Collision Care Guide based on Large Language Models
基于 LLM 的交通事故陳述記錄輔助助理

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

摘要

本研究提出了一個基於大型語言模型（LLM）的交通事故陳述記錄助理系統（CCG），旨在協助車禍當事人釐清事故經過，減輕警方與保險專員詢問的負擔。CCG 系統的核心設計包括三個主要模組：提問模組、資訊擷取模組和事故經過生成模組。提問模組利用預設的問題模板引導用户提供完整資訊；資訊擷取模組將用户回答準確擷取至預先定義的資訊儲存格式（TARF）中；事故經過生成模組則將結構化資料轉化爲連貫的事故敘述。這種模組化設計實現了高效、準確的事故資料收集和處理。

實驗結果顯示，CCG 系統在人工評估中 F1 分數達到 0.909，在 LLM 對話評估中準確性和完整性評分均達到 10 分中的 7 分以上。我們同時採用人工評估和 LLM 自動評估方法，驗證了系統在實際應用中的有效性。這些結果證明 CCG 系統具有良好的實用性和應用前景，能夠協助事故當事人精確記錄所提供的資訊，爲後續的法律程序和保險理賠提供可靠依據。

關鍵字：大型語言模型、對話系統、資訊擷取、交通事故

1 Introduction

根據台灣交通部路政及道安司的統計，臺灣交通事故數量逐年上升，從 2019 年發生 34 萬件事故，至 2023 年統計已突破 40 萬件，平均每天超過 1000 件事故發生[1]。當事人在發生交通事故後，通常會立即報案並聯繫保險公司，但警方與保險公司在處理時，首要任務是釐清事故經過，然而常見問題包括：當事人描述不完整或遺漏關鍵事實，導致責任歸屬難以確定；警力與保險公司人力有限，難以同時服務所有當事人，造成漫長等待；而大多數人在事故發生後往往不知如何應對，需仰賴有經驗者協助處理賠償與責任問題。在現今基於大型語言模型（LLM）的智能對話系統雖廣泛應用於教育與醫療領域 (Dam et al., 2024)，但針對交通事故領域的應用卻相對稀少。目前在交通事故領域的研究，多集中於事故預測與數據分析 (Behboudi et al., 2024)，針對當前事故對於事故當事人的記錄、蒐集信息等方面的研究也是較爲缺乏。

鑑於大型語言模型 (LLM) 在對話能力與資訊整合上的進步，我們選定由 OpenAI 所提出的 GPT 爲核心設計，因爲他在對話生成與訊息理解廣爲人知與表現出色。我們提出了一個基於 ChatGPT 的智能聊天系統 CollisionCare Guide (CCG)，設計了一個完整的問答流程，包含了三個不同的模組: 提問模組、資訊擷取模組和事故經過生成模組，並使用自定義的 JSON 表單模板，資訊儲存格式 (TARF)，以引導當事人逐步記錄成結構化的事故細節，最終也可以透過模組轉化爲連貫的敘述形式，以便警方或保險理賠員進行責任判斷與賠償計算。此系統研究目的是爲了減少多次重複問答的負擔，提升事故處理效率。爲驗證整體 CCG 的效能，我們進行了人工評估與 LLM 自動評估。在人工評估中，測試人員根據判決書模擬當事人進行描述回答，最終 F1 分數達到 0.9，顯示 CCG 能準確針對使用者回覆的資訊，擷取至正確的欄位中，值得注意的是，AI 代理人在相同的測試方法中也達到了 0.9 的 F1 分數，這表明 AI 代理人與眞人測試員有高度的相似度，這一結果證實了 AI 代理人可以有效地替代眞人進行更多對話上的模擬測試。在 LLM 自動評估中，AI 代理人與 CCG 進行互動對話，利用 LLM 自動評估結果顯示其完整性與精準性達至約 7/10 分，證明 CCG 能有效引導當事人並正確擷取所提供的資訊。

我們此論文的貢獻如下:

- 我們提出一個交通事故代理人 CCG，結合擷取與問答模組的交互合作，與可自定

[1] https://ba.npa.gov.tw/statis/webMain.aspx?k=defjsp

義的事件資訊模板，將記錄轉化成結構化的表達方式和交通事件的完整敘述，具有高度的應用彈性。

- 我們設計並實施了一套創新的評估方法，結合人工評估和 LLM 自動評估，全面驗證了 CCG 系統在實際應用中的有效性和準確性。

- 我們的研究設計一套完整問答流程，爲交通事故資訊蒐集和處理提供了一個範例，展示了 LLM 在特定領域應用的潛力，爲未來在其他法律案件中的應用奠定了基礎。

2 Related Work

專門處理車禍案件的市場規模相當龐大，尤其在交通事故頻繁的地區。根據上述官方資料，台灣每年發生數萬起交通事故，這些事故涉及的法律問題和賠償需求促使車禍律師的需求不斷增加。此外，隨著人們對法律權益的重視程度提高，越來越多的人選擇在發生車禍後尋求專業律師的協助，以確保自己的權益得到保障。

