

Design and Evaluation of a Courtroom Examination AI Simulation System with Behavioral Fidelity

具行為擬真度之法庭詰問 AI 模擬系統設計與評估

Liao, Hsien-Jyh
Independent Researcher
hjliao123@gmail.com

摘要

本研究提出一套以「行為擬真度 (Behavioral Fidelity)」為核心的法庭詰問 AI 模擬系統，並將語音互動設計納入以提升沉浸感與臨場再現；惟為確保評估之標準化與可重現性，本輪前導評估採文字筆錄。系統整合語用—心理規則與臺灣刑案卷宗，以模擬證人在詰問壓力下的語言行為；效能評估採經優化之「專家圖靈測試」框架，涵蓋四個面向：專業準確度、情境適應、人味化與邏輯一致性。作為前導研究 (pilot)，結果顯示：在相同題庫與知識依據下，客製化 GPT 相較於 GPT-Vanilla 於「情境適應」與「人味化」呈現較高評分；同一框架應用於另一主流模型 (Gemini 2.5 Flash) 時亦達可比水準，惟在本樣本規模下差異未形成一致結論。整體而言，研究提供「行為擬真度」作為評估指標之初步證據，並顯示生成式 AI 於法律訓練場域具可規模化應用之潛力；語音條件下之評估及多案件、多角色擴充留待後續研究。

Abstract

We present a courtroom cross-examination AI simulation system centered on Behavioral Fidelity, with speech interaction included as a design feature to enhance immersion. For standardization and reproducibility, the present pilot evaluation uses transcripts. The system integrates pragmatic-psychological rules with Taiwanese criminal case files to simulate witness behavior under cross-examination pressure. Using an optimized Expert Turing Test framework with four dimensions—professional accuracy, situational

adaptability, human-likeness, and logical consistency—we conduct a pilot study. Under identical prompts and knowledge sources, the customized GPT condition received higher ratings than GPT-Vanilla on adaptability and human-likeness. Applying the same framework to another mainstream model (Gemini 2.5 Flash) yielded comparable performance, while differences remain inconclusive at this sample size. Overall, the results provide preliminary evidence that Behavioral Fidelity is a feasible evaluation target and indicate the scalability of generative AI for legal training; speech-condition evaluation and multi-case, multi-role extensions are left for future work.

關鍵字：法庭詰問、行為擬真度、法庭語言學、人工智能模擬

Keywords: courtroom examination, behavioral fidelity, forensic linguistics, AI simulation

1 引言 (Introduction)

在法律實務訓練中，交互詰問是塑造訴訟攻防與臨場決策的核心能力；然而，交互詰問難以實作練習，而傳統模擬法庭也因人力與環境限制，難以長期、規模化實施，但近年生成式人工智能 (Generative AI, GenAI) 提供了新的路徑：除可低成本產生多樣情境，亦能扮演特定專業角色，支援反覆演練與即時回饋的教學流程。基此，本研究結合 GenAI、法庭真實語境及語音互動，開發一套可透過口語進行沉浸式詰問演練的法庭詰問 AI 模擬系統，並以「行為擬真度」(Behavioral Fidelity) 做為本系統的核心設計與評估理念。

然而，如何評估此類系統的有效性，仍是關鍵挑戰。現行主流評估（如 MMLU：

Hendrycks et al., 2021; HELM: Liang et al., 2022) 與傳統文本品質指標 (ROUGE: Lin, 2004; BLEU: Papineni et al., 2002) 多著眼於內容正確性與流暢度。相較之下，交互詰問的關鍵在於語用策略、權力互動與壓力下的臨場反應 (Cotterill, 2003; Coulthard & Johnson, 2017; Gudjonsson, 2003; Slater & Sanchez-Vives, 2016)。若無法重現這些行為特徵，模擬的教學價值將受限。因此，本研究將焦點前移：從「答得對」轉向「像真人一樣在法庭上作答」。

為了在可重現前提下檢驗系統效能，本輪前導評估 (pilot) 採文字筆錄作為評估材料 (語音互動僅為系統功能，非本輪評估對象)，並設計「專家圖靈測試」式的人評框架，涵蓋四個維度：專業準確度、情境適應、人味化、邏輯一致性。我們以去識別化之臺灣刑案卷宗為知識依據，結合語用—心理規則與五大題型分類，生成在不同詰問階段下可控的證人行為。

