EasyChair Preprint

№ 15243

To What Extent Do LLMs Understand a Verdict? A Case Study on Traffic Accident Information Extraction

Huai-Hsuan Huang, Chia-Hui Chang, Jo-Chi Kung and Kuo-Chun Chien

October 18, 2024

大型語言模型對判決理解的探討：以交通事故資訊擷取爲例
To What Extent Do LLMs Understand A Verdict? A Case Study on Traffic Accident Information Extraction

黃懷萱 **Huai-Hsuai Huang** 張嘉惠 **Chia-Hui Chang**
龔若齊 Jo-Chi Kung 簡國峻 Kuo-Chun Chien
國立中央大學資訊工程學系
chrbezz0487@gmail.com, chia@csie.ncu.edu.tw
z1a2x3s4c5d6v7f8b9g@gmail.com, qk0614@gmail.com

摘要

本研究探討大型語言模型 (Large Language Model, LLM) 在台灣地區車禍判決書資訊擷取任務中的應用。我們比較了大型參數模型如 GPT 和 GEMINI，以及小參數模型如 LLAMA-8B，並設計了三種提示詞 (Basic、Advanced 和 One-Shot) 來評估各模型在不同情境下的表現。

研究結果顯示，不同提示詞對於不同模型的效能有顯著差異，這可能與模型處理長文本的能力相關。具體來說，GPT 在使用 One-Shot 提示詞時，由於提示詞包含較多上下文，在字串的表現顯著優於其他提示詞，達到 89.2% 的平均準確率。然而，對於 GEMINI 模型，長提示詞反而導致效能下降，特別是在處理較長文本時表現不佳，顯示該模型對於過長提示詞的處理能力有限。這表明提示詞設計與模型架構的匹配度對效能有重要影響。

微調結果表明，GPT 在字串和數值資料的擷取中，經微調後的效能顯著提升，特別是在「折舊方法」和「修車費用」欄位上分別達到 97.9% 和 95.3%。相較之下，已經微調過的 chinese-llama 雖然初始效能較好但微調後效能提升有限，顯示其對微調的響應較低；而 instruct-llama 這類原型模型在微調後對字串資料的準確性大幅提升，從 63.7% 提升至 79.8%。

總結來說，提示詞設計和微調策略是提升模型效能的關鍵因素，未來可通過更大規模的模型和更精細的微調技術來進一步優化 LLM 在特定領域的應用效能。

關鍵字：法律判決書、LLM 微調、資料標註、資訊擷取、交通事故

1 Introduction

傳統的自然語言理解任務（如命名實體辨識、關係擷取）需要大量人工標記資料進行訓練。隨著大型語言模型（如 ChatGPT）的發展，許多研究開始利用這些模型替代人工進行資訊標記 (Li et al., 2023; He et al., 2024)。雖然人工標記準確性高，但成本和時間消耗大，而大型語言模型能有效降低這些成本，特別是在專業知識需求不高的文本上。

儘管 ChatGPT 在美國律師考試中表現不俗 (LeCun and Socratic, 2022)，但在處理如交通事故判決書這類結構複雜且高度專業的法律文本時，仍面臨挑戰。每份判決書因案件不同而有所變化，需要高度的準確性和一致性，因此大型語言模型的效能仍需進一步驗證與改進。

本研究評估大參數大型語言模型 (Mega Large Language Models, MLLMs) 和小參數大型語言模型 (Small Large Language Models, SLLMs) 在處理複雜交通事故判決書的能力。我們選擇車禍判決書作爲資料集，因爲它們包含法律專業知識、事實描述、法理分析及判決結果，有助於檢驗模型在法律文書上的表現，並支持交通事故責任認定和法律推理。

我們針對判決書中與賠償金額相關的 18 個欄位（如修車費用、賠償金額總額等）進行擷取，測試多種語言模型的原生能力和不同提示詞的效能，包括閉源的 GPT-4o-mini、Gemini-1.5-flash，及開源的 Meta Llama-3-8B 等。針對 GPT 和 LLAMA 進行了微調，但僅限於提示詞調整，以測試其在特定任務上的效能提升。

研究結果顯示，不同提示詞對於模型的效能有顯著影響。具體來說，GPT 模型在使用 One-Shot 提示詞時，由於提示詞包含更多上下文，在字串類型任務中表現最佳，達到 89.2% 的平均準確率。然而，GEMINI 模型在處理較長提示詞時，效能反而下降，特別是在長文本處理上的表現不佳，顯示其在處理過長提示詞時能力有限。這表明提示詞的設計和模型架構之間的匹配度對效能有重要影響。

此外，微調結果顯示，GPT 模型經微調後，對字串與數值資料的擷取效能均有顯著提升，特別是在「折舊方法」和「修車費用」欄位上的準確率分別達到 97.9% 和 95.3%。相較之下，chinese-llama 雖然初始效能較好，但微調後的提升有限，顯示其對微調的響應較

低；而 instruct-llama 在字串資料的準確率從 63.7% 提升至 79.8%。

實驗結果顯示，即使成功微調，參數較少的模型 SLLMs(LLAMA-8B) 效能仍不如 MLLMs（GPT、GEMINI）。這強調在訓練資料受限或任務更複雜的情況下，MLLMs 的還是更加具有優勢。

