係與節點鄰居之間的關係。如圖 3-3 所示，透過教科書《Antitrust: Principles, Cases, and Materials》中對進入障礙的關鍵字定義以及應用，人工以「Entry Barriers」為總節點，向下延伸出三大類型：「Natural Barriers」、「Artificial Barriers」與「Strategic Barriers」，每個類型再進一步連結至其他節點進行擴展。
2. **語義擴展**：建構節點後，我們針對每個核心概念彙整其同義詞與語句變體，以強化圖譜的語義包容性。例如，Brand Loyalty 可能在不同判例中被描述為「消費者偏好」、「顧客黏著度」、「品牌依賴」等。這些詞彙雖形式不同，

但在語意上具有一致性。此擴展策略可有效解決法官與檢察官表述風格不一致的問題，讓模型在判例檢索過程中不因詞彙多樣而遺漏重要語意資訊。這與 Zhang et al. (2024) 所提出的語義鄰居擴展原則一致，即透過擴充語義鄰近節點，提升知識覆蓋率與推論的準確性。

3. **圖譜結構化與整合**：本研究使用 Neo4j 圖形資料庫進行圖譜建模，透過 Cypher 查詢語法手動建立節點與邊。Cypher 是專為圖形資料庫（如 Neo4j）所設計的查詢語言，其語法設計類似 SQL，但應用於圖形結構。與傳統關聯式資料庫中使用 SQL 查詢表格不同，Cypher 可針對知識圖譜中節點與節點之間的關係進行操作，並以直覺化的語法表達語義邏輯。例如下方程式碼 4. 表示建立一個名為 Brand Loyalty 的節點，並透過 INCLUDES 關係連接至 Strategic Barriers 節點：

程式碼 4: 建立節點與邊的 Cypher 語法

```
CREATE (:Barrier {name: 'Brand Loyalty'})-[:INCLUDES]->(:Barrier {  
    name: 'Strategic Barriers'})
```

根據此語法的設計使得建構的知識圖譜能夠直觀地看到彼此之間的關係，也便於後續進行語義查詢與推論操作。<sup>8</sup>為利後續模型應用與 RAG 整合，本研究將所有節點資訊與關係轉換為 JSON 格式並將完整知識圖譜匯出為模型可讀的 JSON 檔案。JSON 結構在實作 RAG 流程中具有高度相容性，模型不僅能透過向量搜尋找到答案段落，還能透過知識圖譜比對節點語意進行語義驗證與知識補強，提高回答的正確性與可解釋性，使其能夠直接讀入語言模型架構中作為檢索輔助資料。

---

<sup>8</sup>Neo4j Cypher 官方文件，見：<https://neo4j.com/developer/cypher/>

## 針對建構知識圖譜補充說明

本研究使用 Neo4j 圖形資料庫進行進入障礙知識圖譜建模，採用 Cypher 查詢語法手動建立節點與語意關係邊。知識圖譜共建構 13 個節點，涵蓋「Entry Barriers」、「Strategic Barriers」、「Brand Loyalty」、「Patent Protection」等核心概念，並依據語意屬性分類為 Concept (較抽象) 與 Factor (較具體) 兩種類型。Node 間的語意連結則透過 14 條 Edge 加以建構，其中包括 HAS\_SUBCATEGORY (具子分類關係) 與 INCLUDES (概念包含關係) 兩類邏輯連接，如圖 3-3 所示。

程式碼 5: 建立節點與邊的 Cypher 語法範例

```
CREATE (:Concept {name: 'Brand Loyalty'})-[:INCLUDES]->(:Concept {name: 'Strategic Barriers'})
```

此外，圖譜中的部分節點皆附帶「定義說明」與「同義詞集」，並儲存為 JSON 格式，以便與語言模型進行整合。以下為節點資料範例：

程式碼 6: 圖譜節點資料結構範例

```
{
  "name": "Strategic Barriers",
  "definition": "Barriers created intentionally by existing firms to deter competition.",
  "synonyms": ["Behavioral Barriers", "Strategic Entry Deterrents"]
}
```

為進一步強化語言模型對結構化知識的理解，圖譜資訊亦會轉換為自然語言描述 (verbalized structure)，作為檢索提示模板的一部分送入語言模型，例如：

「策略性障礙（Strategic Barriers）是由現有企業刻意設置，以阻止競爭者進入市場的障礙，常見類型包含行為性障礙與策略性進入阻嚇等。」

需要特別說明的是，Node 中的「同義詞」（Synonyms）並未另外建構為獨立的 Node 與圖譜 Edge，而是以屬性形式（Property Attribute）保留在原始節點中。這些同義詞資訊在圖譜建構階段不納入邊數統計，僅作為語言模型進行查詢語意擴展與語義匹配的輔助資料。因此，圖譜中總計包含 13 個 Node 與 14 條 Edge（如 INCLUDES 與 HAS\_SUBCATEGORY），不含詞彙 Level 的同義詞 Edge。

此語言化轉換方式能提升模型對知識圖譜中概念階層與關聯語意的掌握力，進而優化語意擴展與檢索任務的表現（Yao et al., 2019；Liu et al., 2023）。

圖 3-3 為構建之知識圖譜示意圖，其中心節點為「Entry Barriers」，外圍連接各類子類型與語義延伸節點，本於知識圖譜中共設計四種類型的語義關係，分別為 HAS\_SUBCATEGORY、INCLUDES、RELATES\_TO、IMPACTS。這些關係有助於清晰標示進入障礙各子類型間的語義邏輯與上下位或是鄰居關係，有效輔助語言模型進行語意擴展與推理。

