基於微調開源大型語言模型的交通事故資訊蒐集代理人系統研究
The Study of a Traffic Accident Information Collection Agent System Based on Fine-tuned Open-Source Large Language Models

龔若齊 **Jo-Chi Kung**　　張嘉惠 **Chia-Hui Chang**
z1a2x3s4c5d6v7f8b9g@gmail.com, chiahui@g.ncu.edu.tw

摘要

本研究提出了一套名爲「交通事故資訊蒐集代理人」（Collision Care Guide, CCG）的系統架構，專注於事故初期階段的結構化資訊蒐集。CCG 整合三大模組：問題生成、資訊擷取及事故重建，透過多輪對話引導使用者敘述事故細節並轉換爲結構化資料格式（TARF），同時生成可讀性敘述供核對。爲滿足成本效益、隱私保護及部署彈性需求，本研究比較開源 Llama 模型（3B/8B 參數，完整微調及 4-bit PEFT 方法）與商業基準 GPT-4o-mini 的效能表現。結果顯示，資訊擷取模組欄位準確率高於 0.94，JSON 語義相似度達 0.995；問題生成模組語義相似度介於 0.85-0.88，問題表達更加精煉。微調模型在對話品質與資訊擷取的 LLM 評估中均獲得 4 分以上（滿分 5 分），與商業基準差距小於 0.5 分。研究證實開源模型經微調後能逼近商業模型效能，且量化版本在資源受限場景中具備高效能與部署潛力。CCG 的設計填補了事故初期互動式資訊蒐集的技術空白，爲交通事故處理提供了高效且具成本優勢的解決方案。

Abstract

This study introduces the "Collision Care Guide" (CCG), a system designed to collect structured traffic accident information during the early stages of an incident. CCG integrates three core modules: question generation, information extraction, and accident reconstruction. Through multi-turn dialogues, users are guided to describe accident details, which are then transformed into a structured format (TARF), alongside readable narratives for verification. To address cost efficiency, privacy protection, and deployment flexibility, this study compares the performance of open-source Llama models (3B/8B parameters with full fine-tuning and 4-bit PEFT methods) against the commercial baseline GPT-4o-mini. Results show that the information extraction module achieves field accuracy above 0.94 and JSON semantic similarity of 0.995, while the question generation module attains semantic similarity between 0.85-0.88 with more concise expressions. Fine-tuned models scored 4 (out of 5) in dialogue quality and information extraction evaluations, with differences from the commercial baseline within 0.5 points. Findings confirm that fine-tuned open-source models can achieve performance comparable to commercial models, with quantized versions demonstrating high efficiency and deployment potential in resource-constrained scenarios. The CCG design bridges the technical gap in interactive information collection during the early stages of accidents, offering a cost-effective and efficient solution for traffic incident management.

關鍵字：交通事故資訊蒐集、大型語言模型、對話式 Agent、資訊擷取、模型微調

Keywords: LLM、Conversation Agent、Information Extraction、Finetuning

1 緒論

道路交通安全問題長期爲全球公共衛生與基礎設施治理的重大議題[1]。在台灣，年度交通事故總件數自 2019 年約 34 萬件攀升至 2024 年將近 40 萬件[2]，平均每日逾千件事故。事故初期的人工筆錄流程面臨多重挑戰：當事人在事故後受情緒與壓力影響，難以完整準確地有條理敘述事故經過；事故量的攀升亦造成執法人員的人力資源負擔，影響後續責任釐清的效率與公正。上述背景凸顯了交通事故資訊蒐集的需求，尤其是在高事故量與人工處理效率間的矛盾，爲自動化解決方案提供了切入點。

[1] https://www.who.int/publications/i/item/9789240087712

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

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 71–79
November 20-22, 2025

現有交通人工智慧研究多著重事故風險預測 (Zhou et al., 2020)、多模態重建與責任分析 (Wu et al., 2024)，法律科技領域聚焦判決預測與文本檢索 (Chalkidis et al., 2022)。然而，針對事故發生後初期之即時、互動式、結構化的資訊蒐集相對缺乏。現有技術主要假設事故事實已充分記錄，忽略了事故初期口語敘述的稀疏性及不確定性，這形成了前置事實蒐集的技術缺口。

爲填補此缺口，本研究提出基於大型語言模型的交通事故資訊蒐集代理人 Collision Care Guide（CCG），透過多輪對話引導提問與結構化模板，實現使用者口語敘述與結構化紀錄的雙向轉換。CCG 系統由三個協同模組構成：(1) 問題生成模組：依據缺失資訊與前輪回覆動態生成聚焦提問；(2) 資訊擷取模組：自當事人口語敘述中擷取並填入結構化 JSON 欄位，處理模糊與不確定表述；(3) 事故重建模組：將完成之結構化紀錄重建爲條理清晰的敘述文本，以支援後續人員閱讀與理解。這些模組相互協作，形成完整的資訊蒐集流程，旨在提升事故初期事實蒐集的效率與準確性。