國際上專門處理車禍案件的律師事務所如 Alexander[2]等，通過結合地理定位和在線聊天機器人，提供律師匹配與法律諮詢的服務，如 Accident Consults[3]、FindLaw[4]和 Legal010[5]。這些平台目標在於提供專業見解並幫助受害者爭取賠償。除此之外，一些系統也運用了人工智能技術來提高法律服務的效率。舉例來說，LAW-U (Socatiyanurak et al., 2021) 採用了有限狀態機控制對話流程，爲性犯罪受害者提供精確法律建議，並能準確推薦相關最高法院判決。而 DoNotPay[6]作爲「第一位機器人律師」，讓使用者能夠處理法律糾紛如爭議罰單和隱私保護等問題。

針對交通事故分析的技術也逐漸發展，AccidentGPT (Wu et al., 2024) 爲一個多模態基礎模型，能處理多種輸入數據，包括音頻、影像、文本等，並自動生成多任務分析結果，從重建事故過程到生成責任歸屬報告皆涵蓋其中。此外，基於 BERT 的大型語言模型也被應用於交通事故的嚴重程度分類 (Grigorev et al., 2024)。這項研究使用了超過 75 萬份事故敘述，並達到 84.2% 的預測準確率，強調了 LLM 在處理大型事故數據集時的潛力，且降低了計算複雜性，提升了可擴展性。

隨著 LLM 技術的進步，研究發現這些模型可以替代人類進行自動化評估。根據 Chiang and yi Lee (2023) 的研究，LLM 在語言生成任務的評估結果與人類高度相關，顯示了其在自動化評估中的潛力。同樣地， Lin and Chen (2023) 提出了一種基於 LLM 的自動對話系統評估方法，該方法與人類評估結果顯示高度一致性，進一步證明了 LLM 在對話系統評估中的應用價值。

綜上所述，雖然有關交通事故領域的研究相當稀少，不過現有的研究展示了在法律方面與交通事故後續處理的解決方案，以及對於大型語言模型的評估方法。然而，這些系統普遍受限於特定的領域如法條處理或是事故分析。本研究提出的 CCG 系統，通過結合大型語言模型的對話能力和信息擷取技術，旨在解決這些問題，提供更高效、更準確的交通事故處理解決方案。

3 Collision Care Guide (CCG) Agent

一般而言，交通警察處理車禍事故的筆錄過程主要是透過提問以及當事人的回答來獲取相關資訊，會依序向當事人提問以下有關交通事故的問題：

- 事故發生的時間地點、使用的交通工具

- 事故前的起點、行駛道路、行進方向、交通號誌、標誌是否清楚、和事故經過

- 天候、路況、交通流量、障礙物等

- 碰撞位置、車損情形、人員受傷情形

我們參考警察處理交通事故[7]筆錄時的標準程序，定義事故重點結構模板（包含事件細節及事件細節解釋），詢問車禍當事人各項細節，確保不會遺漏任何重要資訊。

本研究中，我們使用 OpenAI 提供的 GPT-3.5-turbo 作爲 CCG Agent 的核心基礎，結合設計的 prompt 與事故處理模板格式，來了解交通事故經過。

根據 OpenAI 提供的建議中[8]，比起在一個指令下執行兩種不同任務，將單一複雜任務拆分成兩項子任務，更能使大型語言模型更專注

[2] https://shunnarah.com/practice-areas/car-accident-lawyer/

[3] https://www.accidentconsults.com/

[4] https://lawyers.findlaw.com/

[5] https://laws010.com/blog/car-accident/car-accident-lawyer/car-accident-lawyer-01

[6] https://donotpay.com/

[7] https://td.police.gov.taipei/News_Content.aspx?n=9CDDA66829FF2249&sms=5E5FFE038245884F&s=9D88B5BC9452512F

[8] https://platform.openai.com/docs/guides/prompt-engineering

於當前任務，不僅能降低錯誤率，更易於控制與修正。

如圖 1所示，我們將 CCG 系統中處理交通事故的問答過程，分爲提問與資訊擷取兩個模組，前者任務是負責確認是否有無缺漏的資訊進而提出相關問題，後者任務是負責將使用者回覆的資訊，擷取資訊並正確紀錄至相應的問題欄位中。透過一問一答的方式和當事人互動，將這些車禍事故的回覆資訊擷取至自定義的儲存模板中。最後，當模板中所有所需資訊都獲得完後，透過事故經過生成模組還原成事故經過。