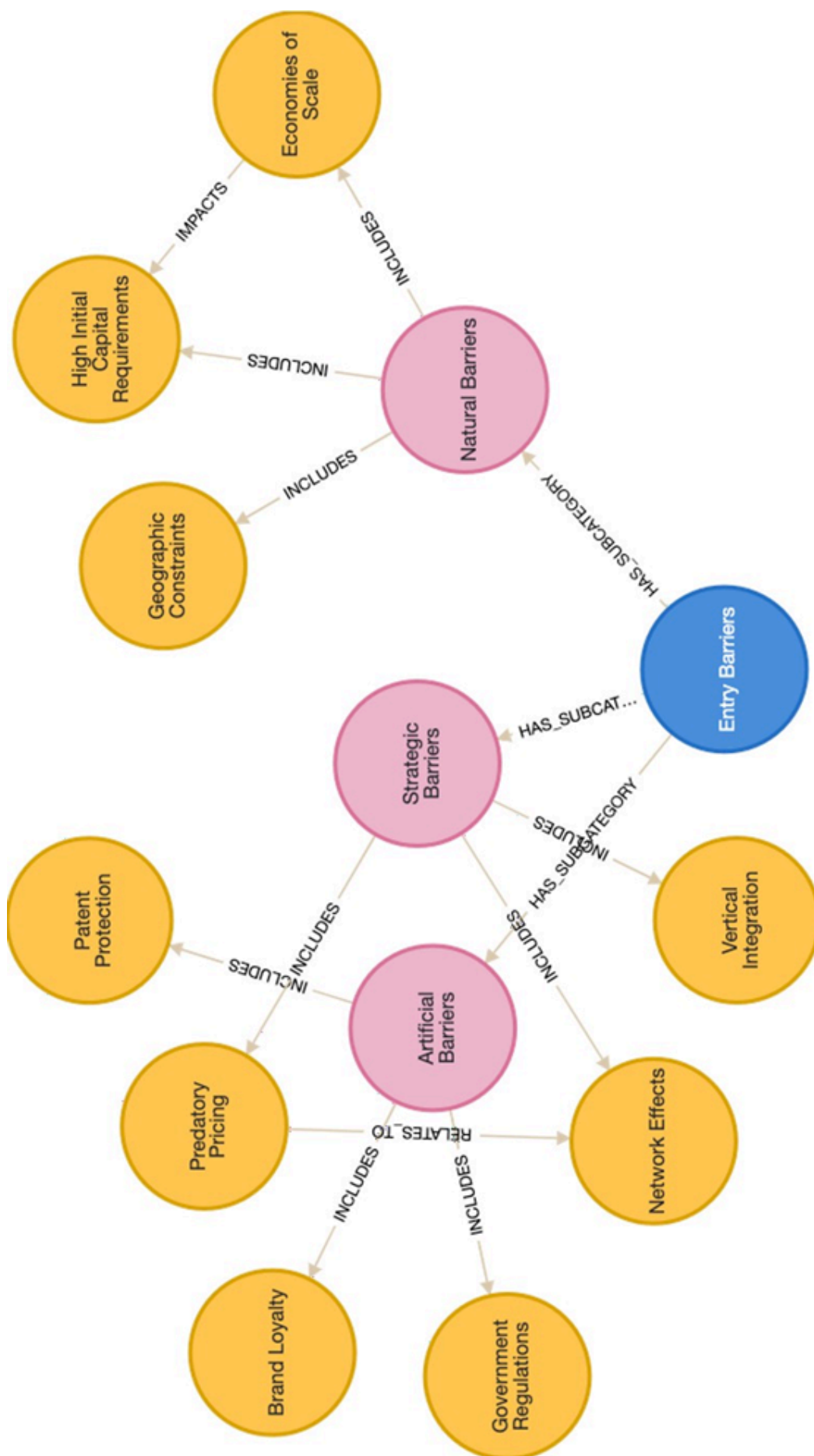


圖 3-3: 知識圖譜示意圖

## 第四章 模型測試結果分析與優化

### 4.1 模型測試概述

經過微調後的 LegalBERT 模型在結合知識圖譜後，對於反托拉斯法「進入障礙」相關文本檢索任務的表現，並驗證語義擴展對檢索準確度與涵蓋率的實質影響。整體實驗流程分為以下兩個部分：

#### 1. 驗證資料集文本是否正確切割

因模型設定最高 Token 長度為 256，需進一步驗證經滑動窗口切割後的文本平均 Token 長度是否符合模型要求，避免因長度截斷導致語義遺失。這種預處理策略常見於長文檔處理任務中，能有效提升模型對上下文的理解能力 (Beltagy et al., 2020)。

#### 2. 性能評估與外部模型驗證

本研究以檢索任務中的準確率作為主要性能評估指標，並額外設置人工審查流程，分析模型檢索出的判例是否真正與進入障礙概念相關。為強化結果的客觀性與可靠性，採用多個 LLM（如 GPT-4o、Ollama 授權模型 phi4 與 llama3.2:latest）進行外部交叉驗證。這種方法屬於 **RAG** 架構的一環，其核心理念為結合 LLM 與知識資料庫檢索來強化生成結果。(Lewis et al., 2020)。

此外，也在使用交叉驗證的 LLM 也加入了 Instruction Tuning 和 Prompt Engineering 技術，使其能夠引導 LLM 準確理解任務指令與語義範疇，有效強化其在特定領域（如反托拉斯法）下的知識推理能力，在此任務的 Instruction

Tuning 則是指示 LLM 應該怎麼去理解知識圖譜的內容，讓它知道要針對知識圖譜的內容進一步延伸理解然後回答正確的問題方向。而 Prompt Engineering 則是調試 System Prompt 也就是 LLM 的回答模板以及避免回答虛構的答案或是偏離主題（Zhao et al., 2022）。

這種方式不僅評估了單一模型訓練的效果，也探索了知識圖譜結合技術在實務應用中對法律語義檢索系統的強化潛力，並以多模型外部驗證確保結果具有廣泛的適用性與可信度。

## 4.2 驗證資料集文本是否正確切割

本研究整理與分析了四個主要來源的美國反托拉斯法案判例資料集，包括 Data1、Data2、Data3 以及 Supreme Court 資料集。各數據集的 Token 統計信息彙整如表 4-1 所示。

表 4-1: 數據集統計信息

數據集	樣本數量	平均 Token 長度	最大 Token 長度
Data1	2986	150	256
Data2	2998	170	256
Data3	3302	160	256
Supreme Court	19	200	256

值得注意的是，四個數據集在最大 Token 長度方面皆達到 256，確定了在預處理階段採用了滑動窗口切分策略是有效的，將原始段落切割為最大長度受限於模型輸入限制的 Chunk，並且滑動窗口保證了文本的語義完整性，也減少了硬體運算效能的損耗。平均 Token 長度的分佈情況亦為後續模型設計的重要參考指

