反托拉斯法案文本分析

指導教授：佘建源

指導學生：唐嘉宏

目錄
Contents

01 研究背景
Subject Background

02 研究問題
Research Question

03 研究方法
Results Method

04 結論
Conclusion

01

研究背景

Research Background

研究背景

反壟斷

大數據和機器學習能夠協助反壟斷機構迅速分析市場數據，來發現潛在的違規行為，並且還可以幫助企業更好地符合反壟斷法的規範，但預測模型無法直接確認某行為是否違反反壟斷法，還需要結合經濟分析的框架來推論實際影響

進入障礙

進入障礙（Entry Barriers）隨著時代變遷，進入障礙的形態也發生改變，包括技術壁壘、網絡效應、資金限制等新的進入障礙形式，使得法律專業人士面臨挑戰，必須深入理解並準確定位案例中與進入障礙相關的段落和概念。

[1]Siying Cao "ECONOMIC ANALYSIS IN ANTITRUST LAW: AN AUTOMATED APPROACH APPLIED TO US APPELLATE COURTS"

資料概述

法院類別

共13個類別

法院類型概述

盧憶老師提供資料和
李逸華同學協助清理

LexisNexis

上訴法院

負責審理對下級法院判決的上訴案件

最高法院

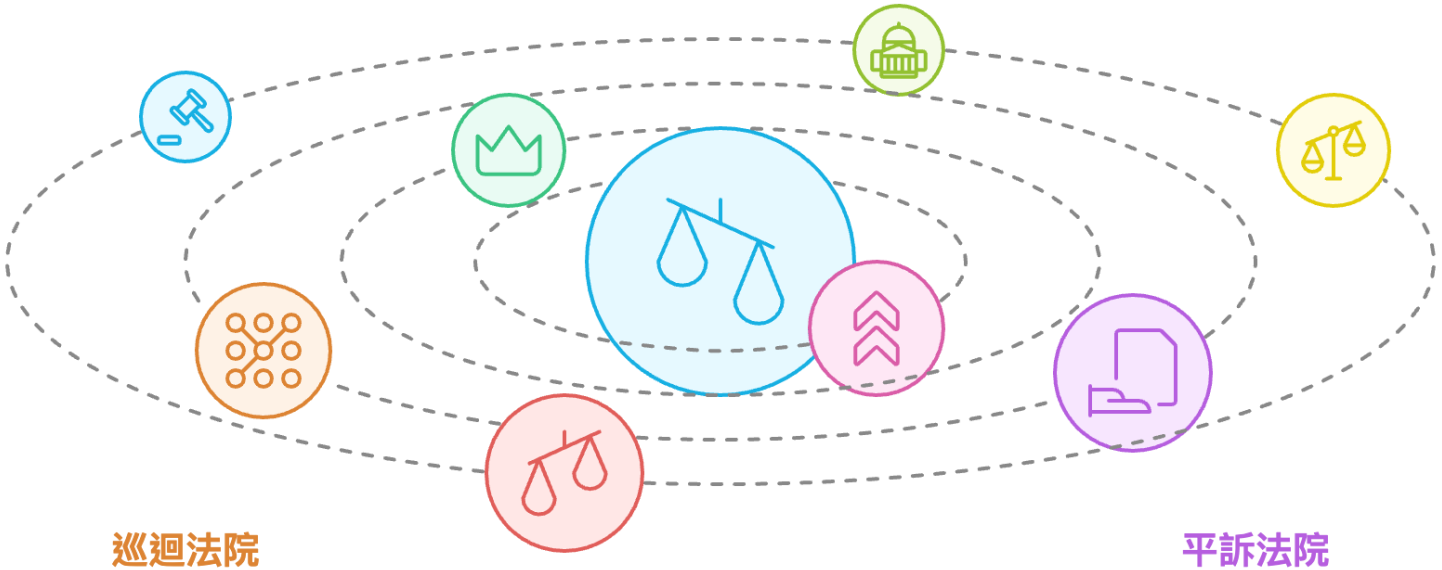
國家最高級別的上訴法院,審理關鍵法律和憲法問題

高等法院

負責刑事、民事、家事等案件的初審和上訴

衡平法院

專門處理信託、遺囑等衡平法案件



巡迴法院

負責中級上訴案件

民事上訴法院

專門處理涉及民事案件的上訴事宜

上訴法院

審理對下級法院的判決提出上訴的案件

平訴法院

主要處理民事和刑事案件的初審法院

資料概述

LexisNexis

	文件	文件標籤	年份範圍	判例數量（單一PDF）
Data 1	a	a1-a10	1891-01-01 ~ 1978-01-30	１００個判例
	b	b1-b10	1978-02-13 ~ 1984-12-14	b2有93個判例 ， 其餘皆100個判例
	c	c1-c10	1984-12-17 ~ 1991-10-01	c2有93個判例， 其餘皆100個判例
	d	d1-d19	1891-01-01 ~ 2022-12-31	1 個判例
Data 2	e	e1-e10	1998-09-14 ~ 2004-10-26	１００個判例
	f	f1-f10	1991-10-15 ~ 1998-09-10	f2有99個判例， 其餘100個判例
	g	g1-g10	2004-10-26 ~ 2009-04-23	g2有99個判例， 其餘100個判例
Data 3	h	h1-h10	2009-04-23 ~ 2013-07-19	h2有98個判例， 其餘100個判例
	i	i1-i10	2017-07-07 ~ 2021-08-30	i2有99個判例， 其餘100個判例
	j	j1-j10	2013-07-22 ~ 2017-07-05	１００個判例
	k	k1-k4	2021-08-31 ~ 2022-12-30	k1五個判例， 其餘１００個判例
	a-k	a1-k4	1891-01-01 – 2022-12-30	共９３０５個判例

02

研究問題

Research Content

研究問題

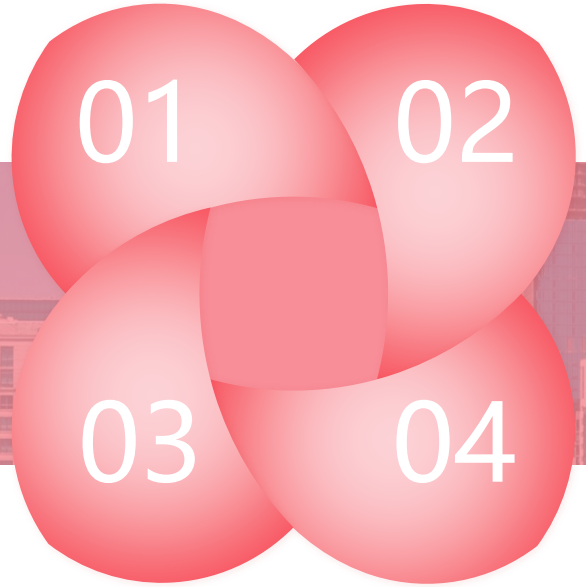
判例中是否有提到進入障礙的概念?
在哪一個段落提到的?