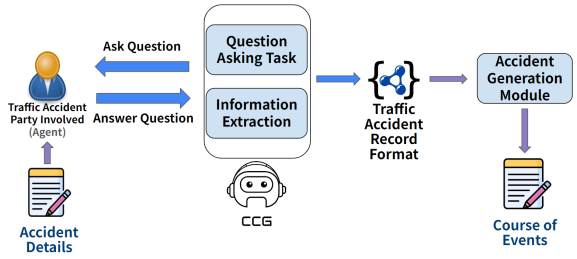

Figure 1: 多指令控制及事故經過生成流程

3.1 提問模組

在提問模組中，我們除了期望 CCG 能夠依照我們事先定義的資訊格式來詢問問題，也能夠依照當事人的狀況來適當地回覆。首先，CCG 要能夠根據當事人的回答，適切地以友善的語氣回應他們。如果當事人偏離了我們所要問的問題，CCG 會以專業且友善的態度引導當事人回到正題，正確地回答問題。如果當事人顯得緊張或不知所措，CCG 也能夠以安撫和鼓勵的話語幫助他們放鬆，並鼓勵他們提供所知的信息。當然若當事人正確地回覆了我們所提出的問題，CCG 也會適時地鼓勵他們繼續完成問答。

如同交通警察在筆錄時需要透過多輪對話了解事故經過一樣，CCG 的核心工作是提出問題並根據當事人的回覆獲取所需資訊。車禍事故記錄會以一個 JSON 格式來儲存，如表 2左欄交通資訊儲存格式 @TARF (Traffic Accident Record Format) 所示，這些交通資訊都是以警方會在車禍筆錄中所提問的常見問題，當然這些資訊都可以自行定義增減問題。另外，問題解釋格式 @QEF (Question Explaination Format) 中，則是對應著這一系列事故屬性鍵的名詞解釋。

在此任務中，我們會主動檢查資訊儲存格式 @TARF 中尚未提問的屬性鍵，然後將這些屬性鍵對應到問題解釋格式 @QEF 中的屬性解釋，與表 1中的第一欄 Prompt 組合後，輸入給大型語言模型，讓語言模型生成一個不僅能回應當事人的回覆，並且提出下一個問題資訊給予當事人作回答。

3.2 資訊擷取模組

當使用者回覆問題後，我們再透過資訊擷取提示，依靠著大型語言模型的統整能力，將使用者的答案填入相對應的資訊儲存格式中。如表1第二欄所示，資訊擷取提示除了擷取任務的提示詞之外，還包括使用者收到提問模組的上一則問題、使用者的回覆以及當下的資訊儲存格式 @TARF。上一則問題及使用者的回覆有助於大型語言模型更了解當下的事故擷取任務是針對哪一個資訊，從而可以正確擷取至對應到儲存格式中，並且輸出一個更新完整的資訊儲存格式 @TARF。

3.3 CCG 模型控制流程

CCG 控制流程如 Figure 1所示。首先，CCG 會詢問使用者事故發生時的情形概況，請使用者簡要描述案件狀況，包括發生日期、時間、地點和事發經過。接著，透過初始資訊提取提示，CCG 會將使用者提供的資訊輸出成預設的資訊儲存格式 @TARF，完成第一輪詢問及擷取。接著 CCG 會檢查資訊儲存格式 @TARF 中是否尚有空屬性值，加上屬性的定義後，CCG 透過提問模組提示產生一個問題給使用者。使用者回覆有關這個問題的資訊後，資訊擷取模組會根據使用者的回覆和資訊擷取提示擷取相關資訊，並將其對應到資訊儲存格式中正確的屬性鍵值上，然後更新格式，形成一次完整的問答。如果未能將交通事故記錄儲存格式中的所有資訊擷取出來，則會重複使用提問模組和屬性擷取模組，直到獲得所有所需的資訊。

3.4 事故經過生成模型

如 Figure 2所示，事故經過生成任務可以看成是交通事故資訊擷取問題的逆向工程，若能夠將這些原本在交通事故陳述上擷取下來的資訊用 Json 儲存，透過事實經過生成模組來還原回陳述形式的交通事故，就可以將使視爲正確完整的擷取。如 Table 1中第三欄所看到的，我們將 CCG 與當事人交互對話後，最終的資訊儲存格式 @TARF，將此 JSON 結果格式配合設計的提示詞，共同輸入給 GPT，就能得到敘述形式的交通事故陳述，整體流程如同 Figure 1後半部所示。最後不管是利用資訊儲存格式 @TARF 還是交通事故陳述，在後續不管是警方或是保險理賠人員，抑或是律師，都能透過兩種格式來因應不同場合需要，能夠