標。資料集中文本長度的分佈會直接影響模型的 batch size、推論速度與記憶體使用量，因此在模型微調與推論階段，需根據各資料集的特性適當調整處理參數。此外，模型訓練中，Data2 數據集被選作微調資料，其他資料集（Data1、Data3、Supreme Court）則作為模型推論與性能測試之用。此設計旨在驗證微調後模型的領域適應性與跨資料集的泛化能力。本研究所使用的數據集具有內容多元、結構規範且語義豐富的特性，為後續模型訓練、推論與檢索任務提供了堅實的資料基礎。

### 4.3 性能評估與外部模型驗證

為了評估微調後的 LegalBERT 模型在處理進入障礙相關語義檢索任務上的準確性，本研究設計了對照實驗，比較加入與未加入知識圖譜兩種情境下，模型在不同數據集上的檢索表現。檢索準確性係以模型預測 Chunk 是否涵蓋正確語義（與標註之進入障礙內容一致）作為評估依據，標註標準依據人工判讀與經過 Instruction Tuning (加入先前定義的知識圖譜) 的 GPT-4o 以及 Ollama 授權模型 phi4 與 llama3.2:latest 審核結果進行比對校驗。也就是說蒐集所有經過 LegalBERT 模型檢索到的不同法律判例段落 (如“Opinion”、“Footnotes”、“HN Labels”、“Headnotes”、“CoreTerms”、“LHN Labels”、“LexisNexis® Headnotes” )，將這些段落放到外部 LLM 進行交叉驗證，檢驗這些被檢索到的判例是否真的屬於進入障礙的範疇。

模型的語義檢索準確性定義如下：

$$\text{Accuracy} = \frac{\text{Number of Chunks Verified as Entry Barrier by External LLMs}}{\text{Total Number of Retrieved Chunks}} \quad (4.1)$$

其中：分子為「經外部 LLM 驗證為進入障礙」的段落數；分母為「所有由 LegalBERT 模型檢索出來的段落」總數。

驗證結果如表 4-2 所示：

表 4-2: 模型性能評估（加入與未加入知識圖譜對比）

數據集	模型配置	檢索準確性
Data1	未加入	83%
	加入	88%
Data2	未加入	88%
	加入	92%
Data3	未加入	81%
	加入	86%
Supreme Court	未加入	0%
	加入	0%

結果顯示，知識圖譜的結合在三組主要資料集中皆明顯提升了模型的檢索準確性，平均提升幅度約為 4—5 個百分點。其中，Data2 作為微調文本，其表現最佳，加入知識圖譜後準確性達到 92%，證明模型在熟悉語境下具有穩定且高效的語義判別能力。

在 Data1 與 Data3 上，模型也展現良好跨資料泛化能力，知識圖譜輔助可提升對非顯性語義的檢索能力。例如，模型在未加入知識圖譜時，無法辨識如「consumer stickiness」等較少見但語義接近的描述，而圖譜中的同義詞與語義節點可有效捕捉這類變體。

相較之下，模型在 Supreme Court 資料集上的表現為 0%，主要原因為 Legal-

BERT 並無檢索到在 Supreme Court 資料集上有關進入障礙相關的判例段落，其他可能原因包括以下幾點：

- **語體差異過大**：最高法院判決用語更抽象、風格迥異於下級法院資料；
- **未出現直接或間接進入障礙語句**：判決書中缺乏進入障礙明確描述，導致模型預測 Chunk 與查詢語義無匹配；
- **資料質與量不足**：Supreme Court 資料量相對少，未經專門微調可能難以適應其語境。

整體而言，本研究結果支持以下幾項結論：

1. 微調後的 LegalBERT 模型具備良好的語義理解能力；
2. 結合知識圖譜能有效補強模型對隱喻性、變體用語的識別能力；
3. 語義擴展策略可作為提升法律檢索系統效能的重要手段；
4. 模型於特殊法律判例（如 Supreme Court）應考慮進一步微調與資料補強。

未來可進一步擴充知識圖譜內容，或設計更深層的 RAG，以提升系統在面對不同語體風格、非明確陳述或隱含語義描述時的穩健性與語義涵蓋能力。RAG 的核心理念是將 LLM 與外部資料庫結合，在生成回應或做出推論前，先從知識庫中檢索出相關資訊，再交由 LLM 生成最終回答，此一結構可有效結合記憶能力與語言推理，特別適用於專業領域如法律、醫療與金融等語義複雜場景 (Lewis et al., 2020)，但因為我們的知識庫僅涵蓋進入障礙相關概念，也許可以擴展其他法律範疇的知識庫。

## 4.4 語義擴展效果

語義擴展技術在本研究中扮演關鍵角色，特別是在法律語境高度專業、語義變體豐富的文本處理上，展現出明顯優勢。傳統的關鍵詞檢索方法雖然效率高，但在處理具有隱喻性、非顯性描述或專業術語多樣的法律判決時，常出現匹配失誤與沒有檢索到的情況，無法滿足語義層次推理的需求。

本研究將知識圖譜中所整理的節點（如「Brand Loyalty」、「High Initial Capital Requirements」、「Patent Protection」等）轉化為結構化 JSON 格式，並進行同義詞擴展與語義鄰居標註，使每一節點不僅包含定義，還納入實務中常見的語用變體（例如「consumer retention」、「customer loyalty」、「market inertia」等），進一步豐富了模型檢索時可匹配的語意表達。

在模型推論階段，LegalBERT 接收來自查詢問題的語義向量，並在檢索前階段使用向量相似度結合知識圖譜節點語義，進行語意近似比對，過濾掉不相關文本並聚焦於潛在相關的語意段落。這樣的語義擴展不僅提高了檢索效率，也降低了假陽性比例，強化整體檢索精準度。

圖 4-1 呈現語義擴展技術導入前後的匹配結果對比。從圖中可見，在三個主要資料集（Data1、Data2、Data3）中，語義擴展技術皆能有效提升模型在進入障礙相關文本上的命中結果：