本研究的貢獻如下：

1. 探討大型語言模型是否能從專業的交通事故判決書文本中擷取所需資訊，並評估其在專業領域任務中的應用，以及提高法官處理資訊擷取的效率和準確度。

2. 分析 GPT、LLAMA 等模型處理判決書文本的效能，實驗顯示微調後的 SLLMs 效能仍沒有比 MLLMs 更好。

3. 我們也提出數值與字串類別標記一致性的計算方法，以評估不同模型在資訊擷取任務中的效能。

4. 實驗提示詞的敘述程度，驗證了並非提示詞越詳盡越好，而是需要根據語言模型的能力進行適度的調整。

2 Related Work

過去已有使用 GPT-3 標註 NLP 數據集的實驗。據 Zhou et al. (2022)，GPT 作爲標註工具仍有改進空間，但以低成本提供有效標註，甚至在預算有限時達到人類標記效果。然而，語言模型處理結構化資料或需專業知識的資料仍不及專家。另據 He et al. (2024)，GPT-3.5 在多種 NLP 任務上的績效可匹敵甚至超越人類標註者。這些研究表明，大型語言模型潛力巨大，尤其在有限資源下提供高效標註。

爲了更有效的提升大型語言模型做標記的效能，Li et al. (2023) 提出 CoAnnotating 框架來減輕了人工標註者的工作量，在保持標註準確性的同時，減少了 40% 以上的工作量。該框架先由模型生成初步標註，再由人工修正，並通過投票機制提高標註多樣性和準確性。不過，根據 Shen et al. (2023) 的說明，語言模型在標註或評估上存在位置偏差，也就是評估時如果是多個選項，可能會因爲選項位置而影響評估效能。所以在作評估時需要透過多種評估方式消除偏差，或對選擇進行隨機排序，最後使用經過人工評估的數據集進行評估，可以減少這種影響。

上述研究大多以分類標記爲主，然而在某些任務中，大型語言模型不一定優於傳統模型。例如，在命名實體識別（NER）任務中，Wang et al. (2023) 提出了 GPT-NER 方法，將 NER 任務轉爲生成任務，但生成特性使其更容易出現錯誤與幻覺。該研究結論是監督式學習仍優於大型語言模型。

Extraction-JSON
{ ” 事故日期”: ””, ” 事發經過”: ””, ” 事故車出廠日期”: ””, ” 傷勢”: ””, ” 職業”: ””, ” 折舊方法”: ””, ” 被告肇責”: ””, ” 塗裝”: 0, ” 工資”: 0, ” 烤漆”: 0, ” 鈑金”: 0, ” 耐用年數”: 0, ” 修車費用”: 0, ” 賠償金額總額”: 0, ” 保險給付金額”: 0, ” 居家看護費用”: 0, ” 居家看護天數”: 0, ” 每日居家看護金額”: 0 }

Table 1: 擷取格式設定，包含欄位與預設值（字串類型的預設值爲空，數值爲 0）

在硬體限制環境下使用 8 Billion 的 LLAMA 模型需要借助 QLoRA(Dettmers et al., 2023) 來降低運行成本，使其在資源有限的環境中高效運行。

LLAMA 是 Meta 開發的開源大型語言模型，旨在提升自然語言處理任務的表現。根據Touvron et al. (2023)，LLAMA 通過多層次訓練，顯著提升了語言生成和理解的能力。憑藉廣泛上下文和更多參數，LLAMA 能更精確地捕捉語義，在實體識別、文本分類和機器翻譯等任務中表現出色。相比其他大型語言模型如 GPT，LLAMA 在模型精度和資源利用效率方面具有優勢。

總之，儘管大型語言模型在多種 NLP 任務中具有潛力，但在進行信息抽取和實體識別等任務時仍面臨挑戰。LLM 容易生成不相關或不正確的信息，特別是在實體識別任務中。此外，LLM 微調需要大量標註數據，這可能成爲某些領域的瓶頸。最後，LLM 的訓練和推理需要大量計算資源，限制了其在實際應用中的普及。

3 Method

儘管研究已展示大型語言模型能減輕人工標註工作，實現人機協同，但依賴語言模型自動完成複雜任務的標註仍面臨挑戰。因此，本研究旨在評估現有大型語言模型在複雜任務上的表現。

我們選用車禍判決書作爲文本資料來源，這些判決書詳述了事故經過、時間和賠償金額等資訊。判決書內容豐富且包含許多計算性質的資訊，從中準確擷取數據是一項挑戰，尤其是與賠償金額相關的 18 個不同類別欄位，根據擷取的資料特性分爲數值類型與字串類型。在初始擷取時，字串類型設爲空字串，數值類型設爲 0，整體架構爲字典格式（如表1）。本研究旨在探索語言模型在此情境下的效能。

我們採用以下方法，並以正規表示式擷取結果爲基準，比較 GPT、GEMINI 和 LLAMA 模型的表現。

3.1 Different Prompt Types

爲增加實驗的多樣性與隨機性，我們設計了多種類型的提示詞來描述欄位擷取任務，如表2所示。提示詞分爲基本 (Basic)、進階 (Advanced)、範例 (One-Shot) 三個難度等級，差異在於任務規則的描述程度。目的是探討語言模型在不同規則限制與提示數量下的效能差異。以下提示詞的字數數量不包含判決書與擷取格式：基本提示詞爲 223，進階提示詞爲 553，範例提示詞爲 2772。設計方法是從基本提示詞出發，進階提示詞是針對不足的欄位進行補充，最終以範例提示詞作爲標準答案示範可能的擷取結果。