文本數據龐大

人工檢索與分析不僅耗時且容易出現遺漏或是錯誤

大型語言模型

大型語言模型（LLM）在理解法律語言，可能存在著語意偏差和識別困難，需要更精確地定義反托拉斯的法律概念。



NLP

自然語言處理（NLP）技術的應用為法律文本分析提供了新的可能

LLM

利用微調後的大型語言模型（LLM），來準確識別法律判例中與進入障礙相關的段落，並能夠理解其背後的法律概念。

人工標註耗時

總共9305個判例, 若採人工標註將會耗費大量人力和時間

知識圖譜的盲點

法律和經濟領域對進入障礙和反壟斷的定義不一樣，不能單單依靠其一知識，需要整理正確的知識圖譜建立網絡。

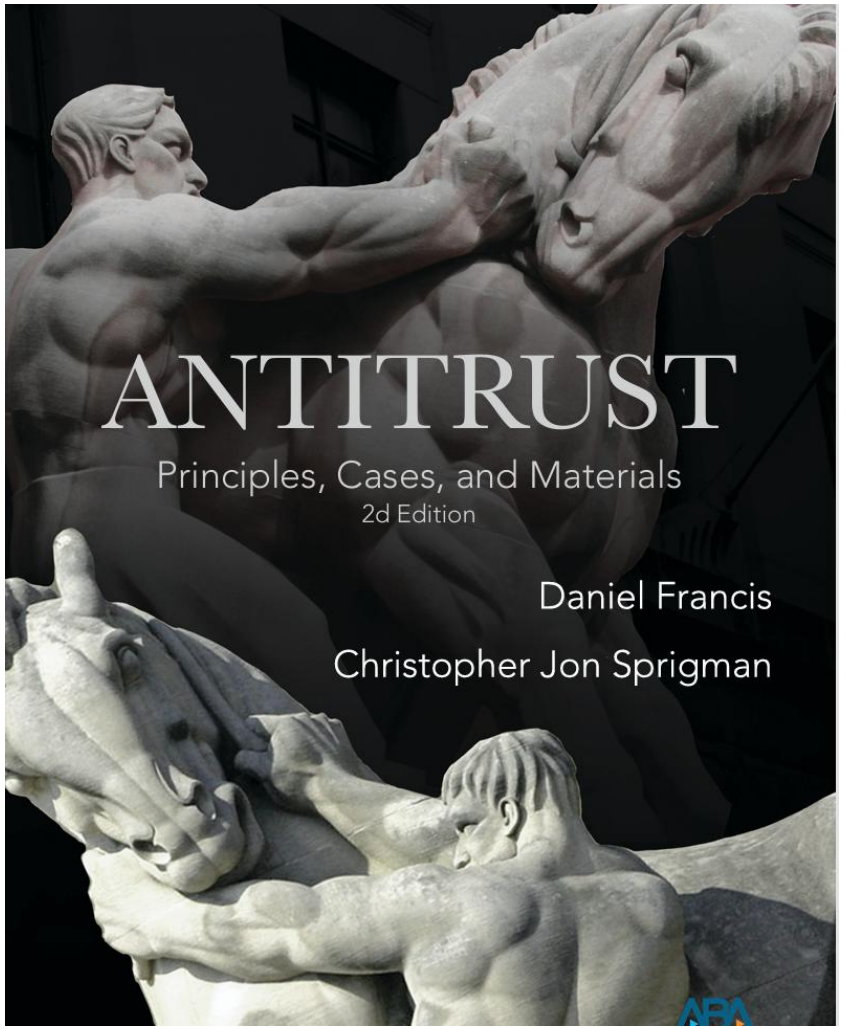
研究方向

建立LLM模型

對預訓練模型（Pre-trained Model）加上檢索增強生成（RAG）加強檢索知識圖譜，進行提示工程（Prompt Engineering），讓語言模型知道我們的任務和修改回答內容，最後進行微調（Fine-tuning）讓模型輸出更符合預期。

研究目標

能讓大型語言模型識別某段法律段落是否屬於反托拉斯法案或進入障礙的範疇，並能夠精確定位該段落是屬於該篇判例的哪個段落。並通過多個大型語言模型進行比較效能（GPT-4o、GPT-3.5 、 Legal-BERT…）



[2] Ruiwen Zhou et.al. 2024. TRAD: Enhancing LLM Agents with Step-Wise Thought Retrieval and Aligned Decision.
[3] Yu Zhao et.al. 2024. Contrast then Memorize: Semantic Neighbor Retrieval-Enhanced Inductive Multimodal Knowledge Graph Completion.
[4] Jiawei Chen et.al 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation
[5] Ilias Chalkidis et.al 2020. LEGAL-BERT: The Muppets straight out of Law School

03

研究方法

Research Method

資料清理

- 定義**正則表達式**(Regular Expression)
- 以“**Core Terms**”开始的文本段落，直到遇到**大寫字母開頭**的句子，或者到**“Opinion by:”**為止。

```
r'Core Terms\s+([\s\S]+?)(?=\n(?:[A-Z][^\s]*(?:\s[A-Z][^\s]*)*[:\n]|Opinion by:))'
```

CoreTerms

正則表達式

```
# 匹配分類標記及編號
category_patterns = [
    re.compile(r"LexisNexis® Headnotes"),
    re.compile(r"(HN\d+)"),  # 抓取HN後面的數字
    re.compile(r"Headnotes"),
    # 抓取LEdHN後面的數字
    re.compile(r"LEdHN\[(\d+)\]"),
]
```

PDF	Category	Content
a1	LexisNexis® Headnotes	Antitrust & Trade Law > ...
a2	HN1	Antitrust & Trade Law > ...
. . .	. . .	. . .
k4	HN13	Antitrust & Trade Law > Sherman Act > Scope > Monopolization Offenses

Headnotes

- 帶有**" * "和" + "以及" 數字標號 "**開頭的

```
6.0： "*" , "+" , "footnotes開頭的數字標號"
9.0：footnotes內容
```

- 抓取目標**黑線以下文本塊**

```
x_range = (50, 563)  長度範圍
color = (0.0, 0.0, 0.0)  顏色為黑色
width_range = (0.72, 0.73) 寬度範圍
```