本研究基於Kung et al. (2024) 的研究，進一步考量部署成本、隱私與可控性需求，針對開源中小參數 Llama 模型（Llama 3.2 3B、Llama 3.1 8B）進行任務特化微調，並與商業基線 GPT-4o-mini 比較於兩項核心任務：資訊擷取與問題生成，同時評估多任務（Combined）訓練設定下之效能穩定性。測試集結果顯示，資訊擷取模組整體 JSON 語義相似度約 0.995，欄位完全準確度最高達 0.95；問題生成模組平均語義相似度達 0.85，提問長度平均差異約 20%，保留必要語義而減少冗餘同理心鋪陳。此外，4-bit 量化模型在保持高效能的同時，顯著降低了部署成本，驗證其在私有化部署中的潛力。

本文的主要貢獻如下：

1. 提出交通事故多輪對話資訊蒐集系統（CCG），能從當事人敘述中引導並擷取預先定義之關鍵資訊，實現結構化資料的雙向轉換，降低重複詢問與遺漏風險

2. 探索任務特化微調與多任務聯合訓練對兩核心任務（問題生成與資訊擷取）的穩定性與效能影響

3. 建立雙層評估（測試集指標與 LLM 評分），驗證小參數模型部署的成本效益與技術可行性

綜上所述，CCG 的設計不僅填補了事故初期資訊蒐集的技術空白，也通過實驗結果證實了開源模型的效能逼近商業模型的可行性，爲交通事故場景提供了一個高效且具成本優勢的解決方案。

2 相關研究

交通人工智慧技術在事故風險預測、事故重建與責任分析等領域取得了顯著進展。例如，Zhou et al. (2020) 等人提出的事故風險預測模型利用道路特徵與歷史事故數據進行風險評估，準確率達 90% 以上。Wu et al. (2024) 等人基於影像與感測器數據進行事故場景的重建，有效支持責任分析。然而，這些模型依賴完整的結構化輸入數據，難以應對事故初期口語敘述的稀疏性與不確定性。

法律科技研究主要集中於判決預測與文本檢索。例如，LexGLUE(Chalkidis et al., 2022) 基準系統能有效支持法律決策，但假設輸入事實已整理完備，難以處理事故初期的模糊敘述。保險系統則著重於理賠流程的自動化，依賴完整事故報告，缺乏即時處理口語敘述的能力。

開源模型微調研究顯示，針對特定領域任務的參數高效微調（PEFT）(Hu et al., 2021) 與量化技術 (Dettmers et al., 2023) 能在保持效能的同時降低部署成本。PEFT 技術通過調整少量模型參數，使模型能快速適應特定任務，顯著降低訓練成本。例如，在交通事故場景中，PEFT 技術能支持模型快速適應口語敘述的資訊擷取任務，提升事故初期資訊蒐集的效率。量化技術則通過使用低精度數據格式（如 4-bit 或 8-bit）來減少模型計算需求，保持效能的同時降低硬體資源消耗。Llama 系列模型 (Touvron et al., 2023) 在多任務學習與領域適應方面展現了潛力，爲私有化部署提供了高效且成本友善的替代方案。然而，針對事故初期資訊蒐集的特定場景，開源模型與商業模型的效能差異仍需系統性驗證。

綜上所述，現有研究在事故預測與分析方面已趨成熟，但事故初期的即時互動資訊蒐集仍缺乏系統化方案。同時，開源 LLM 在特定領域任務上的微調效能與部署可行性需要進一步驗證。因此，本研究聚焦於填補前置資料蒐集缺口，並探索成本效益與隱私友善的開源模型替代方案。

3 Collision Care Guide (CCG)

本章介紹交通事故資訊蒐集對話代理系統 Collision Care Guide (CCG)，其目標是在事故發生初期透過多輪互動提問取得事故關鍵資訊，並將資訊結構化，同時提供可讀性敘述供當事人核對。CCG 系統以模組化設計爲基礎，

整合三個核心模組：問題生成模組（Question Generation Module）、資訊擷取模組（Information Extraction Module）、事故重建模組（Accident Reconstruction Module），形成完整的資訊蒐集流程。

3.1 資料格式：TARF 與 QEF

TARF（Traffic Accident Record Format）是一個包含 18 個欄位的結構化資料格式，用於儲存事故相關資訊，包括基本情境、行爲路徑、環境條件、事件結果與動機用途等（如表 1 所示）。此格式確保事故資訊能以結構化方式進行存儲與檢索。

QEF（Question Explanation Format）則爲 TARF 中各欄位提供語義解釋與提問對齊說明，確保問題生成的語義精準性，減少術語歧義。例如，針對「事故發生地點」欄位，QEF 提供了具體的提問方式與語義參考。爲 TARF 各欄位提供語義解釋與提問對齊說明，確保問題生成的語義精準性，減少術語歧義。