本研究的主要貢獻如下：

- (1) 提出並操作化「行為擬真度」作為評估核心：補足既有評估偏重內容正確性的不足，將臨場語用與行為一致性納入量化與質化的綜合指標。
- (2) 建構一可重現的跨域系統架構：整合語用—心理規則、五大問題分類機制與本土卷宗知識庫，以支援高擬真度的口語互動演練。
- (3) 以專家圖靈測試框架完成前導實證：在相同題庫與知識依據下，比較客製化 GPT 與 GPT-Vanilla 之差異，並提供與另一主流模型的敘述性對照；結果呈現估計導向報告 (含效果方向與不確定性)，為後續語音條件、跨案件與多角色之擴充研究奠定基準。

2 文獻回顧

本研究旨在建構一套法庭詰問 AI 模擬系統，其核心設計理念為行為擬真度 (Behavioral Fidelity)。此目標連結法庭語言學、法庭心理學與人機互動／AI 角色模擬三大脈絡，焦點從「知識答對」前移至「在壓力下像真人一樣作答」，以支援可重現且具臨場感的專業訓練。

2.1 法庭詰問的挑戰：策略、權力與心理壓力

法庭詰問並非單純事實問答，而是結合提問策略、話語權力與心理壓力的互動博奕 (Mauet, 2017; Cotterill, 2003)。律師得以透過封閉式／誘導式提問、重述與打斷等手法，塑造敘事框架與焦點 (Cotterill, 2003; Coulthard & Johnson, 2017)。證人在高壓情境下常出現遲疑、模糊化、記憶偏移等自然反應 (Gudjonsson, 2003)，這些語用與心理特徵正是法律訓練的目標能力之一。據此，一個擬真的 AI 證人，必須能在題型與情境變化間，展現權變性的語用行為，而非僅生成流暢文本。

2.2 從角色扮演到「行為擬真」：AI 模擬的設計要件

專業訓練中常見的「標準化角色」 (Barrows, 1993) 提供了可重複的演練基線；但在 AI 應用場域，僅具語言流暢不足以令人信服。過度完美與缺乏內在情感邏輯的回應會引發恐怖谷效應 (Mori, 2012)，降低沉浸與信任。人機互動研究指出，可被感知的行為真實性能顯著提升臨場感與可信度 (Slater & Sanchez-Vives, 2016)。因此，系統設計應將「恰當的不完美」納入 (如遲疑詞、修辭回溯、語速與停頓)，並透過可控的人設／語氣設定與程序性規則，使行為在不同詰問階段具一致的內在因果與邏輯邊界。

2.3 人類評估的必要性

當 AI 生成回應的目標是「行為擬真」而非僅「內容正確」，傳統自動指標 (如 ROUGE、BLEU) 對「是否像真人」的辨識力有限。對話系統評估文獻強調，人類評估仍是風格、自然度、權變性與一致性等主觀構念的黃金標準 (Celikyilmaz et al., 2020; Deriu et al., 2021)。本研究據此改造了圖靈測試思路，引入專家圖靈測試框架，並將評估構面收斂為四類：專業準確度、情境適應、人味化、邏輯一致性。此處的人評並非臨時主觀打分，而是有理論錨定且可操作化的評分規準 (見表 1)。

表 1：評估構面之理論錨定與可操作化指標

評估構面	理論鑄定	主要觀察點（可操作化）
專業準確度	忠實性 / 事實一致性	僅援引卷宗或可驗證事實，避免範圍主張
情境適應	權變性 / 對話依存	（封閉／誘導／推論）題型對應；（誘導、假設不明時）拒答行為觸發；關鍵訊息之澄清與優先順序。
人味化	自然度 / 風格	遲疑詞率與語助詞使用；停頓樣式與句長變化；受控的不完美（冗詞、回溯／自我修正）。
邏輯一致性	前後一致 / 無矛盾	跨回合自我一致；與已述事實不矛盾 (可選) NLI 旗標數以輔助標記矛盾。