模組	**Prompts**
提問模組	作爲一名車禍事故陳述輔助專家，你的兩項主要任務是： 1. 根據 [上一個問題] 與 [當事人回答]，適當的回應當事人。 2. 根據 [下一個欄位]，提出下一個詢問的問題。 若當事人回答不相關，友善提醒他們專注於問題。對緊張的當事人，提供鼓勵。 每次回答後，根據 [問題解釋] 提出清晰的問題。 - [上一個問題]: {上一個問題} - [當事人回答]: {當事人回答} - [下一個欄位]: {下一個問題}的意思是{下一個問題解釋} - [專家的回覆]:
資訊擷取模組	你是資訊擷取機器人，請從 [問題回覆] 中擷取值填入 [Json 格式]。 - [Json 格式]:{'事故發生日期': '108 年 10 月 29 日','事故發生時間': '18 時左右', '事故發生地點': '', '我方駕駛交通工具': '', ...,'我方出發目的是什麼': ''} - [問題]:車禍發生時您駕駛的是什麼交通工具? - [問題回覆]:我當時駕駛的是一輛車牌號碼 000-0000 號的自用小客車。 - [Json]:
事故經過生成模組	你是一個車禍事故陳述記錄專家，根據 [Json 格式] 的事實，敘述車禍經過。 只描述 Json 中提供的事實，不包含其他資訊。 [Json]:{'事故發生日期': '108 年 10 月 29 日','事故發生時間': '18 時左右', '事故發生地點': '高雄市鳥松區松藝路與圓山路口', ...,'我方出發目的是什麼': '未知'}

Table 1: 各模組之提示詞 (藍色字體每輪對話會替換不同欄位資訊)

@TARF 資訊儲存格式	**@QEF** 問題解釋格式
"事故發生日期": "",	"事故發生日期": "具體日期",
"事故發生時間": "",	"事故發生時間": "具體時間",
"事故發生地點": "",	"事故發生地點": "具體地址",
"我方交通工具": "",	"我方交通工具": "交通工具種類",
"對方交通工具": "",	"對方交通工具": "交通工具種類",
"我方行駛道路": "",	"我方行駛道路": "行駛的道路",
"我方行進號誌": "",	"我方行進號誌": "號誌狀態",
"事發經過": "",	"事發經過": "詳細經過",
"當天天候": "",	"當天天候": "天氣情況",
"道路狀況": "",	"道路狀況": "道路狀況",
"我方行車速度": "",	"我方行車速度": "行駛時的車速",
"我方車損情形": "",	"我方車損情形": "車輛損壞情況",
"我方傷勢": "",	"我方傷勢": "傷勢情況",
"對方車損情形": "",	"對方車損情形": "車輛損壞情況",
"對方傷勢": "",	"對方傷勢": "傷勢情況",
"從哪裡出發": "",	"從哪裡出發": "出發地點",
"出發目的地": "",	"出發目的地": "目的地",
"出發目的爲何": "",	"出發目的爲何": "出發目的"

Table 2: 左邊爲資訊儲存格式 @TARF，右邊爲問題解釋格式 @QEF。

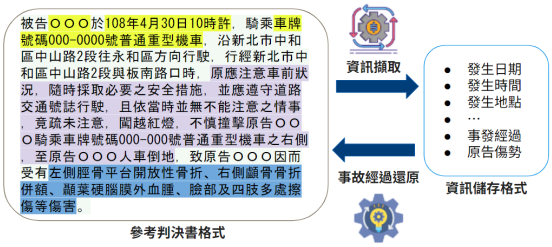

Figure 2: 交通事故資訊擷取 vs. 事故經過生成任務定義

快速獲得當事人在此次車禍事故中的資訊，不必重複的詢問，浪費雙方的寶貴時間。

3.5 系統介面介紹

我們設計的 CCG 系統的介面預覽如圖 3 所示。此系統是通過結合多個模組，包括提問模組、資訊擷取模組以及事故經過生成模組，來實現與使用者的互動對話。本系統包含三個核心組件。當使用者發起對話時，系統首先會通過提問模組向使用者提出問題。這些問題根據預設的交通事故記錄模板設計，旨在收集詳細且準確的事故信息。提問模組能夠靈活應對使用者的各種回應，引導他們提供必要的細節。使用者回答問題後，資訊擷取模組會即時分析並擷取關鍵信息，將這些信息儲存在相應的資訊儲存格式中。這一模組利用大型語言模型的強大能力，確保信息的準確性和完整性。當所有必要信息收集完成後，事故經過生成模組會根據儲存的資料生成完整的事故陳述。這一過程包括對信息的整合和重建，生成一份詳細的事故報告，供警方或保險理賠員使用。