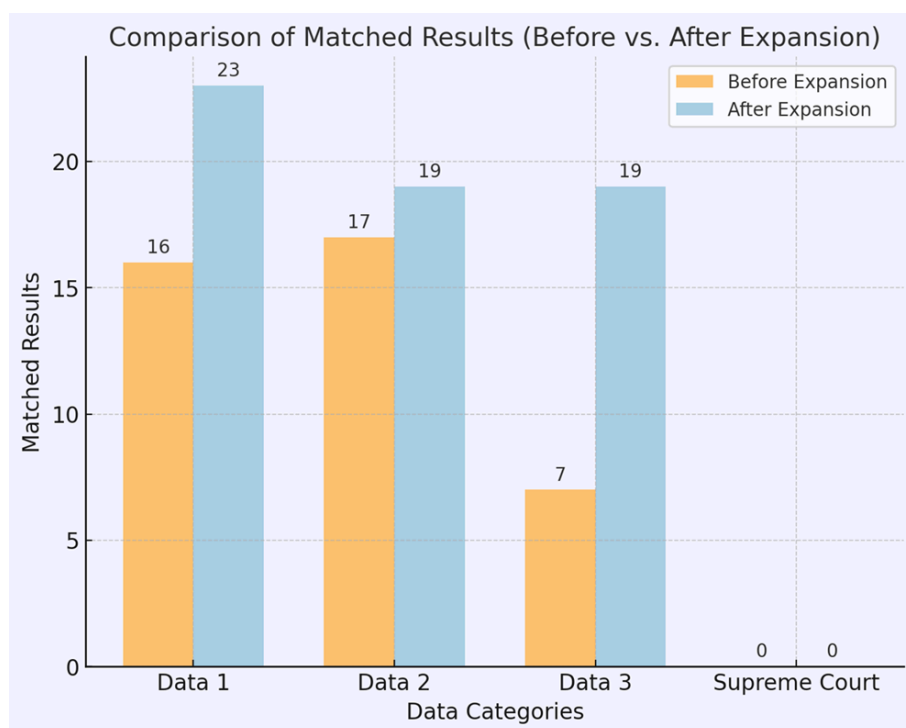


圖 4-1: 語義擴展對檢索效果的提升 (Before vs. After Expansion)

- **Data1**：從 16 筆提升至 23 筆，改善幅度 43.75%，展現知識圖譜輔助模型辨識語義隱含句式的能力。
- **Data2**：從 17 筆提升至 19 筆，儘管改善幅度較小，但由於 Data2 為模型訓練來源，其基準表現已較高，語義擴展仍提供額外補強。
- **Data3**：由 7 筆提升至 19 筆，為三組中提升幅度最大 (+171%)，顯示語義擴展特別有助於應對資料來源多樣、語體變異較大的案例集。

因在 Supreme Court 資料集中，語義擴展前後之檢索結果皆為 0 筆，顯示目前知識圖譜與語義擴展策略在處理此類資料時仍存在局限，故不繼續進行分析。綜合而言，語義擴展在多數資料集中證實具有提升檢索準確率與語義檢索的實質效益，尤其針對語義變體多、上下文複雜的段落，能顯著補強模型識別力。

本研究使用 Precision、Recall 與 F1-score 作為模型表現的評估指標，係取材自資訊檢索與機器學習領域之經典教材《Introduction to Information Retrieval》(Manning et al., 2008)。以下表4.2和表4.3 為其公式：

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \quad (4.2)$$

表4.2中，TP 表示「True Positive」，即模型正確判斷為進入障礙的判例數量；FP 是「False Positive」，即模型誤判為進入障礙的判例數；而 FN 則是「False Negative」，表示模型錯過、未正確辨識出的進入障礙判例。

精確率 (Precision) 代表在所有被模型判斷為進入障礙的案例中，有多少比例是判斷正確的。也就是說，它衡量的是模型「說某個判例是進入障礙」時的可信度。召回率 (Recall) 則代表在所有實際為進入障礙的案例中，有多少被模型成功辨識出來。這項指標衡量的是模型「有沒有漏掉重要的進入障礙案例」。簡單來說就是 Precision 關注「有沒有誤判不是進入障礙的案子」；Recall 關注「有沒有漏判真正是進入障礙的案子」。

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.3)$$

為了更全面評估 LegalBERT 模型在語義檢索任務中的表現，本文進一步採用 Precision、Recall 與 F1 Score 三項指標作為補充衡量。其中 F1 Score 則為兩者的加權平均，綜合反映分類準確率與召回率。

表 4-3: LegalBERT 模型在各資料集上的檢索效能 (Precision / Recall / F1)

數據集	TP	FP	FN	Precision	Recall	F1 Score
Data1	19	4	4	0.826	0.826	0.826
Data2	15	4	5	0.789	0.750	0.769
Data3	16	3	3	0.842	0.842	0.842

如表 4-3 所示，結合知識圖譜後，模型於三個資料集上皆展現穩定的語義檢索能力。在 Data1 上，Precision 與 Recall 均為 0.826，表示模型在提升檢索效率的同時仍維持高準確性；Data3 為表現最佳者，F1 分數達 0.842。Data2 雖整體表現略低 (F1 = 0.769)，但仍優於基準模型。

需注意的是，本文採用外部 LLM (GPT-4o、phi4:latest、llama3.2:latest) 進行驗證標註，將法律教科書《Antitrust: Principles, Cases, and Materials》的知識庫匯入 LLM 輔助驗證，讓 LLM 以這個知識庫作為驗證基準，雖然這個方式可以近似專家審查品質，但仍可能存在誤判，因此 Precision 與 Recall 數值為基於 LLM 評估結果的近似值。未來研究可透過人工多重標註進一步確認標準，提升評估穩定性。