註：NLI (Natural Language Inference) 可作為半自動化輔助，用於標示回答間的矛盾候-選，供評估者複核。

2.4 近期相關工作：AI 駕動之法庭模擬與辯論系統

近期開始出現若干以 GenAI 支援法學教育或辯論策略的系統，但研究目標與評估焦點與本研究不同：

- (1) 教學框架類：如 Moot MentorAI (Serra, 2024) 著重教學部署與文字論述回饋，評估主軸多在文本品質與學習成效，較少觸及「擬真行為」的量化。
- (2) 多代理對抗類：如 AgentCourt (Chen et al., 2025) 專注於對抗式演化以提升辯論策略，目標是「勝率／推理能力」而非「擬真行為」。
- (3) 原型公告類：如 Stanford CodeX × Three Crowns 公告之訓練原型，尚未公開可重現之系統設定與評估數據。

整體而言，現有研究或偏向教學流程或文字技巧，或僅瞄準策略最適化，能同時兼顧行為擬真與語音互動設計，並提供透明可重現的評估框架者仍少見。本研究即補上此一縫隙：在單一 AI 證人的詰問場景中，以行為擬真為首要目標，提出可操作的人評標準與可重現的系統實作。

2.5 小結

綜合上述脈絡，既有研究或著重教學流程與文字技巧，或以多代理對抗提升辯論策略，較少將「行為擬真」作為首要評估目標，亦

缺乏可重現之框架以量化法庭情境下的語用行為。基於此缺口，本研究提出以行為擬真度為核心的法庭詰問 AI 模擬系統，結合本土卷宗 RAG、語用—心理規則與角色／語氣設定，並採用專家圖靈測試式人評作為前導評估。需強調的是：語音互動為系統功能，本輪為確保標準化而採文字筆錄進行評估；外部效度（多案件、多角色）留待後續擴充。以下（第 3 章）據此說明系統架構與實作細節。

3 系統架構與實作

3.1 整體流程與技術核心

本系統以大型語言模型 (LLM) 為核心，整合語用—心理判斷規則、五類問題分類機制與本土刑事卷宗知識庫，形成一跨學科的智能模擬系統。為提升臨場感，系統整合了語音生成 (TTS) 與辨識 (ASR) 模組，讓使用者能以口語和 AI 證人互動。

整體互動流程如圖 3-1 所示，主要可分為以下三個階段：

- (1) 輸入階段：使用者（如學生、律師）以自然語言提問。系統首先會自動識別詰問階段（如主詰問、反詰問），並運用分類規則判定問題類型。
- (2) AI 應答階段：系統根據專屬提示詞 (Prompt) 與檢索增強生成 (RAG) 技術，調用知識庫與語用規則。AI 證人會依據問題類型、提問意圖及法庭情境，生成高度擬真的答覆。
- (3) 輸出階段：系統即時以帶有語調、停頓變化的語音呈現 AI 證人的回應，模擬真實法庭中證人緊張、遲疑或堅定的口氣，以確保學習者能沉浸於真實的互動情境。

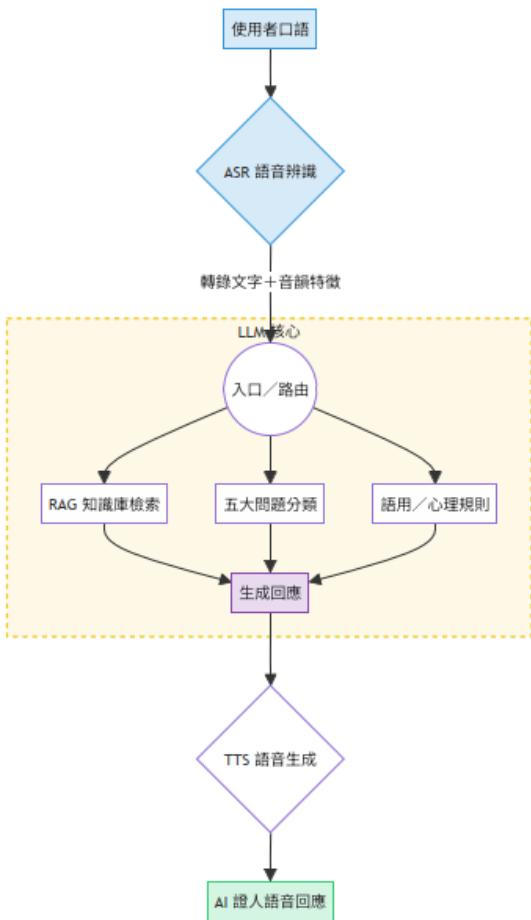


圖 1：法庭詰問 AI 模擬系統之架構與互動流程

本圖呈現使用者從發出口語提問，到接收 AI 證人語音回應的完整處理流程。使用者的語音輸入先經由 ASR 模組轉換為文字，接著進入「LLM 核心」進行處理。此核心整合了 RAG 知識庫檢索、五大問題分類與語用—心理規則三大元件，以生成兼具專業性與行為擬真度的回應文本。最後，文本再經由 TTS 模組轉換為帶有語氣變化的語音，完成一次詰問互動。