Table 1: TARF 主要欄位與簡述

“事故發生日期”: 事故之具體日期
“事故發生時間”: 事故之具體時間
“事故發生地點”: 發生道路或地址
“對方駕駛交通工具”: 對方車種
“我方駕駛交通工具”: 我方車種
“我方行駛道路”: 我方所行經道路
“事發經過”: 簡述事故情節
“我方行進方向的號誌”: 相關號誌狀態
“當天天候”: 天氣情況
“道路狀況”: 施工／濕滑等狀態
“我方行車速度”: 事發時速度
“我方車輛損壞情形”: 我方車損
“我方傷勢”: 我方傷害
“對方車輛損壞情形”: 對方車損
“對方傷勢”: 對方傷害
“我方從哪裡出發”: 出發起點
“我方出發目的地”: 目的地
“我方出發目的是什麼”: 出發動機

3.2 模組化架構與多輪流程

CCG 系統採用模組化設計，由三個模組協同運作，形成完整的資訊蒐集與驗證流程。整體流程採用缺失導向的迭代策略（如圖 1 所示），具體包括以下步驟：**1.** 初始詢問：系統通過開放式問題獲取事故概要。**2.** 迭代循環：系統檢測 TARF 中的缺失欄位，並動態生成聚焦問題以補全資訊。**3.** 結構化更新：根據使用者回答更新 TARF 欄位，直至所有欄位填寫完成或達到最大輪數限制（20 輪）。**4.** 敘述重建：流程最後階段將結構化紀錄重建爲自然語言敘述供使用者核對，必要時進行局部修正。此設計確保資訊完整性與一致性，並形成可驗證的雙向轉換閉環。

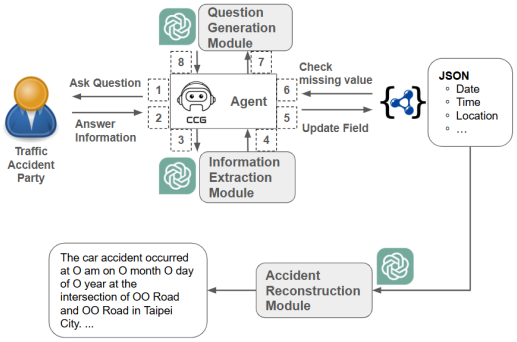

Figure 1: CCG 架構與多輪採集流程

3.3 問題生成模組

問題生成模組負責根據 TARF 填寫狀態與使用者回覆動態生成適當的回應與後續問題。該模組整合以下三項功能：在互動回覆方面，模組能對偏題回答進行專業引導，並對正確回覆給予正向回饋，從而維持良好的互動氛圍。動態檢查使模組能主動識別 TARF 中的未填欄位，結合 QEF 語義解釋生成自然且具體的提問，確保資訊蒐集的完整性。此外，語境維持透過參考前次對話內容（表 2），有效避免重複提問並保持語義連貫性。

例如，當 TARF 中「事故發生地點」欄位缺失時，模組會生成如下問題：「請問事故發生的具體地點是在哪裡？」

Table 2: 問題生成模組提示詞摘要（藍色文字表示每輪對話中動態替換的欄位資訊）

Prompt
作爲車禍事故敘述助理，主要任務包括：
1. 根據 [上一個問題] 和 [當事人回答] 給予適當回應
2. 根據 [下一個欄位] 詢問下一個問題
若回答不相關則重新引導；對焦慮當事人給予鼓勵
輸入參數：
- [上一個問題]：{previous_question}
- [當事人回答]：{user_response}
- [下一個問題]：{current_question}
- [問題解釋]：{qef_attributes}

3.4 資訊擷取模組

資訊擷取模組負責將使用者的自然語言回覆轉換爲結構化的 TARF 資料。該模組依據上下文理解回覆的語義脈絡，並遵循以下原則：（如表 3）1. 僅處理與當前問題直接相關的欄位，不進行跨欄推測。2. 面對「不記得」、「不知道」等回覆時，將欄位標記爲「未知」。3. 確保填入資訊忠實於原始回答內容。

例如，當使用者回答「高雄市楠梓區左楠路機車道」時，模組會更新 TARF 中「我方行駛

道路」欄位爲該內容。

Table 3: 資訊擷取模組提示詞摘要（藍色文字表示每輪對話中動態替換的欄位資訊）

Prompt
專業事件資訊擷取助理，從 [當事人回答] 中擷取資訊 並填入 [JSON 格式] 對應欄位 執行規則： • 僅處理與當前問題直接相關的 JSON 欄位 • 「不記得」、「不知道」等回應填入「未知」 • 確保填入資訊忠實於原始回答 輸入參數： - [JSON 格式]：{current_tarf} - [問題]：{previous_question} - [當事人回答]：{user_response}

3.5 事故重建模組

事故重建模組將結構化的 TARF 資料逆向轉換爲自然語言敘述，以驗證系統的雙向語義保眞度（如圖 2所示）。此模組的設計體現了雙重資料格式的價值：結構化 **TARF** 格式：適合自動化處理與快速檢索的應用場景，如保險理賠處理與法律文件生成。自然語言敘述：更適合人員閱讀理解的情境，包括事故報告撰寫與當事人陳述確認。

例如，根據 TARF 中的資料，模組生成如下敘述：「事故發生於 2024 年 7 月 15 日早上 8:30，地點爲高雄市楠梓區左楠路。當時天氣晴朗，路面乾燥，事故涉及我方機車與對方轎車。機車行駛速度約 40 公里每小時，事故導致我方車輛右側損壞。」