這套模組化系統通過整合提問模組、資訊擷取模組和事故經過生成模組，提供了一個高效、準確的交通事故信息記錄與報告平台。這一系統不僅能夠提高事故處理的效率，還能顯著減少警方和保險理賠員的工作量，爲交通事故處理帶來革新性的改變。

4 Experiments

我們期望 CCG 在與當事人對話時，不僅可以使用人性化的語句詢問當事人，還能夠將當事人提供的車禍資訊準確地記錄到資訊儲存格式中。爲了達成這一目標，我們採用了兩項方式來評估：人工評估與 LLM 對話自動評估。在 CCG 中，最終目的也是需要眞實人類的使用，因此我們讓人類受測員參考判決書內容進行回覆，並以精確度 (Precision)、召回率 (Recall) 以及 F1 分數 (F1 Score) 爲指標，這

Figure 3: CCG 系統架設畫面預覽，(1) 爲對話內容，根據聊天內容藉由資訊擷取模組產生交通事故資訊儲存格式。(2) 的對話內容根據提問模組產生。(3) 爲事故經過生成模型。

些指標能全面反映 CCG 在資訊擷取的準確性與完整性。此外，我們也在人工評估中驗證 AI 當事人的表現，與眞實受測員的相似度，確保其能夠準確模擬眞實場景。

在 LLM 對話自動評估中，我們依照過去的研究如 LLM-Eval 自動評估的做法 (Lin and Chen, 2023)，利用大型語言模型資料分析的能力，來協助評估擷取與完整性這兩個指標任務。因爲我們本身 CCG 系統是採用 GPT，爲了避免潛在的偏見，在評估階段分別使用 GPT 和 Gemini 兩種 LLM 來進行獨立的綜合評估，以避免依賴單一 LLM 的評估結果來做出定論。我們希望這樣的綜合評估能夠更全面地了解 CCG 在與 AI 代理人對話中的表現，以及對車禍資訊的準確記錄能力，以獲得更具代表性和可信度的評估結論。

4.1 Human Evaluation

在本次實驗中，我們針對人類代理人評估進行設計，旨在評估 CCG 系統在眞實場景中的表現。因爲在現實我們無法取得眞實車禍的資訊，因此讓受測員扮演了參考判決書中車禍場景的原告，並且依照原告的主述，誠實不虛假的回覆每個 CCG 的問題，就像在眞實與警察做筆錄時需要誠實的回答所需的資訊。此實驗方法考慮了 CCG 是否正確擷取了當事人提供的信息，但在參考判決書中的原告敘述內容通常只會針對事發經過與傷害損傷賠償爲重點做論述，大多資訊如天氣、道路狀況與行車目的等內容並未提及，在目前我們自定義的 18 個欄位中平均約只有 8 至 10 個欄位有提及。因此我們在評估計算時，會以每筆判決書中提及的欄位內容爲主，其餘未談及的欄位將不列入計算。

實驗設計流程包括以下步驟：

1. 資料準備：首先，我們使用 GPT-4o 對判決書共 25 篇進行資訊擷取至 @TARF 中，並由人工檢查確認擷取結果的正確性，這些人工修正後的結果作爲參考。這些正確擷取的欄位數（去除未提及的欄位）記錄爲 a_i。

2. 受測員問答：受測員會隨機抽取判決書內容，模擬當事人進行回答，並提供相關的車禍信息，CCG 系統根據受測員的回答進行擷取。CCG 詢問的欄位數（去除 GPT 未擷取的欄位）記錄爲 q_i。