3.2 三大核心架構

為讓 AI 模擬證人，實現「行為擬真度」，本系統由三大核心架構構成：

3.2.1 語用-心理判斷規則模組

本模組結合法庭語用學與認知心理學理論 (Cotterill, 2003; Coulthard and Johnson, 2017)，使 AI 能在壓力、誘導或澄清等情境下展現遲疑、拒答或修正等行為，避免傳統對話系統

「過度流暢」「過度迎合」而失真的問題。其核心為大規模提示詞工程 (Prompt Engineering)，涵蓋角色設定、行為規則、語用風格與情境化策略，全文逾萬字：主詰問中設計為誠懇配合，反詰問則轉為謹慎閃避，並在誘導性提問下展現「拒答」或「模糊化」的自然反應。由於提示詞篇幅龐大且細節繁複，若全文刊載將嚴重影響篇幅與可讀性，因此本文僅呈現設計原則與部分範例。相較於僅關注正確性與流暢度的既有方法 (Ji et al., 2023; Belz and Kow, 2011)，本研究更強調在高壓法庭場景中重現臨場語用與行為反應。為此，本模組採取可控文本生成 (Controllable Text Generation) 的設計路徑 (Hu et al., 2017)，將語言學與心理學理論轉化為結構化提示詞的 meta-rules。例如，「反詰問」標籤會觸發簡短句式與不確定詞彙 (如「也許」「不一定」)，並避免主動補充，從而確保生成結果符合專業語境的「行為擬真度」。

3.2.2 五大問題分類機制

為充分重現真實法庭詰問現場，系統將所有提問自動歸類為以下五大類型，五大問題分類設計，主要參考 Cotterill (2003) 對於法庭詰問話語結構與敘事權力的分類分析，同時融入本研究者在台灣模擬法庭教學與檢察官公訴實務經驗，強化在本土法庭詰問語境下的實用性與臨場擬真度。每類問題均訂有專屬語用規則，並可依主詰問/反詰問進一步細緻調整，問題共分為五類，詳細範例放在附錄 A：

(1) 背景問題 (Background Questions)

此類問題重在建立證人專業信任，故要求 AI 證人能以冷靜、自信、條理清晰之口吻，簡要說明自身資格與經驗

(2) 證據問題 (Evidence Questions)

此類問題是法庭詰問的核心，發問者希望證人基於本案事實或證據回答問題，因此，此階段的應答規則以抑制模型產生學界所稱的「幻覺」 (Hallucination) 為首要目標 (Ji et al., 2023)，嚴格限制其回應不得脫離卷證範圍或憑空捏造。

(3) 推論或專業判斷問題 (Inference Questions)

此類問題在專家證人的詰問中非常常見，發問者希望證人做出推論，故常以：「你認為..」、「依照你的經驗…」做為問題開頭。AI 證人對此類問題的回答必須基於專業及本案證據回答。

(4) 澄清問題 (Clarification Questions)

此類問題在詰問中亦很常見。在詰問中，發問者會對證人已經回答過的問題繼續追問；或針對關鍵事實希望證人補充；或希望專家證人對專業術語以白話說明或舉例解釋。

(5) 題組問題 (Grouped Questions)

此類問題通常包含許多子題，發問者是以題組方式進行詰問，此時，證人不一定要每一題都詳細回答，但必須維持邏輯一致性 (Self-Consistency)，且必須在後續回答中適時補充前一問題的細節，以建構完整的證詞敘事。

3.2.3 本土卷宗知識庫與 RAG 檢索技術應用

本系統的知識核心來自去識別化與結構化處理後的臺灣真實刑案卷宗，目前涵蓋 2 件案件與 4 個角色（被告、目擊證人、被害人、專家證人）。在技術上，我們利用 OpenAI GPTs (OpenAI, 2023) 的知識庫檢索功能，將卷宗切分為片段並轉換為向量嵌入，以便檢索。使用者提問後，系統會檢索最相關片段並輸入大型語言模型生成回應，屬於典型的檢索增強生成 (Retrieval-Augmented Generation, RAG) 範式 (Lewis et al., 2020)，能降低「幻覺」(Ji et al., 2023) 並確保回答有所依據。