Footnotes

資料清理

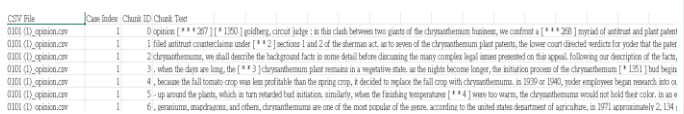
正則表達式

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

- 檢察官、法官、寫意見的人
- Judges和Counsel混淆，**將正則表達式界設定得更嚴謹**一點，確保遇到下一個標記"Opinion"時停止

Opinion by

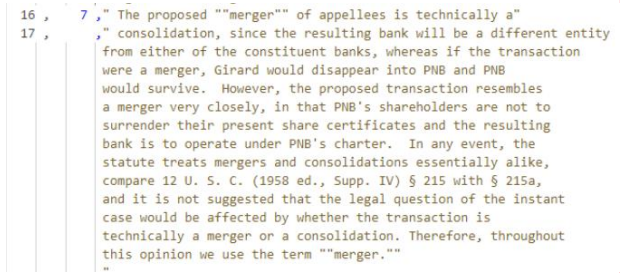
Opinion



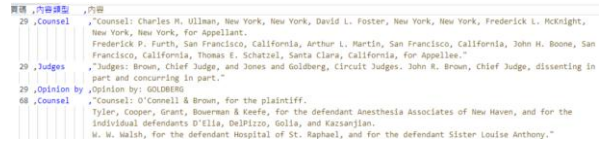
- **排除其他抓取方法**，像是Footnotes, Opinion by
- 抓取以”**Opinion**”開頭的標題和”**End of document**”結尾的文本