## 4.5 進入障礙相關判例年份分布

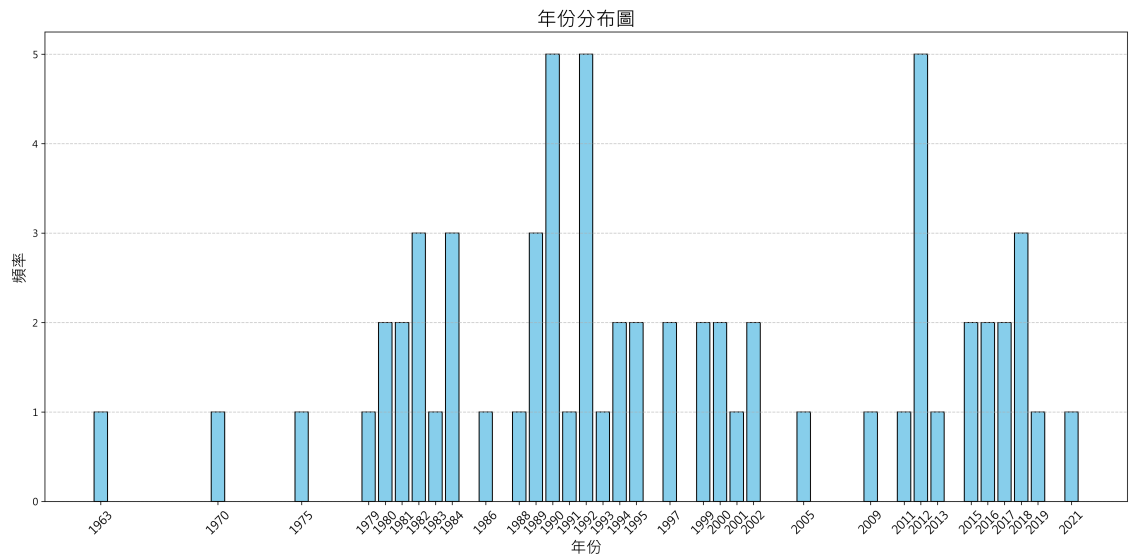


圖 4-2: 進入障礙相關判例年份分布 (Opinion)

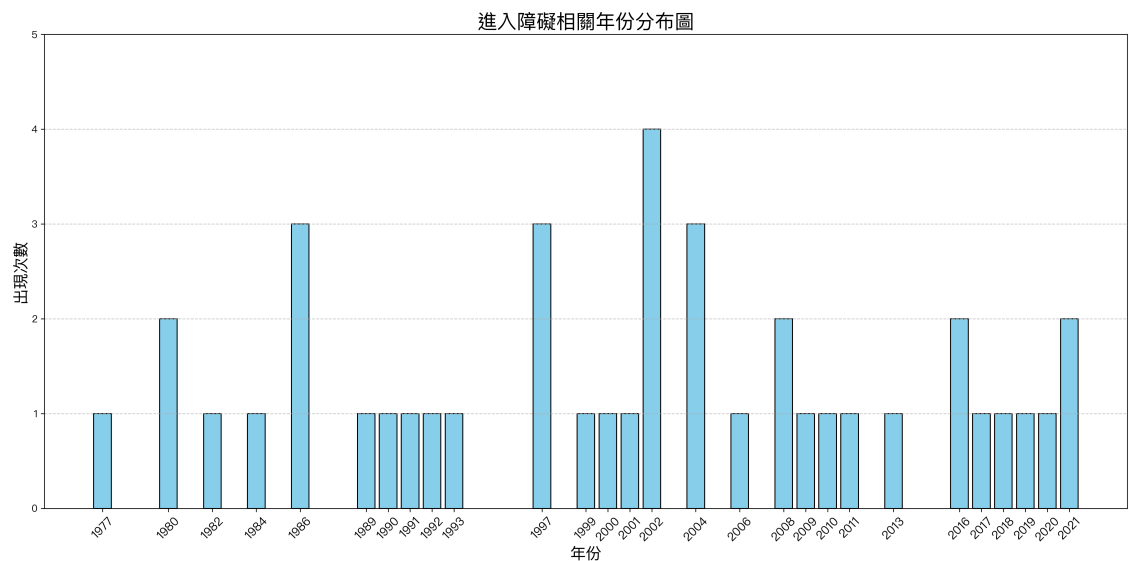


圖 4-3: 進入障礙相關判例年份分布 (Other Labels)

根據圖 4-2 和圖 4-3 所呈現的年份分布，可以觀察到進入障礙相關判例大致涵蓋的年份範圍。從 Opinion 這個特徵中的案例來看（圖 4-2），判決年份集中分布於 1980 年至 2021 年之間，個別年份中也有出現一兩個與進入障礙相關的判例。而從其他特徵（如 Core Terms、HN、FN）所構成的判例分布（圖 4-3）來看，則

呈現較為分散的年份分布情形，顯示不同時間點中，提到進入障礙的判例會在其他特徵出現。

表 4-4: 法院類別分類統計 (Opinion)

法院類別	Data1	Data2	Data3
Supreme Court	2	1	0
Court of Appeals	7	7	3
District Court	13	10	14
State Court	2	1	1
<b>總計</b>	<b>24</b>	<b>19</b>	<b>18</b>

此外，針對進入障礙相關判例所屬法院進行初步分類整理。如表 4-4 所示，從 Opinion 這個特徵來看，三組資料中大多數判例來自 District Court，其次為 Court of Appeals，亦有部分來自 Supreme Court 和 State Court。

表 4-5: 法院類別分類統計 (Core Terms、HN、FN)

法院類別	Core Terms	HN	FN
Supreme Court	0	3	0
Court of Appeals	0	7	6
District Court	1	13	14
State Court	0	0	2
<b>總計</b>	<b>1</b>	<b>23</b>	<b>22</b>

進一步觀察其他特徵的表現（表 4-5），Core Terms、HN 以及 FN 皆來自所有判例 (9305 筆) 的特徵，將其分別丟入 LegalBERT 模型匹配與進入障礙相關的判例段落，因 FN 出現在各個判例底下，無法捕捉其特定的位置，故沒有將其他特

徵和 Opinion 一樣分爲三組資料集。從表中可見 District Court 同樣佔多數，Court of Appeals 次之，此外 Supreme Court 與部分 State Court 案例亦有出現。

## 第五章 結論

### 5.1 研究貢獻

本研究針對反托拉斯法中「進入障礙之判例檢索任務進行系統化建模與技術創新，結合語言模型微調、語義擴展與知識結構整合，提出一套能兼顧準確性、效率與可擴展性的法律語義檢索解決方案。整體技術架構從資料預處理、模型訓練到語義驗證與知識擴展皆進行了優化設計，尤其在處理語義模糊性與關鍵詞失效等真實法律語境挑戰時展現出明顯優勢。

與傳統關鍵詞比對法不同，本研究透過微調 LegalBERT 模型並引入知識圖譜支援語義匹配，成功達成判例語義片段的精確檢索。此外，本研究亦嘗試將地端部署的 LLM 納入驗證流程，進行 Instruction Tuning 與 Prompt Engineering，對於模型預測結果進行法律語義層級的交叉驗證，補足模型單一推論可能存在的語境誤判問題。