4 實驗設計與成果分析

本章透過專為法律專業優化的「專家圖靈測試」，系統性評估本模擬系統之行為擬真度。該測試借鑒圖靈「模仿遊戲」(Turing, 1950)，但在評審組成、評量指標與評分方式上均針對法律專業場域進行調整。由於行為擬真度是一個多維概念，本研究將其拆解為四個核心面向，以全面檢視 AI 證人的表現：

(1) 專業內容準確度 (Professional Content Accuracy)：以事實一致性為基礎，確保 AI 回答忠於卷宗事實 (Ji et al., 2023)。

- (2) 情境適應 (Situational Adaptability)：指 AI 能否根據上下文與詰問策略，展現如真人般的靈活反應 (Deriu et al., 2021)。
- (3) 人味化 (Human-likeness)：評估 AI 是否能模擬真人的語言特徵與「恰當的不完美」 (Celikyilmaz et al., 2020)。
- (4) 邏輯一致性 (Logical Consistency)：要求 AI 在長詰問中，維持證詞前後一致 (Dziri et al., 2022)。

透過對這四個面向的綜合評鑑，本研究希望突破傳統僅重視流暢度的評估框架，推動 AI 評估從「能答」到「像真人在答」的典範轉型。為此，我們邀集具法律專業與一般背景的評審，以盲評方式進行檢驗。

4.1 實驗目的

基於前述四個評估面向，本研究的實驗目的如下：

- (1) 檢驗 AI 證人之行為擬真度：評估系統在四大面向上的綜合表現。
- (2) 驗證本研究框架的有效性：透過比較「客製化 GPT」與「GPT-Vanilla」基線，驗證本研究核心架構的貢獻。
- (3) 評估本研究框架的穩健性：將同一框架應用於不同 LLM (GPT-4.1 與 Gemini 2.5 Flash)，評估其跨模型的潛力。

4.2 實驗設計與資料來源

為確保實驗的嚴謹與可重現性，本前導研究採取下列設計：

- (1) 案件選擇與角色設定：本研究選用臺灣新北地方法院「111 年第 3 場國民法官模擬法庭」之「李家銘殺人案」卷宗為知識依據，設定 AI 證人為該案法醫鑑定人（化名「張開平」）。需說明的是，作為前導研究，本次評估僅聚焦於單一案件，其結論的通用性待後續研究擴充。
- (2) 實驗材料生成：建立四組筆錄作為比較對象。(a) 真人筆錄：由專業書記官紀錄的模擬法庭問答，作為黃金標準 (Gold Standard, GS) (Lin, 2004)；(b) 基於 GPT-4.1 的客製化系統筆錄（後文簡稱『客製化 GPT』）；(c) 套用相同框架的客製化 Gemini2.5 Flash 筆錄（後文簡稱『客製化 Gemini』）。(b)(c)兩者是完全相同的資料庫與提示詞條件；另加入(d) GPT-Vanilla (同一 GPT-4.1、但提示詞中僅僅提供基礎的角色設定 (『你是一位名叫張開平的

法醫』) 與 RAG 檢索的卷宗內容，而未加入任何關於語氣、風格、應對策略的引導。) 作為香草基準。三組 AI 筆錄都是回答與真人筆錄相同的 50 題詰問(涵蓋主詰問與反詰問)。雖本系統具備完整語音功能，但為確保評估的標準化與可重現性，本次評估全程採用書面文字筆錄進行。

- (3) 評審組成與盲評程序：邀集五位評審 ($n=5$)，其中三位具法庭經驗的法律專家，兩位為一般大學畢業者。四組筆錄以隨機順序呈現，並移除所有可能透露來源的線索(如模型名稱、格式化語言)，避免風格偏誤。評審需先辨識出真人筆錄，並將其設定為本次實驗的 GS (各指標自動計為 5.0 分)，以確立基準分數。此設計並非旨在衡量 AI 與真人間的差異，而是確立一個明確且一致的基準分數，作為 AI 系統擬真度比較的參照點。
- (4) 評分方式：針對三組 AI 筆錄，評審使用五點 Likert 量表 (1 = 完全不像真人，5 = 幾可亂真)，並依四個面向——專業內容準確度、情境適應/臨場應變、人味化及邏輯一致性——進行評分，同時並提供質性回饋。