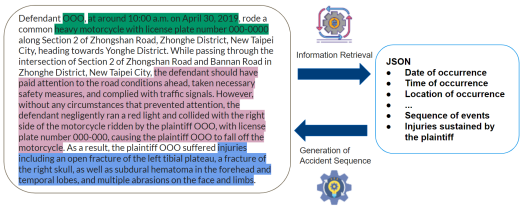

Figure 2: 資訊擷取與事故重建之雙向關係

Table 4: 事故重建模組提示詞摘要（藍色文字表示動態替換的欄位資訊）

Prompt
車禍諮詢專家，根據 [JSON 格式] 中的事實， 用敘述方式重述整個車禍經過 執行規則： • 僅描述 JSON 中提供的車禍相關事實 • 不包含其他無關或未提供的資訊 • 組織成連貫且邏輯清晰的事故敘述 輸入參數： - [JSON 格式]：{final_tarf}

4 模型訓練

本章驗證任務特化微調後之開源模型（Llama 3.2 3B，Llama 3.1 8B）在 CCG 系統的兩核心任務（問題生成、資訊擷取）中，分別進行單任務訓練與多任務聯合（Combined）設定，並透過測試集與 LLM 自動評估指標，檢驗其效能是否能逼近商業基準（GPT-4o-mini）。此外，本研究評估 4-bit 量化參數微調（PEFT）技術在成本、隱私與部署彈性場景中的替代可行性。

4.1 模型選擇與實驗環境

Llama 模型憑藉其在語義理解、生成能力及開源特性，成爲本研究的首選。相比商業模型（如 GPT-4o-mini），開源模型具有更高的成本效益及部署靈活性。爲進一步降低記憶體需求，採用 PEFT 技術（LoRA）進行 4-bit 量化，確保模型在資源受限場景中的實際應用能力。

所有訓練實驗均於 Nvidia RTX A6000 GPU 上執行，該 GPU 提供 48GB 的顯存，能有效支持大規模模型的微調，並顯著縮短訓練時間。整體訓練流程採用 Unsloth 框架[3]，該框架通過記憶體優化技術，顯著降低 GPU 記憶體占用並提升訓練效率。

4.2 訓練資料準備

4.2.1 資料格式設計

爲符合 CCG 系統的推論流程，本研究將對話狀態、缺失欄位提示、問題語境與生成約束整合爲單一 Instruction，並將輸出限定爲目標（下一個問題或更新後的 TARF）。此格式設計有助於模型專注於格式一致性與語義對齊，並便於單任務與聯合訓練模型共享資料結構。

每筆資料包括兩個主要部分：Instruction 與 Output。表 5 與表 6 分別展示了問題生成與資訊擷取的訓練資料示例。

Table 5: 問題生成訓練資料示例（節錄）

Instruction
作爲事故敘述助理，依上下文提出下一則問題。 上一問題：請簡述事故。回答：…… 欄位缺失：我方行駛道路； 欄位說明：事故時我方行駛的具體道路名稱。
Output
請問事故發生時您行駛的道路名稱是什麼？

4.2.2 標籤平衡策略

原始判決文本偏重法律裁判目的，導致 TARF 18 個欄位中僅約 8-10 個被明確提

[3] https://unsloth.ai/

Table 6: 資訊擷取訓練資料示例（節錄）

Instruction
依據回覆更新 JSON。 現有 JSON：{事故發生日期: ...}； 問題：您行駛的道路？ 回覆：高雄市楠梓區左楠路機車道。
Output
{事故發生日期: 民國 108 年 4 月 2 日, 事故發生時間: 07:28, 我方行駛道路: 高雄市楠梓區左楠路機車道, ...}

及。爲避免模型學得過於保守的策略，本研究對訓練資料進行標籤擴增，將資訊分爲三類：`normal`（準確資訊值）、`unknown`（明確未知）、`other`（模糊/無法解析）。經擴增後，訓練集的標籤分布爲 `160:198:20`（約 42.3%/52.4%/5.3%），測試集的分布爲 `39:46:3`（約 44.3%/52.3%/3.4%），確保具體值與未知標記的決策邊界更加平衡。

4.2.3 多樣性與品質控制

爲提升模型的泛化能力，本研究從敘述密度（詳細/中等/精簡）與回答風格（五類語用態度）生成多樣化訓練語料。敘述密度分爲：詳細型：涵蓋大部分 TARF 欄位，內容完整，類似完整事故報告、中等型：僅包含主要事故資訊，省略部分次要細節、精簡型：僅敘述核心事故事實，形式精煉。

回答風格包含五類：驚慌失措型（語序跳躍、重複）、冷靜理性型（邏輯線性、資訊密集）、防禦戒備型（對責任相關細節保留）、創傷恍惚型（不確定詞頻繁、時間順序模糊）、急躁不耐型（回答簡短、易省略修飾）。此考量眞實情形中資訊不完整的場景，同時反映不同當事人的敘述習慣和記憶能力差異，旨在使訓練資料能夠反映眞實世界中當事人的多樣化回答模式。