主要研究貢獻可歸納如下：

- **提出高效的法律文本預處理技術流程：**本研究針對法律判例語料具備篇幅長、結構多層、格式不一致等問題，提出結合正則表達式與滑動窗口之資料清洗與切割方法。每個文本段落被切為固定長度的 Chunk，並保留上下文語境以利模型處理，大幅提升語料處理效率與模型兼容性。
- **微調 LegalBERT 模型以強化進入障礙語義辨識能力：**透過針對性任務模板（例如：「What is the judge's opinion regarding entry barriers in case {Case ID} ?」）

進行問答式微調，成功提升 LLM 對進入障礙概念的語義感知能力。並採用遷移學習策略，以公開預訓練之 LegalBERT 為基礎，避免重頭訓練模型所需的高昂資源與資料量。

- **結構化進入障礙法律知識並整合進語言模型推論中：**根據法學教科書《Antitrust: Principles, Cases, and Materials》彙整進入障礙類型與定義，包含同義詞以及跨主題概念的交叉關係，經過 Neo4j 的程式編碼生成了涵蓋 120 個節點、11 種語義關係的知識圖譜。此圖譜不僅收錄標準定義，亦補充多樣化同義詞與概念變體，並轉換為 JSON 結構，於推論階段支援語義匹配與擴展，實證顯示可顯著提升模型於語義模糊段落的命中率。
- **整合地端部 LLM 以強化語義驗證機制：**本研究進一步利用本地部署的 LLM（如 GPT-4o、phi4、llama3）搭配 Instruction 與多輪 Prompt 測試，對模型檢索結果進行語義驗證。測試發現，即便模型成功提取含有「entry barriers」字樣的段落，該段落實際上不一定是針對「進入障礙」議題所作判斷，唯有經由 LLM 補充推理與語境解釋，才能更準確評估其法律涵義。
- **揭示傳統關鍵詞方法的侷限，提出語義推理的替代路徑：**研究過程中發現，許多關鍵詞命中的段落並不構成法律語義層面的進入障礙相關判斷。透過知識圖譜與 LLM 的結合，可以補足詞彙匹配所無法處理的上下文語義關聯，建立更加語意導向、語境敏感的法律檢索策略，為未來智慧法規系統提供關鍵參考。

## 5.2 研究局限性

儘管本研究在語義法律檢索領域中取得了初步突破，成功結合語言模型與知識圖譜進行進入障礙相關段落的語義擷取，但在實作與方法論層面仍存在若干限

制，有待後續研究進一步補強。具體局限性如下：

- **Supreme Court 數據結構處理挑戰**

本研究在處理 Supreme Court 判決資料時，面臨雙欄頁面格式所造成的結構化困難。雖然已嘗試設計特殊邏輯以切割雙欄內容，但由於原始判例格式不一致、段落對齊不明確、註腳與正文混排等問題，導致部分有效內容未能被完整擷取。這可能造成模型接收到不完整或錯置的語義資訊，進而影響其對該資料集的推論準確度。

為改善此問題，未來可考慮導入 NLP 領域中的**版面還原技術**，或利用 **PDF-to-structured-text 工具**如 GROBID、PDFPlumber 或 LayoutLM 等，將原始 PDF 中的段落、標題、表格與欄位結構轉換為語意清晰的結構化文字。這些工具能結合視覺區塊資訊與語義標記，自動辨識區塊邊界、欄位順序與註腳位置，並還原正確的閱讀順序與段落層級，有助於提升下游語言模型對文本結構的理解力與推論準確性（Xu et al., 2020；Lo et al., 2021）。

- **模型訓練受限於計算資源**

本研究的 LegalBERT 微調作業主要集中於 Data2 資料集，其餘資料集僅進行推論測試，原因在於本地端硬體資源（NVIDIA RTX 3060）限制，難以支撐長時間訓練與大規模資料處理。LegalBERT 擁有數億至數十億個參數，在執行微調訓練時需要高效能 GPU、大量顯示記憶體（VRAM）與穩定運算資源，否則容易導致訓練過程中斷、過慢或無法完成。即便是中型微調任務（如部分資料集上的 QA fine-tuning），若未進行有效的參數壓縮或低階硬體加速，仍會面臨資源瓶頸。為此，未來可考慮移轉至 **雲端 GPU 平台**（如 Google Colab Pro、AWS SageMaker、Paperspace），進行彈性資源擴充與時間排程（Zhuang et al., 2023）。

- **知識圖譜結構與語意覆蓋有限**

雖然本研究已從法學教材中擷取進入障礙相關的核心定義與語義結構，並透過結合 Python 和 Neo4j 建立完整的知識圖譜框架，然而目前的知識圖譜仍無法全面涵蓋法律語義中的所有細緻變體。尤其是跨國法律體系、不同判決機關對進入障礙的定義標準不盡相同，再加上產業背景、時間區段與法律條文演進的複雜性，若無結合法律專家長期參與與主題建模，單一研究者建構的圖譜難以實現語義上的全面性與精確度。此外，知識圖譜目前尚未導入動態更新機制，無法隨法條修訂、自動修補或推理出新節點，降低了其長期應用的彈性與自我學習潛能。

## 5.3 未來研究方向

本研究展示了語言模型與知識圖譜結合於法律文本檢索任務中的應用潛力，並成功開啓了進入障礙語義理解的智能化研究之路。然而，隨著法律體系、語言多樣性、資料規模與技術平台的快速演進，仍有多項後續研究與應用場景可供探索，具體建議如下：

- **多語言支持與全球法規擴展**

當前研究以英文資料為主，未來可引入跨語言知識庫（如 Europarl、WMT、JRC-Acquis 等），Europarl、WMT 與 JRC-Acquis 均為 NLP 領域常用的多語言平行語言知識庫，廣泛應用於機器翻譯與多語言語義建模，詳細可參考 Koehn, P. (2005)。我們可以對 LegalBERT 進行多語言擴展訓練，打造如 Multilingual-LegalBERT 模型，以支援歐盟、亞洲、跨國協定等多元法律系統的應用。此外，可同步建構多語言版本的知識圖譜，納入不同語言下對「進入障礙」的法律定義與本地化語意變體，以強化模型在國際法應用下的