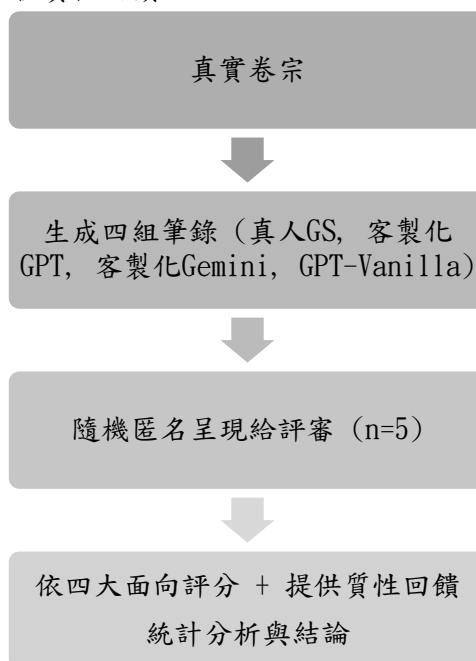


圖 2：專家圖靈測試之評估流程圖

4.3 行為擬真度評量與成果分析

本節呈現專家圖靈測試的量化數據與質性回饋分析。為確保評估的客觀性，我們進行了以下統計檢驗：

- (1) 跨評審一致性：採用 Kendall's W 檢定來衡量五位評審之間的一致性。
- (2) 模型間比較：採用 Wilcoxon 符號等級檢定 ($n=5, \alpha=.05$) 來比較「客製化 GPT」與「GPT-Vanilla」基線之間的差異。

整體擬真度總評分與結論

根據評審的綜合評分，採用本研究行為規訓框架的 AI 系統，其擬真度顯著優於未經規訓的基線。其中，「客製化 GPT」的總平均分數為 4.55 分(見表 2)，表現尤其出色。

「客製化 Gemini」的表現同樣穩健，總平均分為 4.30 分。

詳細的分項評分結果如表 2 所示。從表中可見，本研究的框架在「情境適應」與「人味化」兩個維度上帶來了最顯著的提升。「客製化 GPT」在這兩項的得分(均為 4.6 分)顯著高於「GPT-Vanilla」基線(3.0 及 2.9 分)，此差異達到了統計上的顯著性 ($p < .05$)，與表格中的星號 (*) 標記一致。

行為細節分項評分與定性回饋

為進一步解析差異成因，評審對各項細節指標進行了 1-5 分的李克特量表評分。結果如下表 2 所示：

表 2：專家圖靈測試評分表。

(評分由五位評審進行，採五點李克特量表)

評量面向	真人筆錄 ($M \pm SD$)	客製化 GPT ($M \pm SD$)	客製化 Gemini ($M \pm SD$)	GPT-Vanilla ($M \pm SD$)
1.內容準確度	5.0 ± 0.0	4.4 ± 0.3	4.4 ± 0.4	4.1 ± 0.4
2.情境適應	5.0 ± 0.0	4.6 ± 0.3 *	4.2 ± 0.5	3.0 ± 0.5
3.人味化	5.0 ± 0.0	4.6 ± 0.2 *	4.0 ± 0.6	2.9 ± 0.2
4.邏輯一致	5 ± 0.0	4.6 ± 0.3	4.6 ± 0.2	4.1 ± 0.3
總平均	5.0	4.55	4.30	3.53

註：統計分析以 Python 3.10/R 4.3 進行，可重現。數值為 5 位評審(3 位法律專家，2 位一般參與者) Likert 評分之平均 \pm 標準差；總

平均為四面向等權平均（不另計 SD）。跨評審一致性以 Kendall's W 驗證。* 表示客製化 GPT 系統與 GPT-Vanilla 基線在該維度上的差異，經 Wilcoxon 符號等級檢定達到統計顯著性 ($p < .05$)。本研究的重點比較在於驗證客製化系統相較於香草基線的改進幅度，詳細的效應量與統計值於 4.4 節中闡述。