3. 數據記錄：最終記錄每次受測員與 CCG 對話後，正確擷取至資訊儲存格式 @TARF 的欄位數（去除 GPT 未擷取的欄位）記錄爲 e_i。

4. 計算指標：根據下述公式計算精確度（Precision）、召回率（Recall）和 F1 分數（F1 Score）。

- 精確度（**Precision**）：

$$P(i) = \frac{e_i}{q_i}$$

  精確度表示在 CCG 擷取出的所有欄位中，正確的比例是多少。

- 召回率（**Recall**）：

$$R(i) = \frac{e_i}{a_i}$$

  召回率表示所有應擷取的欄位中，CCG 成功擷取的比例是多少。

- **F1** 分數（**F1 Score**）：

$$F(i) = \frac{2 \cdot P(i) \cdot R(i)}{P(i) + R(i)}$$

  F1 分數是精確度和召回率的調和平均數，提供了一個整體的性能評估。

4.1.1 AI Agent (GPT Agent)

爲了更全面地評估 CCG 的表現，除了眞人受測員，我們同時使用 GPT-3.5-turbo 作爲 AI 代理人進行模擬測試。我們設定 AI 代理人參考與人類代理人相同的判決書內容，並依據經過精心設計，多次迭代的提示詞如 Table4下，確保 AI 代理人會依據判決書的內容，如實的描述判決書上的內容來進行測試。AI 代理人進行的結果，將與眞人測試結果進行比較，評估 AI 代理人是否能與人類受測員結果相近，來進行大規模的對話測試，以更好的評估 CCG 的表現。

4.1.2 結果

我們共蒐集了 25 筆對話，分別使用人類代理人和 AI 代理人進行測試。平均每筆對話的對話輪次爲 12.68 次（人類代理人）和 10.92 次（AI 代理人），對話輪次意思也就是使用者回答一句，加上 CCG 回覆一句，兩句話爲一輪次。結果顯示，CCG 系統在人類代理人測試中的精確度爲 0.923，召回率爲 0.897，F1 分數爲 0.909；而在 GPT 代理測試中的精確度爲 0.919，召回率爲 0.900，F1 分數爲 0.908。從結果中可以看出，CCG 在處理交通事故相關信息時，無論是面對人類代理人還是 AI 代理人，在 F1 分數都超過了 0.9，這代表著只要在使用者能夠正確的敘述資訊時，CCG 就能夠有效地擷取大部分必要的欄位，準確並且完整的紀錄在正確的欄位上。這使得我們有信心的可以說明，CCG 系統在眞實場景中能夠實際應用的潛力，特別是在處理交通事故相關訊息時的特定領域。

AI 代理人和人類代理人在整體 F1 分數上非常接近，但我們觀察到在某些特定類型的問題上存在細微差異。例如，AI 代理人在回答關於精確時間和地點的問題時，只要能在判決書上呈現的都表現優異，而人類代理人在描述主觀感受時更爲自然，這些差異提醒我們在解釋結果時需要考慮 AI 代理人的局限性。雖然 AI 代理人足以驗證能進行大規模、低成本的對話模擬測試，但我們也認識到它可能無法完全捕捉人類行爲的複雜性和不可預測性。因此，我們將後續 AI 代理人在 LLM 自動評估視爲人類評估的延伸補充，而非替代。

	平均輪次	P	R	F1
人類 Agent	12.7	0.923	0.897	0.909
GPT Agent	10.9	0.919	0.9	0.908

Table 3: CCG 系統在人類代理人和 AI 代理人評估中的表現。P 代表 Precision、R 代表 Recall、F1 代表 F1 Score。

4.2 LLM Evaluation

在前面提到了 AI 代理人與人類代理人的相似性，因此爲了評估 CCG 系統，我們需要使用大量的對話資料。我們設定 AI 代理人如上述一樣，設定相同的提示詞與規則 Table4，與 CCG 進行對話互動，以測試 CCG 的擷取模組是否能夠精確地擷取 AI 代理人的回應並將其記錄到正確的欄位中，評估整體對話的完整性和準確性。我們使共使用 587 篇的判決書隨機選擇，使用其中的車禍事實陳述作爲 AI 代理人的輸入，將 AI 代理人模擬成判決書中的原告，讓其與 CCG 進行交互問答，講述車禍相關的資訊。在評估方面，我們設定了兩種評估指標，擷取評估與整體評估，前者專注於 CCG 在單句話的擷取表現，後者專注於最終 @TARF 格式中整體的擷取結果。

4.2.1 擷取評估

在擷取評估中，我們對每一次 CCG 與 AI 代理人的對話進行評估，專注於 AI 代理人根據判決書中的車禍事實所做的回答，以及 CCG 提出的問題是否準確地被擷取並記錄到相應的 @TARF 的空欄中。我們使用 GPT-3.5-turbo 與 Gemini-1.0-pro 兩個 LLM 做爲評估模型，每次對話都會做一次準確性評分，評分範圍爲 1 至 10 分，並取每次對話的平均值作爲對話準確性的評估指標。而擷取評估中的 prompt 如 Table4，並在每次對話中，CCG 的問題、AI 代理人的回答、以及擷取後的 @TARF，一同加入提示詞作爲輸入。爲了進行評估，我們共收集了 587 筆對話，而對話的平均輪次爲 10.8 次，如圖4中 (b) 部分所示。我們將所有對話的準確性評估平均分繪製成分布圖如圖4中 (a) 所示。根據我們的評估，GPT-3.5-turbo 和 Gemini 的平均分別爲 7.93 分和 8.16 分。總體而言，平均準確性達到 8 分，可以表示 CCG 在單句擷取當事人所提供的資訊，大致上可以準確地被擷取到正確的欄位中。值得注意的是，因爲參考判決書中部分資訊並未提供，例如天氣、道路狀況或交通號誌等等，所以 AI 代理人若因爲部分資訊未提供而無法回答的情形，我們的評分將其給予 5 分，即使是這樣的評分標準，大多數情況下，系統能夠精確地擷取所需資訊。