提示詞設計模擬眞實場景中任務描述的多樣性與複雜度。通過對比不同提示詞下模型的表現，我們評估語言模型的泛化能力，並找出影響效能的關鍵因素。提示詞輸入不同語言模型後，生成相應標記結果並進行統計和比對。

此機制避免了單一提示詞描述不當引起的偏差，充分利用語言模型的多樣性與隨機性，提高標記結果的質量與穩健性。

爲確保模型能準確擷取資訊，我們提供詳盡清晰的擷取條件，有時提供範例說明或基本格式要求，讓模型根據知識自行判斷擷取內容。無論方式如何，目標都是提供清晰指示，使模型高效且精準地完成資訊擷取，滿足眞實場景需求。

3.2 Fine-Tuning Models

微調語言模型的目標是提升其在特定領域與任務的專業性和準確度。雖然大型語言模型在許多任務上表現優異，但在處理實體辨識任務或需專業知識如法律領域等時，仍有不足。

爲補足這些不足，我們使用眞實法律判決書作爲輸入資料，並以人工審核的標準答案爲輸出，讓模型學習法律語言的細微差異、專業術語和推理邏輯，以提升其法律文本分析能力。

Meta Llama-3-8B(AI@Meta, 2024) 作爲預訓練模型，並利用 QLoRA 技術壓縮參數，在有限計算資源下進行高效微調。訓練資料由經過預處理的法律判決書文本組成，分段並標記了關鍵詞與句子。考慮到微調模型的 Token 數量限制，我們在訓練前會刪除超過限制的資料，僅保留可行範圍內的資料。

3.3 Similarity Caculation

爲了評估標記人員之間的一致性，以及模型標記結果的正確性，我們根據標記欄位的類型採取不同的評估標準。

對於數值型的資料欄位，如「賠償金額總額」、「修車費用」及「保險給付金額」等，我們會將這些欄位中的數值經過正規化處理，統一轉換爲數字形式，並去除單位等額外資訊。這些欄位中的數值來自不同判決書，會形成一個數值序列。

數值類型完全正確的比例：對於數值型的資料欄位，我們評估完全正確的比例，即標記結果與黃金答案完全一致的數量占比，A 代表標準答案，B 爲擷取資料，計算方式如公式 (1) 所示：

$$\text{Exact Match Ratio} = \frac{\sum_{i=1}^{n} \mathbb{I}(A_i = B_i)}{n} \tag{1}$$

其中，$\mathbb{I}(A_i = B_i)$ 是指示函數，當 $A_i = B_i$ 時取值爲 1，否則爲 0。這個比例衡量自動化標記結果與手動標記結果在數值型資料上的完全一致程度。

字串類型的餘弦相似度：對於文本型的資料欄位，例如「事發經過」、「傷勢」和「職業」等，我們採用計算餘弦相似度的方式來評估標記結果的一致性和準確性。這種方法能夠量化語言模型擷取的文本與人工標記的眞實值之間的相似程度如公式(2)。

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

4 Data Construction

4.1 Evaluation Dataset

本研究的評估資料集來自法務部裁判書公開網站，收錄了 2012 至 2022 年間超過 770 萬筆民事訴訟案例。我們透過關鍵字「駕駛」、「騎乘」和「車禍」篩選資料，最終選出 37884 筆與車禍相關的民事賠償案件，並隨機抽取 1000 筆作爲最終評估集，以確保樣本的代表性和可管理性。圖1顯示選定判決書的字數分佈，大多數集中在 2000 至 4000 字，反映車禍民事賠償案件的規模。當擷取超過字數時則會使用標點符號進行截斷，重複擷取。

我們整理了車禍案件判決書中與賠償金相關的資料，包括事故日期、事發經過、受害者的職業和傷勢等，共計 18 個參數，範例如表3所示。這些欄位的存在與否取決於事故具體情況，不是每起事故都會涉及所有賠償項目。完整的欄位範例如表4，並區分爲法官判決前與判決後的金額，這對語言模型的標記是一大挑戰。

Basic Prompt
[Content]={判決書內容} 依據 [Content] 的內容擷取相關資訊、填入 [Extraction-JSON] 要求如下: + 以賠償給原告的數據填寫 + 擷取折舊前的金額：工資、鈑金、塗裝、烤漆（除此之外擷取折舊後的金額） 返回結果爲一行 JSON 格式字串，無換行或特殊符號 [Extraction-JSON]={Extraction-JSON}
Advanced Prompt
[Content]={判決書內容} 依據 [Content] 的內容擷取相關資訊、填入 [Extraction-JSON] 要求如下: + 以賠償給原告的數據填寫 + 擷取折舊前的金額：工資、鈑金、塗裝、烤漆（除此之外擷取折舊後的金額） + 折舊方法爲定率遞減法或平均法 + 賠償金額總額填入判決書中已列出之總額，爲原告請求被告給付之金額需在... 起至清償日止 + 若無結果則留空白 + 被告肇責爲 0 至 100 + 傷勢爲原告在這場車禍中所造成的傷勢狀況 返回結果爲一行 JSON 格式字串，無換行或特殊符號 [Extraction-JSON]={Extraction-JSON}
One-Shot Model