資料生成流程基於判決書樣式與 TARF 欄位模板，運用多個大型語言模型（Claude 4 Sonnet、Gemini-2.5-pro、GPT-4.1）產出三種事故敘述密度初稿，設定特定風格的當事人角色與基準模型 CCG 進行對話擴展並植入標籤，經人工審核格式合法後形成訓練/測試資料。

4.2.4 資料統計

最終資料集包含：訓練集 40 篇對話（資訊擷取 378 筆樣本、問題生成 338 筆樣本）與測試集 10 篇對話（資訊擷取 88 筆樣本、問題生成 78 筆樣本），資料分布保持與實際多輪流程相近，覆蓋全部密度與風格組合。

4.3 模型選擇與訓練配置

本研究比較 3B 與 8B 兩種模型規模，並採用完整微調（Full Fine-Tuning）與 4-bit 量化 LoRA（PEFT）技術進行對照。具體配置如表 7 所示。

Table 7: 模型訓練配置

名稱	版本	訓練方法	量化
Llama_3B_4bit	3.2 3B	PEFT (LoRA)	4-bit
Llama_3B	3.2 3B	Full FT	None
Llama_8B_4bit	3.1 8B	PEFT (LoRA)	4-bit
Llama_8B	3.1 8B	Full FT	None

5 模型效能評估結果

本章節旨在比較微調模型與基準模型 GPT-4o-mini 在資訊擷取與問題生成任務中的效能表現。評估過程採用 (Kung et al., 2024) 所蒐集的 754 筆對話資料，並透過雙評估器（Gemini-2.0 與 GPT-4o）進行交叉評分，以檢驗模型在對話品質及資訊擷取整體品質上的一致性與穩健性。

5.1 測試集驗證概述

本研究主要聚焦於兩項核心任務：資訊擷取與問題生成。在資訊擷取任務中，模型效能以欄位層級的精確性及語義保眞度進行評估，並透過整體 JSON 一致性來衡量模型跨欄位的語義與結構表現。在問題生成任務中，則著重於模型生成問題的語義覆蓋率及表達精煉程度。此外，爲驗證多任務訓練的成效，研究進一步探討共享語義表示的穩定性是否得以維持或提升。

5.2 資訊擷取效能分析

爲評估模型效能，本研究採用六項指標：完全準確度（**Exact Accuracy**）衡量欄位輸出是否與基準一致；語義相似度（**Semantic Similarity**）以 Sentence Transformer 計算嵌入向量餘弦相似度，範圍 [-1,1]，越接近 1 語義越一致；高語義準確度（**High Semantic Accuracy**）爲語義相似度高於 0.8 的有效欄位比例；未知／空值不匹配率（**Unknown / Empty Mismatch Rate**）檢驗模型對不確定資訊的決策一致性；整體 **JSON** 相似度（**Overall JSON Similarity**）評估跨欄位語義與結構的一致性。

上述指標核心公式如下：

$$\text{Accuracy}_j = \frac{1}{|\mathcal{V}_j|} \sum_{i \in \mathcal{V}_j} \mathbf{1}\big(o_{i,j}^{(f)} = o_{i,j}^{(b)}\big)$$

Table 8: 資訊擷取模組測試結果（欄位層級彙總指標）

模型	Exact Accuracy	High Semantic Accuracy	Unknown Mismatch	Empty Mismatch	Overall JSON Similarity
3B	0.9508	0.9654	0.0139	0.0183	0.9946
3B_4bit	0.9508	0.9694	0.0139	0.0189	0.9948
8B	**0.9571**	0.9621	0.0145	0.0152	**0.9957**
8B_4bit	0.9539	0.9627	0.0139	0.0158	0.9942
Combined_3B	0.9489	0.9673	0.0170	0.0227	0.9942
Combined_3B_4bit	0.9470	0.9618	0.0158	0.0202	0.9937
Combined_8B	0.9558	**0.9702**	0.0139	0.0177	0.9954
Combined_8B_4bit	0.9558	0.9701	**0.0107**	**0.0145**	0.9954

$$\text{Unknown\_Mismatch}_j = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big(u_{i,j}^{(f)} \neq u_{i,j}^{(b)}\big)$$

其中，N 爲測試樣本總數（88）；$\mathcal{V}_j$ 爲欄位 j 之有效樣本集合；$o_{i,j}^{(f)}$、$o_{i,j}^{(b)}$ 分別爲微調與基準模型輸出；$\mathbf{1}(\cdot)$ 爲指示函數。

表 8 彙總了模型在資訊擷取任務中的效能指標。結果顯示，8B 模型在完全準確度（0.9571）及 JSON 相似度（0.9957）方面表現最佳，展現卓越的語義與結構一致性。4-bit 量化版本的 Combined_8B_4bit 模型在未知（0.0107）與空值（0.0145）不匹配率上最低，證明量化技術與多任務訓練未影響穩定性。相比之下，8B 模型效能略高於 3B 模型，而量化版本顯著降低計算資源需求，適合資源受限場景。