評審定性回饋與分析

在質性回饋中，評審普遍認為兩個模型在專業內容準確度與邏輯一致性上均表現出色，GPT-Vanilla 表現一般。然而，細微差異體現在：客製化 Gemini 在提供「2.5 公分」等具體數據上更接近真人筆錄的細緻度；而本研究的客製化 GPT 則在情境適應與人味化上顯著勝出，這也與其和香草基準之間最大的分數差距 (+1.60 及 +1.70) 相吻合。評審特別指出，客製化 GPT 能以「不一定」、「這其實跟角度有關」等更靈活、口語化的方式應對誘導性詰問，並透過大量語助詞與自然的停頓，成功模擬了真人在壓力下的語言行為，甚至在回答中出現了輕微的細節重複或回溯，這些「不完美」反而讓它聽起來像一個活生生、有血有肉的人在發言。更貼近本研究「行為擬真度」的核心目標。不過，有評審補充，客製化 Gemini 在面對帶有錯誤資訊的誘導性提問時，能更直接且準確地予以糾正，展現了另一種形式的應變能力。

4.4 綜合討論與研究限制

綜合量化數據與質性回饋，本研究的成果初步表明，經由專業知識與語用規則客製化的 AI 證人，已能在行為擬真度上展現潛力。如前節所述，本研究框架在「情境適應」與「人味化」兩個維度上帶來了最顯著的提升。

評審的質性回饋為此提供了更深入的解釋。有評審特別指出，客製化 GPT 能以「不一定」、「這其實跟角度有關」等更靈活、口語化的方式應對誘導性詰問，並透過大量語助詞與自然的停頓，成功模擬了真人在壓力下的語言行為。這些「不完美」的細節反而讓它聽起來更像一個真實的人，這與本研究追求「行為擬真度」的核心目標相符。

然而，本研究亦有其限制：

- (1) 案例範圍：實驗僅基於單一刑案卷宗，其結論在跨案件類型上的通用性仍待後續驗證。

- (2) 評估媒介：雖語音互動是本系統的核心創新，但本次為求標準化與可重現性，評估乃基於書面筆錄進行，未能涵蓋語氣、語速等重要的非語言線索。
- (3) 樣本規模：本研究的評審團人數 ($n=5$) 有限，因此其量化結果應被視為探索性的初步發現，有待未來更大規模的評估來證實。

5. 結論與未來展望

5.1 研究結論

本研究的實驗結果表明，所提出的「行為擬真度」框架在法律教育輔助工具的開發上具有潛力。作為一項前導研究 (pilot study)，其成果為後續更大規模的研究奠定了基礎。與過往僅依賴文字互動的模擬系統相比，本研究首次將語音模擬納入本土法庭詰問練習，讓使用者得以進行「口語提問—語音回應」的雙向演練，此特點大幅提升了訓練的沉浸感與真實感。

本研究的主要貢獻可歸納如下：

- (1) 提出並操作化「行為擬真度」作為評估核心：補足了既有 AI 評估偏重內容正確性的不足，將臨場語用與行為一致性納入評估框架。
- (2) 建構一可重現的跨域系統架構：整合了語用—心理規則、五大問題分類機制與本土卷宗知識庫，以支援高擬真度的口語互動演練。
- (3) 以專家圖靈測試框架完成前導實證：透過比較客製化 GPT 與 GPT-Vanilla 基線，驗證了本研究框架對提升擬真度的有效性；並提供了與另一主流模型的敘述性對照，初步展現了方法的潛力。

誠然，如 4.4 節所詳述，本研究作為前導研究，在案例範圍、評估媒介與評審規模上仍有其限制。然而，研究成果仍清晰地驗證了「行為擬真度」作為評估指標的可行性，並展現了生成式 AI 作為可規模化法律培訓工具的巨大潛力。

5.2 未來展望

奠基於本次研究的成果與前述限制，未來的研究可朝以下方向深化：

- (1) 深化行為擬真度 (Deepening Behavioral Fidelity)：針對 AI 在壓力下的微行為（如自然遲疑、語助詞）進行更深入的建模，以克服「數位恐怖谷」效應。
- (2) 整合多模態評估 (Multimodal Evaluation)：在未來研究中，將評估從文字筆錄擴展至包含語氣、語速等完整多模態互動，以更全面地檢驗擬真度。
- (3) 擴充案例與驗證通用性 (Expanding Case Base and Validating Generalizability)：持續擴大本土案例的類型與數量，並招募更多元的評審樣本，以提升系統的通用性與研究結論的外部效度。
- (4) 開發自動化評估指標 (Developing Automated Metrics)：探索半自動化指標（如以 NLI 模型檢測邏輯一致性）的可行性，以輔助專家評分，提升評估效率與規模。