[Content]={判決書內容}
[Rule]={
“事故日期”: “範例：原告主張被告於民國：108 年 10 月 29 日 18 時許，無照駕駛未注意車前狀況而碰撞訴外人... 擷取：108 年 10 月 29 日”,
“零件”: “範例：零件 92,090 元, 擷取：92090”
...
“被告肇責”: 範例：認本件事故被告應負 70% 之過失責任, 擷取：70”
}
遵從 [Rule] 的規則擷取 [Content] 的資訊、填入到 [Extraction-JSON] 中
返回結果爲一行 JSON 格式字串，無換行或特殊符號
[Extraction-JSON]={Extraction-JSON}

Table 2: 爲評估語言模型在不同任務描述條件下的表現，我們設計了多種提示詞範例。其中，[Content] 代表待處理的判決書內容，[Extraction-JSON] 代表擷取格式設定，涵蓋表4所列的所有欄位。基本提示詞僅簡述擷取任務，進階提示詞對擷取欄位進行大致說明，範例提示詞則詳細定義並描述了每個欄位 (共 18 個) 的要求。

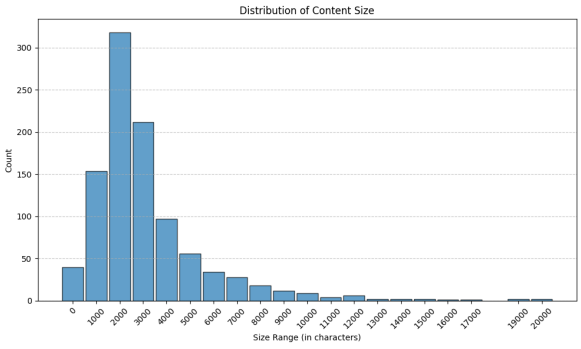

Figure 1: 判決書字數分佈

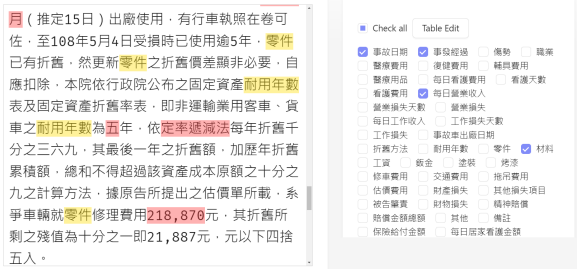

Figure 2: 判決書標記介面。左邊是相對應判決書資料，判決書中黃色螢光部分是與標記欄位的關鍵字相對應的，紅色螢光部分是標記者標記的資料，例如耐用年數爲五年，折舊方法爲定率遞減法，零件費用共 218870 元。右邊的 Label Checked 是標記者需要標記的欄位清單。

判決書範例
臺灣新北地方法院三重簡易庭小額民事判決 108 年度重小字第 941 號.... 原告起訴主張：緣被告於民國106 年 8 月 25 日上午 9 時 30 分 許駕駛車牌號碼 000-0000 號自小客車，在新北市蘆洲區仁愛街 93 巷 B2 停車場車道斜坡駛入 B1 停車場入口處時，因有轉彎車未讓直行車先行之過失，致擦撞原告所承保、由訴外人 OOO 駕駛之車牌號碼 0000-00 號自小客車（下稱系爭車輛），系爭車輛因而受損，經送修後，計支出修復費用新臺幣... 含鈑金 11,065 元、烤漆 12,060 元、零件 14,940 元），... 查系爭車輛係於99 年 5 月（推定 15 日）出廠使用，有行車執照附卷可稽...，依定率遞減法... 其折舊所剩之殘值爲十分之一即 1,494 元（元以下四捨五入）。此外，原告另支出鈑金 11,065 元、烤漆 12,060 元，無需折舊，是原告得請求被告賠償之修車費用，共計24,619元（... 。查本件事故之發生，被告固有過失。... 本院綜合雙方過失情節及相關事證，認原告及被告之過失程度各爲十分之三、十分之七，是被告應賠償原告之損害金額應減爲 17,233 元（計算式：24,619 元 ×7/10 = 17,233 元，元以下四捨五入），始爲適當。...

Table 3: 判決書範例 (藍色爲事故日期、紅色爲事發經過、橘色爲車損費用細項、紫色爲出廠日期、棕色爲折舊法、綠色爲最終車損賠償金額，粉紅色爲肇責比例)