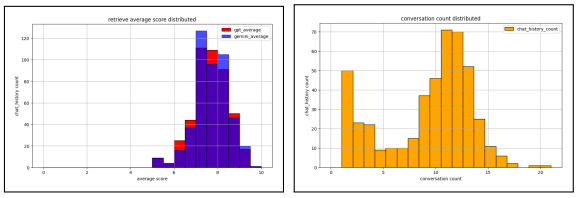

Figure 4: (a) 擷取評估準確性平均分數分布 (b) 對話輪次分布

4.2.2 整體評估

在整體評估中，我們的目標是確保最終的資訊儲存格式 @TARF 盡可能地與最初提供給 AI 代理人的判決書事故描述相符。爲了評估這一目標，我們對完整性和準確性分別給予了 1 至 10 分的評分。其中，完整性評分評估了是否將判決書事故描述中的所有資訊都完整地擷取至最終 @TARF 格式中，而準確性評分則評估了擷取到的資訊是否與判決書事故描述中所描述的一致且正確。我們一樣與上段擷取評

Agent 與評估	Prompts
AI 代理人	扮演 [車禍事故] 中的原告，簡答警察問題，未提及者回答' 不記得' 或' 忘記了'。 [車禍事故]:109 年 9 月 27 日 18:56，被告駕駛車牌 000-0000 自用車， 在新北市 0 路 0 段與華鋒三街口撞損原告車牌 000-000 重型機車，致修理費 14,200 元。 [警察的問題]:描述車禍當時的情況（時間、地點、經過等）
擷取評估	你是一位事故資訊分析師，依照以下內容確認使用者回答是否正確擷取到 Json 欄位。 ** 評分準則:** 準確性（1-10 分），資訊錯置扣分，回答不知情者 5 分。 ** 對話內容:** **CCG Agent 問題 **: CCG 提問 ** 使用者回答 **: User 回答 **Json**: @TARF ** 輸出格式:** 擷取評分: 整數分數 解釋說明: 指出正確與錯誤擷取，解釋分數。
整體評估	依照事發經過敘述與 JSON 資料，分析事故敘述與 Json 資料的匹配度， 基於以下標準給出完整性和準確性的整數分數（0 到 10 分）及詳細的評分理由。 ** 輸出格式:** ** 標準:** 完整性分數（根據敘述捕捉關鍵點），準確性分數（擷取內容是否匹配）。 ** 原因解釋:** 說明完整性和準確性評分理由，僅針對敘述提及項目。 ** 輸入:** ** 事發經過敘述 **：參考判決書事發經過 **JSON 資料 **：最終 @TARF 格式

Table 4: AI 代理人提示詞以及擷取評估、整體評估的提示詞。藍色字體在每輪會替換成不同的欄位資訊。

	對話總量	平均輪次	GPT	Gemini
值	587	10.8	7.93/10	8.16/10

Table 5: 對話評估結果

估相同，共收集了 587 筆對話，對話的平均輪次爲 10.8 次，使用了 GPT-4 和 Gemini-1.0-pro 進行綜合評估和比較。根據我們的評估結果，在準確性的部分，分布如圖5中所示，GPT 的準確性評分平均爲 7.41，Gemini 爲 8.31；在完整性的部分，GPT 的完整性評分平均爲 6.76，Gemini 爲 7.59，可以看到大部分的分數都落在 7 至 9 分，Gemini 給予較 GPT 高一點的分數，這表明 CCG 系統能夠大致上將 AI 代理人所敘述的資訊完整且準確地記錄在資訊儲存格式 @TARF 中。