Opinion

Footnotes

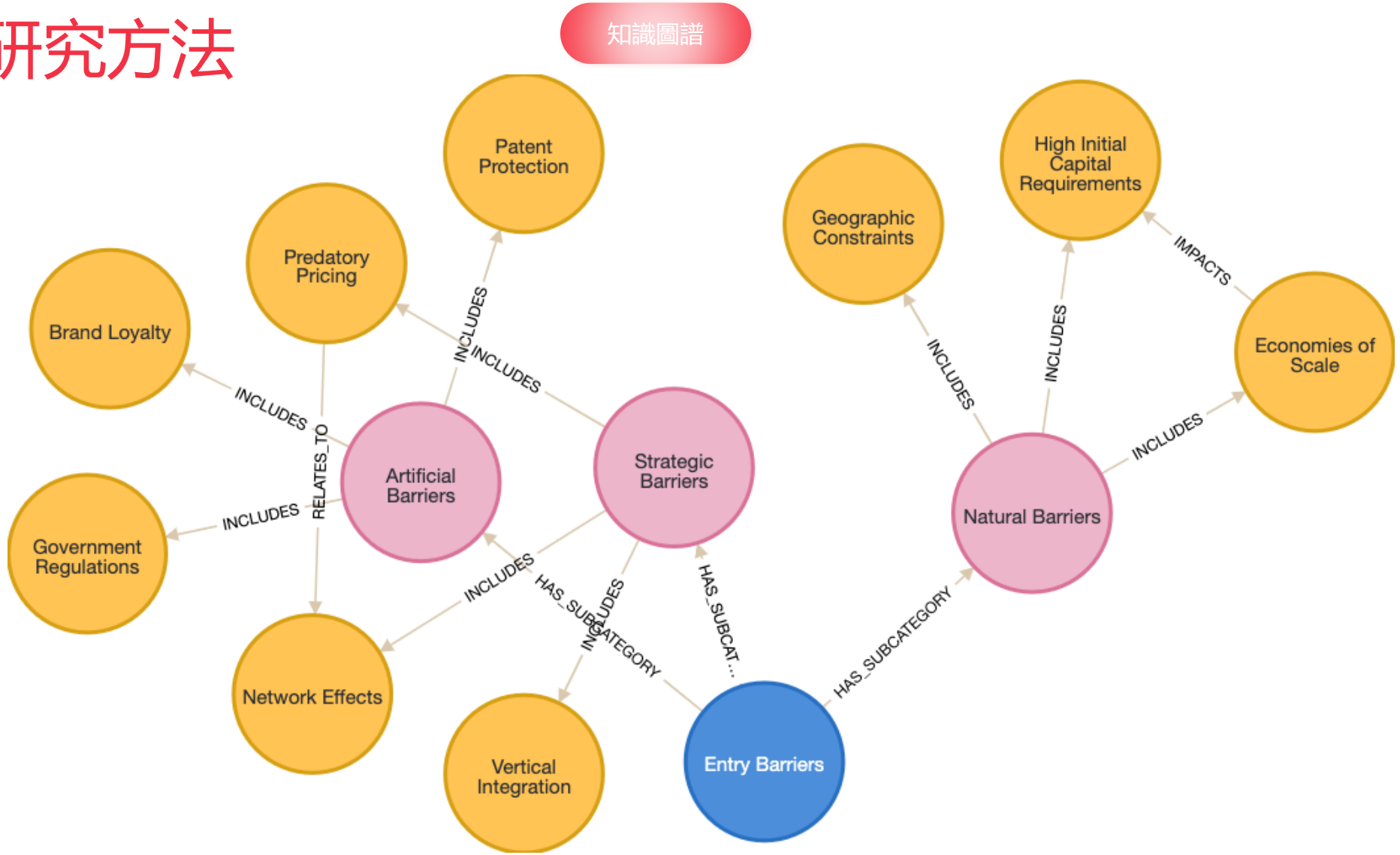


Opinion by



範例文本

研究方法


知識圖譜
Patent Protection
Predatory Pricing
Brand Loyalty
Government Regulations
Artificial Barriers
Strategic Barriers
Network Effects
Vertical Integration
Entry Barriers
Natural Barriers
Geographic Constraints
High Initial Capital Requirements
Economies of Scale
INCLUDES
RELATES_TO
HAS_SUBCATEGORY
IMPACTS