欄位	欄位說明
事故日期	事實及理由一、原告主張被告於民國：108 年 10 月 29 日 18 時許，無照駕駛未注意車前狀況而碰撞訴外人...
事發經過	被告於民國 108 年 10 月 29 日 18 時許，無照駕駛車牌號碼 000-0000 號自用小客車，未注意車前狀況而碰撞訴外人普通重型機車...。
事故車出廠日期	其次，系爭汽車係96 年 5 月出廠...
傷勢	兩車因此發生碰撞，致原告人車倒地，受有胸壁鈍挫傷合併兩側肋骨骨折（右側：第六肋肋骨，左側：第四、五、六肋肋骨）、頭部外傷合併腦震盪以及左側頭皮、臉部撕裂傷，各約 1 公分、左側肩部、骨盆處以及四肢多處挫擦傷與瘀腫傷等傷害。....
職業	查原告主張系爭汽車修理須 3 日，致其另受有營業損失 4,500 元等語，已提出記載系爭汽車所須維修工作日爲 3 日之估價單，及高雄市政府專職計程車駕駛人每天營業總收入平均值爲 1,514 元... 其尚受有營業損失 4,500 元...
耐用年數	有關汽車耐用年數爲5年之期限...
折舊方法	依平均法計算其折舊結果（即以固定資產成本減除殘價後之餘額... 並參酌營利事業所得稅查核準則第 95 條第 6 項規定「固定資產提列折舊採用定率遞減法者，以 1 年爲計算單位其使用期間未滿 1 年者，...」
被告肇責	依兩造過失程度，原告應負擔與有過失責任百分之四十，被告應負擔過失責任百分之六十...
塗裝 工資 烤漆 鈑金	查系爭汽車修繕所須費用分別爲塗裝費用3,500 元、工資6,000元、烤漆費用5,500元一節，鈑金費用4300元...
修車費用	原告得請求修復系爭汽車所須之必要費用應爲12,083元...
賠償金額總額	原告依據侵權行爲損害賠償及保險代位權之法律關係，請求被告給付63,734元，及自起訴狀繕本送達被告之翌日（即 108 年 9 月 20 日，參見本院卷第 109 頁送達回證）起至清償...
保險給付金額	原告因系爭車禍得請求被告賠償之金額，自應扣除其已受領之2 萬 4,703元，於扣除後，原告尚得請求被告給付 5 萬 2,600 元【計算式：77,303 − 24,703 = 52,600（元）】。
每日居家看護費用 居家看護天數 居家看護費用	是以，依兩造不爭執之全日看護費爲...（計算式：1,200元 ×30=36,000）

Table 4: 賠償金相關欄位定義及擷取範例。橘色部分爲擷取的重點，而欄位中黑色的部分則是原告原始的金額，藍色是最終判決金額，這些最終判決金額是經過法官判決後的結果，總共欄位包含 18 個。

4.2 人工標記答案

爲了評估不同技術方法在資料標記上的準確度，我們從原始資料集中隨機選取 1000 筆資料進行標記，包含表4中的 18 個欄位。兩位標記人員進行標記，並計算技術方法與人工標記的一致性分數，作爲方法優劣的依據。相似度分數越高，表示標記準確度越高。

初次標記一致性爲 0.68，顯示該任務的複雜度高，人類標記者難以在初期達到良好一致性。我們進一步討論和審視分歧，最終達到相似度分數 0.93。

在標記過程中，我們發現某些欄位的複雜性較高，如表4中的「每日居家看護」、「居家看護天數」與「居家看護費用」，通常透過公式表示，因此未明確表示欄位與數值的關聯性，需標記者依常識判斷。另外，「折舊方法」欄位提及平均法，但文章內會同時說明參考定率遞減法，最終公式則用殘值計算，易誤導標記者，導致兩位標記人員理解差異過大，進一步加大了分歧。我們保留這些不一致的情況作爲分析參考。

我們探討了一致性不佳的原因，包括欄位定義不明、資料的模糊性以及標記人員理解的差異。對於這些理解上的衝突，我們在達成共識後進行了重新標記。這一過程突顯了在資料標記中保持一致性和準確性的重要性，並爲改進標記策略提供了參考。在標記過程中，我們也發現金額標記存在挑戰，因爲金額可能分爲原告索取的金額與法官實際判決的金額。這一差異對標記人員來說容易造成混淆，進而增加標記的難度，我們認爲這是金額標記上的一大挑戰。

標記介面如圖2所示，我們選取了與車禍相關的法院判決書進行資訊擷取測試。透過人工標記的方式將車禍相關的欄位都擷取出來，我們稱此資料集爲 TAVCD（Traffic Accident Verdict Compensation Dataset），其中標記總數爲 1000 筆，不僅包括擷取的結果，還保存了擷取結果在原文中的位置及其前後文脈絡等關鍵資訊，以保留訓練中或未來可能面臨的問題。

4.3 微調訓練資料的準備

在微調模型時，我們使用 TAVCD 作爲訓練資料，並將其分爲 80% 的訓練資料和 20% 的驗證資料，如表5所示。

分析標記結果時，我們發現相同的擷取信息可能在不同位置重複出現，這反映了文本的冗餘性及其分佈模式可能對模型性能有影響。

這種重複現象對提升模型的理解和泛化能力可能有重要意義，如幫助模型更好識別並權衡

資料類型	數量	佔比
訓練資料	800	80%
驗證資料	200	20%
總計	**1000**	**100%**

Table 5: 資料集的數量分布表

相同信息在不同上下文中的重要性，從而提高回答的準確性。我們希望通過這樣的訓練，使語言模型的標記更接近人類理解，提升準確性和實用性。具體的輸入資料格式如下：

```
<s>
    [INS] {提示詞} [\INS]
    [CONTENT] {判決書} [\CONTENT]
    {標準答案}
</s>
```

5 Results

我們選擇正規表示法作爲資訊擷取的基準方法。然而，由於判決書是人類撰寫，關鍵字和前後文的表述方式經常變化，影響了擷取的準確性。例如，「職業」欄位因缺乏固定模式，如「職業」欄位可能出現「該人專職 XXX」或「此人爲 XXX 職位」等多種描述形式，增加了擷取難度導致正規表示法表現不佳；只會是「定率遞減法」或「平均法」，易被正規表示法識別。這些觀察顯示，前後文的穩定性在一定程度上影響了正規表示法的效能。