表 9 顯示，Combined_8B_4bit 模型大多數欄位的完全準確度超過 0.95，標準化欄位（如車種、天候）語義相似度達 0.985 至 1.000。長敘述欄位（如「事發經過」）的完全準確度較低（0.898），但語義相似度達 0.995，顯示模型能有效掌握長文本語義。道路狀況與速度欄位因表達多樣性導致未知不匹配率略高（≥ 0.034），但仍在可接受範圍。

綜上，微調模型在結構化欄位填寫與語義重現上接近商業基準，量化技術未對主要指標造成影響，多任務設定進一步提升語義泛化能力與穩定性。

5.3 問題生成效能評估

本研究針對問題生成任務進行了全面的性能評估，測試集包含 78 筆樣本，主要分析模型在語義覆蓋及表達簡潔性方面的表現。評估指標包括：高語義準確度（語義相似度閾值 0.8 的比例）、中等語義準確度（語義相似度閾值 0.6 的比例）、平均語義相似度（Avg Semantic Similarity），以及長度相似度（Avg Length Similarity）。

長度相似度（**Length Similarity**）用於衡量微調模型生成問題的長度與基準模型生成問題長度的相似程度，反映問題表達的完整性與精煉性。該指標的計算公式如下：

$$\text{AvgLengthSim} = \frac{1}{N}\sum_{i=1}^{N}\frac{\min(|q_i^{(f)}|, |q_i^{(b)}|)}{\max(|q_i^{(f)}|, |q_i^{(b)}|)}$$

其中 $q_i^{(f)}$ 和 $q_i^{(b)}$ 分別代表微調模型與基準模型生成的第 i 個問題的長度，N 爲樣本總數。

Table 9: Combined_8B_4bit 各欄位指標

欄位	完全準確度	語義相似度	未知不匹配率
事故發生日期	0.955	0.982	0.011
事故發生時間	0.955	0.982	0.000
事故發生地點	0.955	0.984	0.011
對方駕駛交通工具	0.989	0.994	0.000
我方駕駛交通工具	0.977	1.000	0.000
我方行駛道路	0.977	0.996	0.000
事發經過	0.898	0.995	0.000
我方行進方向的號誌	0.966	1.000	0.011
當天天候	0.989	1.000	0.011
道路狀況	0.932	0.973	0.034
我方行車速度	0.955	0.910	0.034
我方車輛損壞情形	0.943	0.988	0.000
我方傷勢	0.932	0.992	0.000
對方車輛損壞情形	0.955	0.984	0.000
對方傷勢	0.932	0.965	0.023
我方從哪裡出發	0.989	1.000	0.000
我方出發目的地	0.977	0.985	0.023
我方出發目的是什麼	0.932	0.987	0.034

表 10 彙總了各模型在問題生成任務中的測試結果。所有模型的平均語義相似度均達到 0.8323 以上，其中 Combined_8B_4bit 在高語義準確度（0.7949）及平均語義相似度（0.8841）方面表現最佳，而 Combined_8B 和其量化版本在中等語義準確度上均達到 0.9872，展現了近乎完整的語義覆蓋能力。此外，所有模型的長度相似度集中於 0.78 至 0.80，顯示微調模型生成的問題平均比基準模型短約 20%，但仍保留了核心語義，呈現更直接且精煉的提問風格。

以下以道路狀況欄位生成問題爲例，展示基準模型與微調模型的語言特徵差異：

Table 10: 問題生成模組測試結果

模型	High Semantic Accuracy	Avg Semantic Similarity	Medium Semantic Accuracy	Avg Length Similarity
3B	0.6538	0.8486	0.9103	0.7919
3B_4bit	0.6667	0.8323	0.8974	0.7848
8B	0.7179	0.8601	0.9487	0.7779
8B_4bit	0.6923	0.8545	0.9487	0.7850
Combined_3B	0.6667	0.8508	0.9359	0.7876
Combined_3B_4bit	0.7308	0.8671	0.9615	**0.7990**
Combined_8B	0.7564	0.8761	0.9872	0.7822
Combined_8B_4bit	**0.7949**	**0.8841**	**0.9872**	0.7939

- **GPT-4o-mini**：「理解您的情況，燈號有時可能會讓人感到模糊。那麼，當時您所行駛的道路狀況如何？例如，那條路上是否有施工、破損或者是路面濕滑？」
- **Llama-3B**：「請您回想一下當時行駛的道路狀況如何？例如，有沒有施工、路面是否濕滑或有其他障礙？」

基準模型的問題通常包含情境鋪陳及同理語句，旨在增加對話的自然性與情感連結；而微調模型則省略冗長的轉接語，直接聚焦於核心語義，符合任務功能性需求。這種語言特徵的差異表明微調模型在提升表達簡潔性的同時，仍能維持語義完整性。