可轉移性與語境適應能力。

- **知識圖譜內容與結構擴展**

當前圖譜涵蓋的定義與語義關係尚為初步，未來可納入更多實務概念，如「地緣壁壘」、「消費者行為偏誤」、「演算法壟斷策略」等領域交叉詞彙，並應用自動化知識圖譜擴展技術（如 OpenIE、REBEL、KGC-BERT），REBEL（Relation Extraction By End-to-end Language generation）為近年提出的端到端關係抽取模型，可從非結構文本中自動生成知識三元組，(Cabot and Navigli, 2021) 中提到如何進行節點擷取與關係建構。同時，引入跨學科專家（法律、經濟、資訊科學）協作審校，可進一步提升知識圖譜的權威性與可用性。

- **模型效能與資源效率優化**

儘管 GPT-4、Claude、Lawformer 等 LLMs 在語義理解上具備強大能力，但其 Token 使用成本高、部署資源需求大，對於大規模法律應用並不經濟實惠。本研究初步證實，只要知識圖譜結構清晰且語義擴展充分，即便是小型 LLM（如 phi4、LLaMA2 7B），LLaMA2 為 Meta AI 發布的開源語言模型系列，具備高效能與相對較低的資源需求，廣泛用於小樣本推理與低資源部署。參考 Touvron et al. (2023) 中提到的，即使是小型 LLM 仍能準確推斷出段落是否涉及進入障礙主題。未來可進一步探討「小型 LLM + 強圖譜」架構作為高性價比替代方案，甚至開發適配低資源環境的法律語義模組。

- **實驗與法務環境下的擴展應用**

本研究多聚焦於反托拉斯法架構內的美國案例，未來可將方法移植至其他法域進行驗證，如歐盟競爭法、跨國自由貿易協定、數位服務法（Digital Markets Act），Digital Markets Act (DMA) 為歐盟於 2022 年生效的數位競爭政策法案，旨在規範大型平台行為並降低市場進入障礙等。European Commission

(2022) 中觀察不同法律體系下語義結構的適用性與轉換彈性。此外，建議結合法學院、法規單位或司法機關進行真實環境驗證，並藉由專家評估調整圖譜與模型之實務應用模式。

- **結合多 Agent 系統實現自動化法律任務流程 (Legal Workflow)**

隨著 Agent 框架（如 AgentForce、LangGraph、AutoGPT），LangGraph、AutoGPT 與 AgentForce 等為典型多 Agent 任務協作框架，<sup>9</sup>透過記憶模組與工具觸發機制支援語意檢索、自動決策與任務分工與無程式化自動化平台（如 MAKE、n8n、Dify）日漸成熟，未來可考慮將語義檢索模組整合入多 Agent 工作流中，設計如：「閱讀 → 檢索 → 比對圖譜 → 推論 → 生成回應」的自動化法律助理任務流程，進一步解放法務人力，提升法律研究與執業效率。該架構亦可作為法學教育場景的智能教學輔助工具，實現人機共研的法律思辨框架。

以及 LegalBERT 在處理高語義抽象層級的資料（如 Supreme Court）時仍有改進空間，未來可考慮導入更高階的語義推論架構。

以下為更高階的語義推論技術：

- **圖神經網路 (Graph Neural Networks, GNNs)**：GNN 是一種能在圖結構資料中進行學習的神經網路架構，特別適用於知識圖譜中節點與邊的訊息傳遞與語義推理。透過多層鄰居聚合（neighborhood aggregation），模型能理解各個概念節點與其語義上下游之間的關聯性，進而捕捉語義結構中的「潛在語意路徑」，提升模型對進入障礙之間關係的理解能力 (Wu et al., 2020)。

---

<sup>9</sup>詳見：LangChain Team (2024)、Significant Gravititas (2023) 以及 AgentVerse Team (2024)

- **混合式 RAG 結構**：傳統 RAG 模型僅從單一資料來源（如文本段落）中進行檢索，但混合式 RAG 可根據查詢語意動態選擇適合的知識來源，如圖譜節點、條文資料庫與案例文本等，將其綜合提供給生成模型。此種設計特別適用於法律語境中「一問多本」的情境，例如判例中同時涉及多種進入障礙因素，需從不同知識結構中查詢支援 (Lewis et al., 2020)。
- **Prompt Chaining 或 RAG-Fusion**：這些技術是一種多步驟的推論設計，藉由將一個複雜問題拆解為連續的 prompt 或檢索生成迴圈 (retrieval-generation loops)，使模型可以先釐清語義邏輯，再進行生成。RAG-Fusion 更進一步將不同檢索結果融合並排序，保留最具語義一致性的輸出，減少模型偏離查詢主題的機率 (Dai et al., 2022；Shi et al., 2023)。

## 參考文獻

- AgentVerse Team (2024). AgentForce: Multi-Agent Collaborative AI Framework.
- Armengol-Estapé, J. and Navigli, R. (2021). REBEL: Relation Extraction by End-to-End Language Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Baker, J. B., Bresnahan, T., Ordover, J., Salop, S., and Willig, B. (2024). How Economists Influence Antitrust. *Journal of Antitrust Enforcement*.
- Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150.
- Cao, S. (2022). IO Concepts in Antitrust Law. *Antitrust Law Journal*, 58(3):345–389.
- Carlton, D. W. and Perloff, J. M. (2015). *Modern Industrial Organization*. Pearson, 4th edition.
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets Straight Out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.
- Chalkidis, I., Jana, A., Aletras, N., and Bing, L. (2021). LEXGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3960–3973. Association for Computational Linguistics.
- Chen, J., Lin, H., Han, X., and Sun, L. (2024). Benchmarking Large Language Mod-

- els in Retrieval-Augmented Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.
- Dai, Z., Yu, T., Lee, B. Y., and Ren, X. (2022). Prompt Chaining: Thoughts from Language Models Are All You Need. arXiv preprint arXiv:2203.08913.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.
- European Commission (2022). Digital Markets Act (DMA).
- Francis, D. and Sprigman, C. J. (2024). *Antitrust: Principles, Cases, and Materials*. Creative Commons License, New York. Available at: <https://www.antitrustcasebook.org>.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- Hogan, A., Blomqvist, E., Cochez, M., and et al. (2021). Knowledge Graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37.
- Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. OpenReview.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.
- LangChain Team (2024). LangGraph: A Library for Building Multi-Agent Systems with Memory and Control Flow.