5.1 不同語言模型的效能

在實驗中，我們測試了不同語言模型在隨機度從 0 到 1 之間（以 0.1 爲間隔）的效能。結果顯示，`GPT-4o-mini` 在隨機度爲 0 時，`Gemini-1.5-flash` 在隨機度爲 0.5 時，產生了最佳的輸出結果。

根據這些最佳表現進行比較，如表6所示，分析了 GPT 和 GEMINI 兩種語言模型在資訊擷取任務上的表現。

在字串類型的資訊擷取中，GPT 模型在 One-Shot 提示詞的情況下表現最佳，特別是在「事故車出廠日期」和「折舊方法」等欄位中達到最高效能。相較之下，GEMINI 模型在字串類型任務中的整體效能較低，但在某些較爲簡單的欄位，如「被告肇責」與「事故車出廠日期」，其表現接近 GPT 模型。

分析提示詞效能差異時，我們發現 One-Shot 提示詞的字節數量（2772 字節）明顯多於 Advanced 提示詞（584 字節）。雖然更多資訊能幫助模型，但也可能增加閱讀負擔。對 GPT 而言，One-Shot 提示詞有助於提高模型在處理不確定文本時的準確性和穩定性；但對 GEMINI 來說，長文本資訊可能成爲負

擔，導致效能下降，兩者在 One-Shot 提示詞上的表現各爲 89% 與 70%。

隨著語言模型對提示詞文本處理能力的提升，One-Shot 提示詞可能成爲最佳選擇。其優勢在於包含完整範例，提供更多上下文和答案參考，使模型在處理複雜字串時更清晰；而 Advanced 提示詞雖說明任務規則，但缺乏具體範例，導致在需要高語境理解的欄位上效能不如 One-Shot 提示詞。

在字串與數值資訊擷取的比較中，數值資料可能需要更彈性的思考空間，如金額的計算與金額表示的說明等，在部分欄位上 Advanced 提示詞不亞於 One-Shot 提示詞，尤其是在 GEMINI 上可能體現出來，可能因 Advanced 提示詞增強了對數值模式的理解亦或是更多的彈性空間思考。相反地，字串類型資料具有較高語境變異和語意複雜度，如事件描述或職業資訊，需要更強的語言理解和上下文解析能力。GPT 模型在 One-Shot 提示詞下表現突出，能有效利用詳細範例和豐富上下文解析複雜字串資訊，而 GEMINI 模型在這方面相對較弱。

總體而言，字串類資料在 GPT 模型使用 One-Shot 提示詞時效能最佳，而數值類資料則在進階提示詞下稍優。這顯示不同提示詞設定會顯著影響語言模型效能，尤其在數值資訊擷取中。這些結果顯示提示詞選擇對語言模型效能至關重要。提示詞的複雜度（Basic, Advanced, One-Shot）可能改變模型對上下文的理解深度，影響輸出質量。未來研究可進一步探討提示詞優化對不同語言模型效能的影響。

5.2 微調模型

針對閉源語言模型 `GPT-4o-mini`、開源模型 `meta-llama/Meta-Llama-3-8B-Instruct`[1]及 `shenzhi-wang/Llama3-8B-Chinese-Chat`[2](中文微調模型) 進行微調，探討在有限硬體資源下，透過微調是否能優化這些模型在車禍判決書資訊擷取任務中的表現，並滿足複雜任務需求。透過微調，希望這些模型能更準確地識別文本與相關欄位的關聯，以提升資訊擷取效能。根據表6的實驗結果，微調模型的資料均使用 `Advanced` 提示詞。在微調實驗中，我們測試了多種參數組合，最終發現最佳配置爲：訓練步數 (steps)600，學習率 (learning rate)0.02，丟棄率 (dropout rate)0.2。若訓練步數過多或丟棄率過低，可能導致模型過度擬合 (overfitting) 並重複輸出相同文字，這可能是因爲模型參數量不足，無法完全理解複雜任務，只能依據前後文推測。此參數組合顯著提升了模型的準確性 (accuracy) 和穩定性 (stability)。結果如表7所示：

在字串類型資訊擷取任務中，`chinese-llama` 未經微調即有較高效能，可能因已針對中文語系初步微調，微調後效能從 0.652 僅提升至 0.748，改進空間有限。相比之下，`instruct-llama` 微調前效能較低，平均僅爲 0.637，但微調後顯著提升至 0.798，顯示對微調高度響應性。這表明 `instruct-llama` 原始模型適應性較弱，但微調後改善了理解和擷取能力。

結果強調了模型初始狀態和訓練語料的重要性：針對特定語系或任務初步調整的模型微調效益較小；而泛用性高但初始表現不佳的模型，微調可帶來顯著效能提升，這對未來選擇微調策略具有指導意義。