模型	平均準確性分數	平均完整性分數
GPT-4	7.41	6.76
Gemini	8.31	7.59

Table 6: 準確性和完整性評估結果

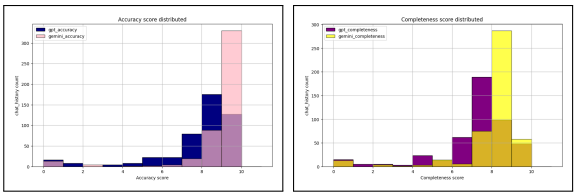

Figure 5: (a) 整體準確性評估分布 (b) 整體完整性評估分布

5 Limitation

在本研究中，我們設計並評估了 CCG 系統，但仍然存在一些限制需要考慮。

- 隱私問題：我們著重於保護當事人的隱私，並不會蒐集使用者的個人資料，若要在某些情況下需要蒐集，必需要進一步的數據隱私保護措施並告知使用者，以確保所有收集到的信息都得到妥善處理。
- 無法進行法律判決：CCG 系統的設計重點在於紀錄當下的車禍事件。在法律判決或賠償金額的決定上，CCG 系統無法替代專業的法律判斷，需要綜合雙方的證詞、影像、客觀錄音以及專家意見才能進行評估。
- 欄位信息不全：我們的實驗主要依賴於模擬環境和預定的判決書資料，在這之中部分欄位如天氣、道路狀況、行車目的等並無提及，這可能無法完全反映眞實世界中的多變情況。
- AI 代理人侷限性: 雖然 AI 代理人在我們的實驗中表現良好，但它們可能無法完全模擬眞實人類在壓力下的反應或不一致的回答，這可能導致在實際應用場景中出現一些未預期的情況。
- 實驗環境局限性：我們當前的實驗主要集中在交通事故場景，且基於中文的特定語言與文化上，在其他法律或非法律領域，

與不同語言和文化的應用效果尚未驗證，這是未來研究可以嘗試的方向。

這些限制說明了目前我們研究中的挑戰，並且提供未來改進的方向，爲後續的研究提供了參考。

6 Conclusion

CCG 系統旨在幫助當事人在車禍事件後，能夠盡速的協助當事人與警方紀錄相關的事件細節，減少重複繁瑣的提問過程，在日後車禍事故判斷或是賠償金額的計算前不必再次詢問。此外，透過我們設計的問答模式與模板設計，都可以根據實際需求進行調整和擴展，在未來可應用至更多法律問答機器人或非法律領域，例如民事、刑事案件的處理，或企業收集用户資訊等應用。

在實驗中，CCG 能夠依照所定義設計好所需的事件細節，準確的紀錄使用者所回答的資訊，進行了人工評估與 LLM 自動評估。在眞人測試中，眞人代理人和 AI 代理人測試與 CCG 的模擬對話中，CCG 系統在資訊擷取能力上表現優良，精確度、召回率和 F1 分數均落在約 0.9 上下。在 LLM 自動評估中，通過 GPT-4 與 Gemini-1.0-pro 的綜合評估下，CCG 在擷取評估與整體評估中也都達到約 7 分的標準，這顯示當使用者誠實的回覆問題下，我們不僅證明 CCG 能有效處理交通事故紀錄，即便在部分如天氣、道路狀況等資訊缺乏的情況下也拿精準將訊息擷取至正確的欄位中，也確保了評估的客觀性與可靠性。然而，我們認識到 AI 代理人能使我們進行大規模、低成本的測試，但 AI 代理人無法完全替代眞人測試，兩者應該相輔相成。

總結來說，CCG 系統展示了在交通事故處理中的潛力，但也存在改進空間。未來的工作將集中在提高 CCG 系統的整體程度，更能夠友善並有效地協助當事人，處理更多樣化的事故場景。我們也將致力於擴大眞實世界的測試範圍，適應更多樣化的事故場景，並探索 CCG 在其他法律和非法律領域的應用可能性，如民事糾紛、醫療諮詢等。本研究不僅爲交通事故信息收集提供了一個創新的解決方案，也爲人工智能在法律和公共服務領域的應用開闢了新的可能性。

References

Noushin Behboudi, Sobhan Moosavi, and Rajiv Ramnath. 2024. Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations?

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots.

Artur Grigorev, Khaled Saleh, Yuming Ou, and Adriana-Simona Mihaita. 2024. Enhancing traffic incident management with large language models: A hybrid machine learning approach for severity classification.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Vorada Socatiyanurak, Nittayapa Klangpornkun, Adirek Munthuli, Phongphan Phienphanich, Lalin Kovudhikulrungsri, Nantawat Saksakulkunakorn, Phonkanok Chairaungsri, and Charturong Tantibundhit. 2021. Law-u: Legal guidance through artificial intelligence chatbot for sexual violence victims and survivors. *IEEE Access*, 9:131440–131461.

Kebin Wu, Wenbin Li, and Xiaofei Xiao. 2024. Accidentgpt: Large multi-modal foundation model for traffic accident analysis.