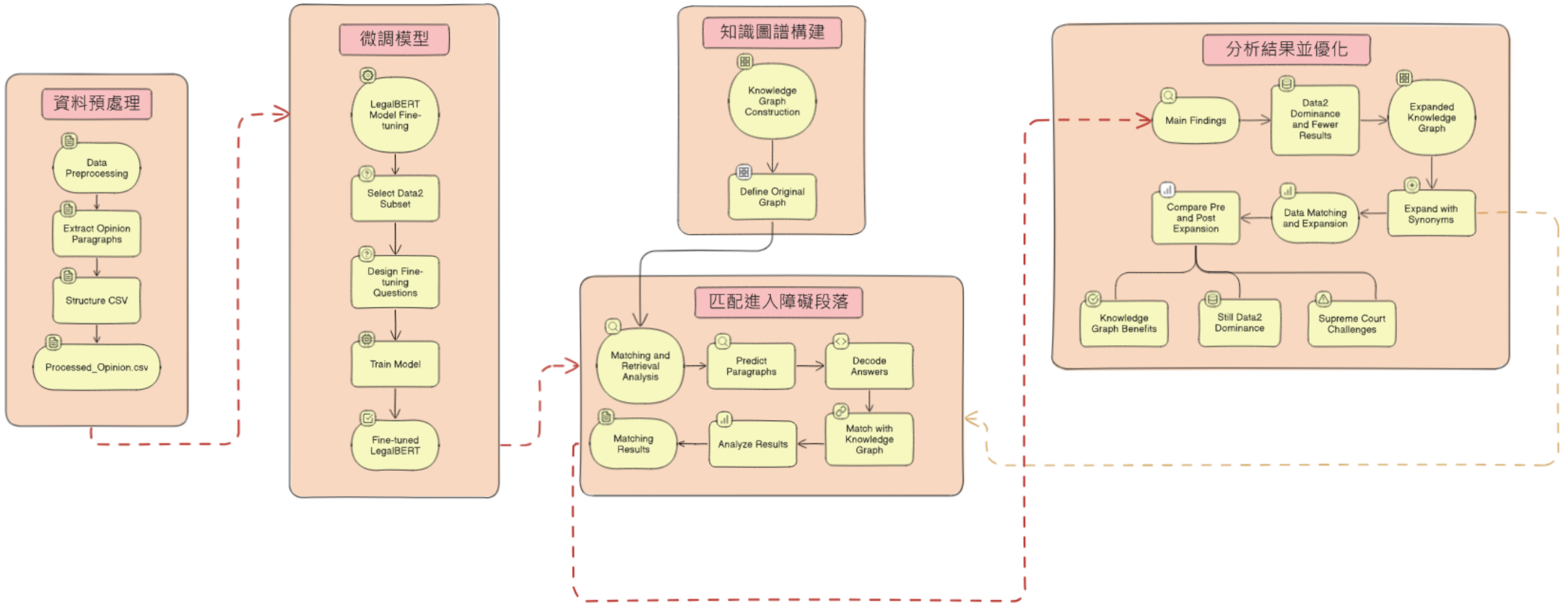
研究方法

段落裁切

算力有限
部分資料

關鍵詞定義
影響重大

微調使用的資料佔
檢索結果比例大



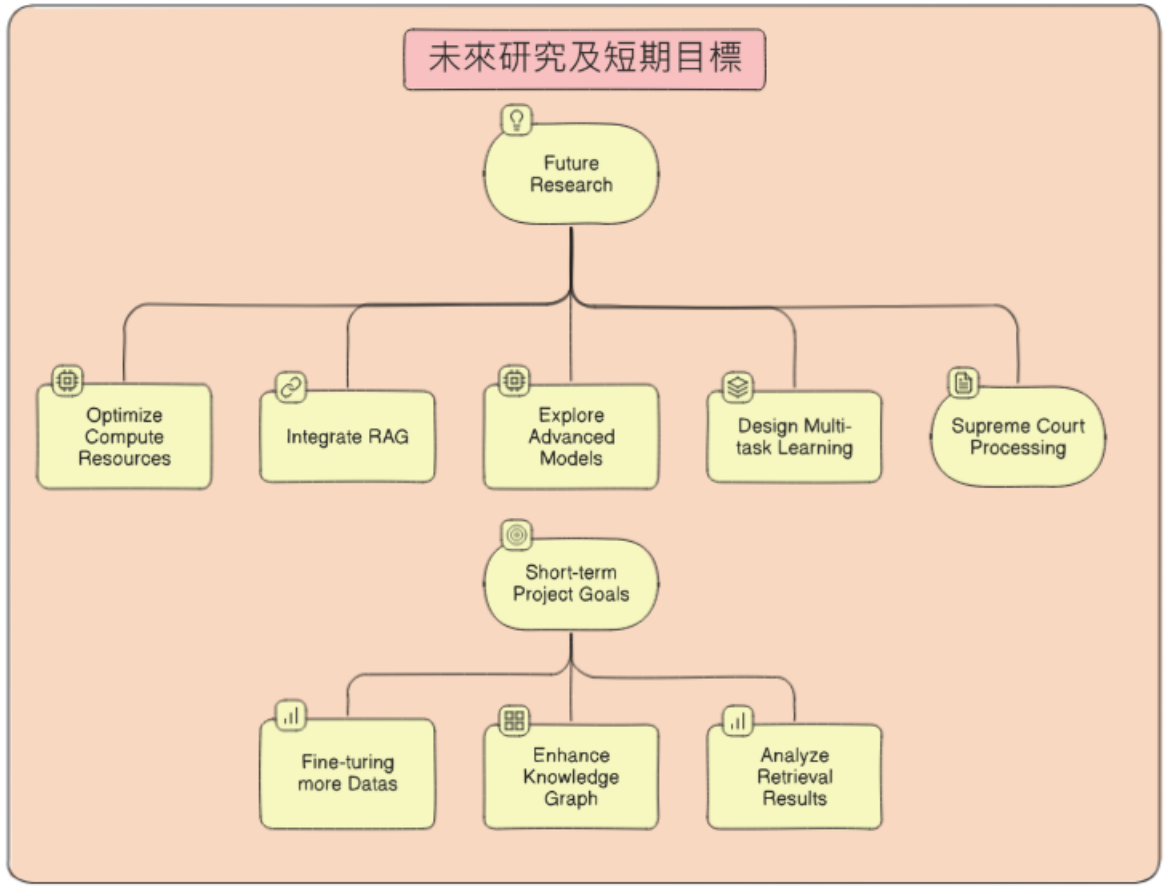
研究方法

未來研究

- 對於優化算力和使用更大模型都是為了解決**微調模型使用的資料有限**
- 可以使用其他文本資料去預測進入障礙相關文本，像是**Footnotes、Headnotes 、Core Terms**
- 針對**Supreme Court**的資料需要更深入探討分析

短期目標

- 先從微調的資料著手、**解決模型預測文本傾向於微調使用的資料**
- 探索更多**與進入障礙相關的關鍵詞和定義**
- 深入分析檢索到進入障礙的判例與結果是否吻合



THANK YOU

報告人：唐嘉宏 JIAHONG, TANG