在數值類型資訊擷取中，`chinese-llama` 微調後數值擷取提升有限，可能是其初始訓練已具備較強數值識別能力；而 `instruct-llama` 對數值類型提升幅度較大，顯示微調增強了其適應性。整體上，`GPT` 微調後效能顯著提升，表現優異，顯示其對微調高度敏感，能快速吸收新資訊以提升資訊擷取能力。而在 `GPT` 微調後效能最佳，特別在「塗裝」、「工資」和「修車費用」等項目表現突出。相比之下，`chinese-llama` 和 `instruct-llama` 微調後雖有改善，但總體提升不如 `GPT`，可能因 `GPT` 架構更靈活精確或是受模型所擁有的參數量影響。`GPT` 的優異表現可能源於其靈活架構和大量參數，使其在微調中快速適應並改善效能，並充分利用訓練數據中的上下文和範例來提高處理複雜語境的準確度。

微調所使用的訓練資料爲 800 筆判決書，模型爲 `gpt-4o-mini-2024-07-18`，訓練參數如下：epoch 數爲 2，訓練的總 token 數爲 6,949,546，批次大小（Batch Size）爲 1，學習率倍率（LR multiplier）設定爲 1.8。整個訓練過程的花費約爲 15.57 美元。

總結而言，微調顯著提升語言模型效能，對 `GPT` 這種 MLLMs 更爲明顯，目前 SLLMs 的前景雖還無法與 MLLMs 相比，但未來研究應探討不同模型和微調方法，以提升 SLLMs 在特定應用中的實用性和效能，讓更多人可以應用。

[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct/tree/main

[2]https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat

	Method	RE	GPT			GEMINI		
	Prompt	-	Basic	Advanced	One-Shot	Basic	Advanced	One-Shot
字串	事故日期	0.810	**1.000**	**1.000**	0.994	0.993	**1.000**	**1.000**
	事發經過	0.608	0.684	0.739	0.789	**0.825**	0.823	0.760
	事故車出廠日期	0.456	0.897	0.869	0.935	**0.946**	0.938	0.935
	傷勢	**0.887**	0.810	0.820	0.803	0.333	0.603	0.285
	職業	0.166	0.891	**0.912**	0.876	0.378	0.632	0.332
	折舊方法	**0.920**	0.854	0.865	0.863	0.560	0.642	0.653
	被告肇責	0.900	0.979	**0.984**	**0.984**	0.953	**0.984**	**0.984**
	Average(字串)	0.678	0.874	0.884	**0.892**	0.713	0.803	0.707
數值	塗裝	**0.938**	0.736	0.731	0.839	0.793	0.782	0.767
	工資	0.528	**0.860**	0.663	0.855	0.788	0.850	0.819
	烤漆	0.803	0.870	0.865	0.933	0.938	**0.953**	0.834
	鈑金	0.912	0.720	0.902	0.953	0.953	0.907	**0.969**
	耐用年數	0.477	0.969	0.922	0.974	0.912	0.907	**0.979**
	修車費用	0.347	0.503	**0.834**	0.487	0.409	0.808	0.420
	賠償金額總額	0.181	0.756	0.731	0.756	0.715	0.689	**0.803**
	保險給付金額	**0.938**	0.653	0.544	0.601	0.679	0.601	0.788
	居家看護天數	0.959	0.964	0.959	0.959	0.959	**0.974**	**0.974**
	居家看護費用	0.948	0.948	0.959	0.948	0.943	**0.969**	0.948
	每日居家看護金額	0.959	**0.979**	0.974	0.964	0.959	**0.979**	**0.979**
	Average(數值)	0.726	0.814	0.826	0.843	0.823	**0.856**	0.844

Table 6: 不同的提示詞對應不同的語言模型效能，粗體代表該欄位的最佳解

	Method	chinese-llama		instruct-llama		GPT	
	Status	pre-trained	fine-tune	pre-trained	fine-tune	pre-trained	fine-tune
字串	事故日期	0.238	0.668	0.166	0.725	**1.000**	0.933
	事發經過	0.386	0.523	0.377	0.533	0.739	**0.897**
	事故車出廠日期	0.736	0.855	0.560	0.834	0.869	**0.964**
	傷勢	0.807	0.841	0.818	0.843	0.820	**0.917**
	職業	0.870	0.881	0.891	0.902	0.912	**0.938**
	折舊方法	0.544	0.487	0.663	0.772	0.865	**0.979**
	被告肇責	**0.984**	**0.984**	**0.984**	0.974	**0.984**	**0.984**
	Average(字串)	0.652	0.748	0.637	0.798	0.884	**0.945**
數值	塗裝	0.896	0.756	0.902	0.798	0.731	**0.995**
	工資	0.513	0.575	0.539	0.720	0.663	**0.953**
	烤漆	0.886	0.793	0.876	0.865	0.865	**0.995**
	鈑金	0.969	0.979	0.964	0.943	0.902	**0.995**
	耐用年數	0.777	0.922	0.725	0.943	0.922	**0.979**
	修車費用	0.497	0.663	0.472	0.642	0.834	**0.953**
	賠償金額總額	0.570	0.736	0.513	0.788	0.731	**0.922**
	保險給付金額	0.762	0.705	0.720	0.741	0.544	**0.990**
	居家看護天數	0.969	0.959	0.959	0.964	0.959	**0.974**
	居家看護費用	0.959	0.959	0.948	0.959	0.959	**0.979**
	每日居家看護金額	0.959	0.959	0.959	0.953	0.974	**0.979**
	Average(數值)	0.796	0.819	0.780	0.847	0.826	**0.974**