5.4 對話品質綜合評估

本節針對 754 筆對話進行雙評估器交叉驗證，分別使用 Gemini-2.0 與 GPT-4o 評估對話品質，涵蓋三項核心指標。評估方法參考 GPTScore (Fu et al., 2023) 與 G-Eval (Liu et al., 2023) 等研究，採用統一提示模板進行綜合評分。評分採用 5 分制量表，指標定義如下：流暢性（**Fluency**）評估系統生成回應的語言自然度與流暢性；關聯性（**Relevance**）衡量回應內容與用户描述或問題的相關性；連貫性（**Coherence**）檢驗整體對話流程的邏輯一致性。

表 11 彙總了各模型在流暢性、關聯性、連貫性及整體評分的表現。結果顯示，Gemini-2.0 評估中，微調後的 Llama-8B 及其量化版本在整體評分上達到 4.74—4.76，與基準 GPT-4o-mini（4.65±0.60）僅有微小差距。相比之下，GPT-4o 的評分普遍低於 Gemini-2.0，下降幅度約 0.4—0.5 分，顯示 GPT-4o 採用了更嚴格的評估標準。

在流暢性（Fluency）方面，所有模型均達到高分，其中 GPT-4o-mini 在 Gemini-2.0 評估中得分爲 5.00±0.06，展現了極高的語言自然度。然而，微調後的 Llama-8B 及其量化版本在關聯性（Relevance）與連貫性（Coherence）指標上的表現更爲穩定，整體評分達到 4.74—4.76，接近基準商業模型 GPT-4o-mini。

相比之下，GPT-4o 評估結果顯示微調模型的整體得分略低，主要體現在關聯性與連貫性指標上。例如，Llama-8B 的 GPT-4o 評分爲 4.28±0.69，低於其在 Gemini-2.0 中的 4.76±0.62。這表明，GPT-4o 更加注重對話回應的語義深度與邏輯一致性，導致評分標準更爲嚴苛。

整體而言，微調後的 Llama 系列模型在對話品質評估中表現穩定，尤其在 Gemini-2.0 評估中接近商業模型的效能，證實了開源模型的潛力及量化技術的實務可行性。

5.5 資訊擷取能力綜合分析

本節針對同批對話資料進行資訊擷取能力的全面評估，涵蓋三項核心指標，均採用 5 分量表進行評分：事實一致性（**Fact Consistency**）檢驗 JSON 格式中提取的資訊是否準確反映對話內容；資訊完整性（**Information Completeness**）評估 JSON 是否涵蓋當事人描述的所有必要事故要素；描述合理性（**Description Reasonability**）判斷生成的資訊描述是否合乎邏輯、客觀且對未提及的資訊正確標記爲未知。

表 12 彙總了各模型在資訊擷取任務中的表現，涵蓋事實一致性、資訊完整性、描述合理性及整體評分。Gemini-2.0 評估結果顯示，所有模型的整體得分均達 4.93 以上，其中 Llama-8B-4bit 模型以 4.97 的得分表現最佳，與基準 GPT-4o-mini（4.96±0.20）差距極小，顯示其卓越的資訊擷取能力。

相比之下，GPT-4o 評估標準更爲嚴苛，模型得分略低於 Gemini-2.0，但 Llama-8B 系列模型仍保持穩定表現，整體得分（4.84—4.86）略高於基準 GPT-4o-mini（4.82±0.40）。此外，量化技術未對模型性能產生顯著影響，量化版本（4-bit）與未量化版本的得分差距不超過 0.03，證實其在資源受限場景中的部署價值。

Table 11: 對話品質評估結果（流暢性、關聯性、連貫性）

模型	Gemini-2.0				GPT-4o			
	Fluency	Relevance	Coherence	Overall	Fluency	Relevance	Coherence	Overall
GPT-4o-mini	5.00±0.06	4.64±0.61	4.63±0.62	4.65±0.60	4.98±0.14	4.50±0.65	4.51±0.62	4.64±0.47
Llama-3B	4.74±0.49	4.68±0.66	4.63±0.73	4.65±0.71	4.54±0.51	4.14±0.75	3.96±0.87	4.14±0.74
Llama-3B-4bit	4.72±0.49	4.67±0.63	4.61±0.71	4.63±0.68	4.51±0.52	4.19±0.76	3.99±0.90	4.20±0.74
Llama-8B	4.85±0.38	4.77±0.64	4.75±0.66	4.76±0.62	4.66±0.47	4.28±0.70	4.14±0.82	4.28±0.69
Llama-8B-4bit	4.82±0.40	4.76±0.61	4.73±0.65	4.74±0.62	4.54±0.50	4.11±0.72	3.94±0.83	4.12±0.71
Llama-3B (Combined)	4.73±0.48	4.67±0.67	4.62±0.73	4.63±0.70	4.54±0.51	4.15±0.73	3.96±0.85	4.15±0.73
Llama-3B-4bit (Combined)	4.68±0.52	4.62±0.68	4.54±0.81	4.57±0.77	4.49±0.51	4.06±0.81	3.86±0.92	4.06±0.79
Llama-8B (Combined)	4.84±0.38	4.76±0.66	4.75±0.66	4.76±0.63	4.60±0.49	4.20±0.72	4.04±0.83	4.19±0.72
Llama-8B-4bit (Combined)	4.85±0.36	4.77±0.65	4.76±0.64	4.76±0.63	4.63±0.48	4.24±0.72	4.09±0.83	4.24±0.72