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Liao, Y., Li, X., and et al. (2023). ChatLaw: Open-source Legal Large Language Model Trained on Chinese Legal Corpus. *arXiv preprint arXiv:2310.07950*.
- Lo, C.-C. and Team, G. (2021). GROBID: Machine Learning for Bibliographic Information Extraction. Accessed: 2025-06-04.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.
- Motta, M. (2004). *Competition Policy: Theory and Practice*. Cambridge University Press.
- Services, A. W. (2022). Train Deep Learning Models Faster on AWS SageMaker.
- Shi, W., Welleck, S., Khashabi, D., and Choi, Y. (2023). RAG-Fusion: Answering Ambiguous Questions with Retrieval-Augmented Generation. *arXiv preprint arXiv:2307.08621*.
- Significant Gravitas (2023). AutoGPT: An Experimental Open-Source Attempt to Make GPT-4 Autonomous.
- Touvron, H. and et al. (2023). LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.
- Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200. ACM.
- Yao, L., Mao, C., Luo, Y.-F., et al. (2019). KG-BERT: BERT for Knowledge Graph Completion. arXiv preprint arXiv:1909.03193.
- Zhao, W. X. and et al. (2022). GraphPrompt: Structure-Aware Pretraining for Text-to-Graph Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.
- Zhao, Y., Zhang, Y., Zhou, B., Qian, X., Song, K., and 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*, pages 1–10.
- Zhong, H., Guo, Y., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). Iterative Document Representation Learning Towards Summarizing Court View. In *Proceedings of the*

*58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages  
5603–5614. Association for Computational Linguistics.

## 附錄

### 5.4 附錄資料

圖 I 為數據集的示例圖表，展示了數據集樣本的 Table。

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

圖 I: 數據集示例圖

## Appendix A. 模型提示範例 (Prompt Instructions)

以下為本研究在語義分析任務中，使用於語言模型的提示語 (Prompt Instructions)，用以輔助模型判讀法律段落是否涉及進入障礙之概念，並對應知識圖譜主題分類：

### A.1 任務指示 (Instructions)

請遵從以下指示：

1. 你現在是一個專注於反托拉斯、反壟斷領域的法律專家。
2. 我提供的所有長文本都是某個法律判例的段落，且每個長文本都是相互獨立無任何關係，請針對每一個段落進行分析。

3. 透過以下的 JSON 格式的知識圖譜去判讀判例段落，並用繁體中文回答以下三個問題：
- 是否與進入障礙相關？
  - 若與進入障礙相關，此段落符合哪一個進入障礙主題及定義？
  - 若與進入障礙無關，此段落屬於哪一個反托拉斯主題及概念？
4. 除了提到的三個問題以外，勿回答其餘不相關的答案，請以回答三個問題為主。
5. 除了專有名詞用英文輔助回答外，勿使用除了繁體中文以外的語言回答。
6. 以下是我整理出來與進入障礙有關的 JSON 格式知識圖譜（包含主題名稱與定義）：

---

程式碼 1: 進入障礙相關知識圖譜 JSON

```
1  [  
2    {  
3      "Name": "Entry Barriers",  
4      "Definition": "Factors that deter or prevent new competitors from  
        entering a market.",  
5      "Synonyms": [  
6        "Barriers", "Barrier", "Barrier to Entry", "Entry Costs", "  
          Entry Restrictions"  
7      ]  
8    },  
9    {
```

```

10  "Name": "Natural Barriers",
11  "Definition": "Barriers due to market characteristics or
    geographical constraints.",
12  "Synonyms": [
13      "Geographic Barriers", "Environmental Constraints"
14  ]
15  },
16  {
17  "Name": "Artificial Barriers",
18  "Definition": "Barriers created by firms or legal structures to
    restrict entry.",
19  "Synonyms": [
20      "Regulatory Barriers", "Institutional Barriers"
21  ]
22  },
23  {
24  "Name": "Strategic Barriers",
25  "Definition": "Barriers created intentionally by existing firms
    to deter competition.",
26  "Synonyms": [
27      "Behavioral Barriers", "Strategic Entry Deterrents"
28  ]
29  },
30  {

```

```

31   "Name": "Economies of Scale",
32   "Definition": "Cost advantages due to increased production,
      making it harder for new entrants to compete."
33 },
34 {
35   "Name": "High Initial Capital Requirements",
36   "Definition": "Significant upfront investments required to enter
      a market."
37 },
38 {
39   "Name": "Geographic Constraints",
40   "Definition": "Barriers arising from location-specific factors
      such as access to resources."
41 },
42 {
43   "Name": "Patent Protection",
44   "Definition": "Legal barriers that prevent competitors from using
      patented technologies or processes."
45 },
46 {
47   "Name": "Brand Loyalty",
48   "Definition": "Consumer preference for existing brands, reducing
      new entrants' market share."
49 },

```

```

50  {
51      "Name": "Government Regulations",
52      "Definition": "Legal or administrative restrictions that limit
                    market entry."
53  },
54  {
55      "Name": "Predatory Pricing",
56      "Definition": "Deliberate underpricing to drive out competitors
                    or prevent new entry."
57  },
58  {
59      "Name": "Network Effects",
60      "Definition": "The value of a product increases as more people
                    use it, deterring new competitors."
61  },
62  {
63      "Name": "Vertical Integration",
64      "Definition": "Control over supply chain that limits competitors'
                    access to resources or distribution."
65  }
66 ]

```

---

## A.2 範例提示語 (Prompt 範例)

根據指示回答是否符合進入障礙，符合的是哪一個段落，符合哪個主題？

本範例分析來源為上傳之判例檔案 e2\_78.pdf 中的段落。

## A.3 模型回答範例 (Sample Answer)

以下為語言模型在輸入某一法律段落與知識圖譜後，所回傳的實際分析回應範例：

**是否與進入障礙相關？**是，此段落涉及到因為鐵路合併而可能產生的市場進入障礙。

**若與進入障礙相關，此段落符合哪一個進入障礙主題及定義？**此段落符合「Artificial Barriers」的主題。這是因為鐵路公司之間的合併可能創造出法律或商業結構上的障礙，限制其他競爭者進入市場。

**若與進入障礙無關，此段落屬於哪一個反托拉斯主題及概念？**不適用。  
因為該段落確實涉及到進入障礙的問題。

註：如需更完整分析，可提供具體段落內容進一步判讀。