Table 7: LLAMA 與 GPT 模型微調效果比較。包含 chinese-llama 與 instruct-llama。底線部分標示的是以 LLAMA 模型爲基準，進行微調與非微調項目的比較，粗體則表示所有項目中表現最優的結果。提示詞皆使用表2中的 Advanced 提示詞。

6 Conclusion

本研究探討了大型語言模型（LLM）在處理台灣地區車禍判決書中的資訊擷取應用。這一任務的挑戰在於文本中自然語言與數值數據的交互使用，且資料多爲非結構化。我們針對如事故日期、賠償金額等 18 種欄位進行模型效能分析，並設計了多種提示詞來評估提示詞設計和微調對效能的影響。

實驗結果顯示，提示詞設計對模型效能的影響顯著，不過並非提示詞越詳細越好。過度複雜的提示詞可能在長文本處理時降低模型的效能。例如，GPT 模型在使用 One-Shot 提示詞時，由於提示詞包含更多上下文資訊，能有效地擷取複雜字串類資料，而 GEMINI 模型則因長提示詞而表現不佳，提示詞的設計應根據模型特性進行調整。

對於語境較穩定的欄位，如「折舊方法」和「修車費用」，模型的擷取效能較高，尤其是在 GPT 模型的微調後，準確率顯著提高，分別達到 97.9% 和 95.3%。然而，對於語境不固定的欄位，如「職業」或「事發經過」，模型效能相對起來較弱，而這類欄位對語言模型的語意解析能力要求更高。

微調模型的結果進一步顯示，GPT 微調後在字串和數值資料擷取上的表現最佳。相比之下，chinese-llama 雖然初始效能較優，蛋微調後提升有限，這可能與其初始已針對中文語系進行過調整有關，而 instruct-llama 的微調響應性較高，特別是字串資料的準確率從 63.7% 提升至 79.8%，顯示其在微調狀態下有顯著的改進空間。

微調過程中發現，適當的微調能顯著提升模型在字串資料上的準確性，但對數值資料的改善則相對有限，可能是由於數值資料格式變化較大，模型難以全面掌握這些變異。結果也顯示，對已經微調的模型進行多次微調不一定會有效果，可能是因爲模型效能在某些任務中達到了飽和。

總結來說，本研究強調了提示詞設計和微調策略對模型效能的關鍵影響，並指出微調策略應根據資料特性和任務需求進行調整，以避免過度訓練或過度擬合的問題。即使在有限的計算資源下，大型參數語言模型（MLLMs）相較於小參數模型（SLLMs）仍具有顯著優勢，但要進一步提升效能，可能需要更大規模的模型和更精細的微調技術。

未來研究應進一步探索不同模型的微調方法，優化提示詞設計，特別針對法律文本等高語意理解任務，尋找更合適的模型架構與訓練策略，以提升 LLM 在特定領域中的應用效能。

7 Future Work

- 目前微調 8B 參數的開源語言模型在性能上仍然難以與現有的閉源大語言模型匹敵。未來可以探討透過投票機制、分層策略（如Zhang et al. (2024) 提出的多層架構）等方法，在硬體環境受限的情況下，進一步提升語言模型在特定領域的效能表現。

- 本研究中使用的標注數據量約爲 1000 筆，數據量相對有限。未來工作可以考慮擴充數據集，以提高模型的泛化性和穩定性，進而提升模型的整體效能。大規模和高質量的數據集將有助於模型更好地理解和擷取文本中的複雜信息。

- 隨著越來越多新的語言模型和架構的出現，未來的研究應持續跟進這些技術的發展。針對這些新出現的語言模型進行實驗和比較，以確定它們在特定應用中的優劣。這不僅有助於選擇最適合的模型，也能促進對不同模型架構和微調策略的理解。

8 Limitation

- 目前標準答案依賴人工標註，由兩位標記者獨立進行並比對結果。然而，兩位專家標註一致但不符合實際答案的情況可能發生，這可能源於人類在閱讀理解上下文時的疏忽或判斷失誤，因此人工標註結果可能與眞實效能產生偏差，需特別注意。

- 雖然進行了實驗來測試模型在不同隨機度下的效能，但仍使用了隨機度參數來提升模型生成結果的多樣性，這種方法可能導致輸出結果的穩定性下降，特別是在模型需要處理具有複雜結構或細微語意差異的資料時。

- 本研究主要聚焦於中文車禍判決書的資訊擷取，因此其結論可能不完全適用於其他語言或文本類型的標記任務，這限制了研究的普遍性和適用範圍。

- 此外，所使用的數據集規模相對較小（僅 1000 筆資料），可能影響模型學習的深度和泛化能力。未來的研究可以擴大數據集規模，以提供更加堅實的結果和見解。

References

AI@Meta. 2024. Llama 3 model card.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Yann LeCun and John Socratic. 2022. A path towards autonomous machine intelligence. arXiv preprint arXiv:2212.14402.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 1487–1505, Singapore. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4215–4233, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Raphaël Couprie, Edouard Grave, Guillaume Lample, and Alexis Conneau. 2023. Llama: Open and efficient foundation language models.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. arXiv preprint arXiv:2304.10428.

Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. 2024. Autocoderover: Autonomous program improvement.

Yilun Zhou, Chunting Zhang, Shuohang Wang, Zhengyuan Liu, and Diyi Yang. 2022. Is gpt-3 a good data annotator? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1051–1061. Association for Computational Linguistics.