Table 12: 資訊擷取評估結果（事實一致性、資訊完整性、描述合理性）

模型	Gemini-2.0				GPT-4o			
	Consistency	Completeness	Reasonability	Overall	Consistency	Completeness	Reasonability	Overall
GPT-4o-mini	4.96±0.20	4.94±0.24	4.98±0.14	4.96±0.20	4.70±0.50	4.80±0.41	4.98±0.17	4.82±0.40
Llama-3B	4.93±0.25	4.95±0.22	4.95±0.23	4.94±0.23	4.72±0.57	4.69±0.50	4.85±0.46	4.75±0.52
Llama-3B-4bit	4.92±0.33	4.95±0.24	4.94±0.29	4.93±0.30	4.73±0.57	4.67±0.53	4.86±0.45	4.75±0.52
Llama-8B	4.94±0.26	4.96±0.19	4.96±0.21	4.95±0.23	4.82±0.46	4.80±0.43	4.90±0.39	4.83±0.45
Llama-8B-4bit	4.95±0.22	4.97±0.18	4.97±0.18	4.97±0.19	4.86±0.42	4.81±0.42	4.92±0.33	4.86±0.40
Llama-3B (Combined)	4.94±0.28	4.94±0.27	4.94±0.27	4.94±0.27	4.76±0.55	4.68±0.53	4.88±0.44	4.78±0.51
Llama-3B-4bit (Combined)	4.94±0.30	4.95±0.28	4.96±0.27	4.96±0.27	4.74±0.58	4.71±0.53	4.86±0.47	4.76±0.53
Llama-8B (Combined)	4.95±0.21	4.96±0.19	4.96±0.20	4.96±0.20	4.84±0.46	4.81±0.41	4.90±0.39	4.84±0.45
Llama-8B-4bit (Combined)	4.95±0.22	4.96±0.19	4.96±0.20	4.96±0.21	4.84±0.46	4.81±0.41	4.90±0.39	4.84±0.44

5.6 結果分析與結論

本研究通過測試集比較及 LLM 自動評估，驗證了 Llama 微調模型在交通事故對話任務中的效能。結果顯示：

綜合測試集與 LLM 自動評估結果：1) 資訊擷取能力：微調模型生成的 JSON 語義相似度達 0.995，欄位準確度超過 94%，量化模型在資源受限環境中保持穩定性，準確度達 952) 問題生成能力：平均語義相似度約 0.85，最佳高語義準確度達 0.7949，問題更精煉但語義完整性未受影響。3) **LLM** 評估結果：多任務訓練未稀釋模型品質，性能與單任務微調相當，部分指標略有提升。

微調模型在特定任務上效能卓越，但對突發場景的適應性仍不及通用模型。本研究證實了微調開源模型的高效性與成本效益，並展現其在資源受限應用場景中的潛力。

6 研究限制

儘管測試集與 LLM 評估結果表明 CCG 微調模型在語義準確性與欄位對齊方面表現出色，其效能及適用性仍受到以下限制的影響。

資料覆蓋有限：訓練資料主要來自判決書文本，細節欄位僅涵蓋約 8-10/18 欄位，影響模型在特定場景的泛化能力。代理生成資料部分降低偏差，但無法完全模擬壓力情境下的語用變異，可能導致回覆不連貫或斷裂。

功能範疇侷限：CG 系統僅限於結構化資訊蒐集，不涉及法律推理或裁量功能，需明確界定用途以避免誤用或誤解。

適用範圍侷限：模型僅適用於繁體中文及台灣法規環境，尚未驗證跨語言或異質法規體系的效能。實務部署中需考量隱私保護與法規遵循，探索技術與法律框架整合以確保合規與安全性。

7 結論

本研究提出 CCG 系統架構，整合問題生成、資訊擷取及事故重建三大模組，專注於交通事故初期的結構化資訊蒐集，爲警政初步紀錄提供技術支撐。

基於 Llama 模型的任務特化微調，系統性能接近商業模型：資訊擷取欄位準確率達 89%，生成 JSON 語義相似度 0.995；問題生成語義相似度 0.85–0.88，提問精煉效果提升 20%。多任務聯合訓練維持語義穩定性，4-bit 量化版本主要指標保持 95% 以上一致性，證實系統適合低資源環境的私有化部署。

不同於聚焦事前預測或事後分析的研究，CCG 專注於「事故後第一時間」的互動式資料收集，爲交通事故前期事實蒐集提供創新解決方案。研究表明，結構化設計與多任務微調使開源模型在專業領域逼近商業模型，展現高部署靈活性與成本效益，並爲 AI 跨領域應用奠定基礎。

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

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

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Jo-Chi Kung, Chia-Hui Chang, Huai-Hsuan Huang, and Kuo-Chun Chien. 2024. A narrative assistant for traffic accidents based on large language models (llm). In *Legal Knowledge and Information Systems*, pages 84–94. IOS Press.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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

Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Hengchang Liu. 2020. Riskoracle: A minute-level citywide traffic accident forecasting framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1258–1265.