參考文獻(References)

- Barrows, Howard S. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine*, 68(6):443–451. DOI: 10.1097/00001888-199306000-00002
- Belz, Anja and Ehud Kow. 2011. Discrete vs. continuous rating scales for language evaluation: Revisiting the human evaluation of NLG systems. In Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 98–102. Association for Computational Linguistics. ACL Anthology ID: P11-2040
- Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799. <https://arxiv.org/abs/2006.14799>
- Chen, Zihan, Wei Wang, Fan-Keng Sun, Hong-Han Shuai, and Wen-Chih Peng. 2025. AgentCourt: A multi-agent simulation framework for court debates. In Proceedings of the 39th AAAI Conference on Artificial Intelligence.
- Cotterill, Janet. 2003. *Language and Power in Court: A Linguistic Analysis of the O.J. Simpson Trial*. Palgrave Macmillan.
- Coulthard, Malcolm and Alison Johnson. 2017. *An Introduction to Forensic Linguistics: Language in Evidence* (2nd ed.). Routledge.
- Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Gorka Azkune. 2021. A survey on the evaluation of dialog systems. arXiv preprint arXiv:2106.01254. <https://arxiv.org/abs/2106.01254>
- Dziri, Nouha, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating the factual consistency of large language models through natural language inference. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9716–9741. Association for Computational Linguistics. DOI: 10.18653/v1/2022.emnlp-main.663 (ACL Anthology ID: 2022.emnlp-main.663)
- Gudjonsson, Gisli H. 2003. *The Psychology of Interrogations and Confessions: A Handbook*. Wiley.
- Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2021. Measuring massive multitask language understanding. In Proceedings of the International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/2009.03300>
- Hu, Zhiting, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In Proceedings of the 34th International Conference on Machine Learning (ICML), pages 1587–1596. DOI: 10.48550/arXiv.1703.00955
- Ji, Zhijing, Nanyun Lee, et al. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38. DOI: 10.1145/3571730
- Lewis, Patrick, Ethan Perez, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pages 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Liang, Percy, Rishi Bommasani, et al. 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110. <https://arxiv.org/abs/2211.09110>
- Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the ACL-04 Workshop on Text Summarization Branches Out, pages 74–81. Association for Computational Linguistics. ACL Anthology ID: W04-1013
- Mauet, Thomas A. 2017. *Trial Techniques and Trials* (10th ed.). Wolters Kluwer.
- Mori, Masahiro. 2012. The uncanny valley. *IEEE Robotics & Automation Magazine*, 19(2):98–100. DOI: 10.1109/MRA.2012.2192811

OpenAI. 2023. GPTs. <https://platform.openai.com/docs/gpts>. Accessed: October 16, 2025.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02), pages 311–318. DOI: 10.3115/1073083.1073135

Serra, Kevin D. 2024. MootMentorAI: A generative AI-powered platform for law school moot court training. arXiv preprint arXiv:2403.09033. <https://arxiv.org/abs/2403.09033>

Slater, Mel and Maria V. Sanchez-Vives. 2016. Enhancing our lives with immersive virtual reality. Frontiers in Robotics and AI, 3:74. DOI: 10.3389/frobt.2016.00074

Stanford CodeX and Three Crowns LLP. 2025. CodeX and Three Crowns LLP announce AI cross-examination tool. <https://law.stanford.edu/codex-the-stanford-center-for-legal-informatics/codex-and-three-crowns-llp-announce-ai-cross-examination-tool/>. Accessed: October 16, 2025.

Turing, Alan. 1950. Computing machinery and intelligence. Mind, 59(236):433–460. DOI: 10.1093/mind/LIX.236.433

附錄 A：五類問題範例

(1) 背景問題 (Background Questions)

例：

問：請問你擔任法醫多久？

AI 答：我在法醫所任職 20 年，主要負責

(2) 證據問題 (Evidence Questions)

例：

問：請解釋死者死因

AI 答：請看解剖報告第 X 頁，死者死因是

(3) 推論或專業判斷問題 (Inference Questions)

例：

問：一般人肋骨間距大概是多少？

AI 答：一般來說，肋骨的間距大約是 2 到 3 公分，但還是因人而異。

(4) 澄清問題 (Clarification Questions)

例：

問：請說明什麼是單一銳器穿刺傷？

AI 答：就是由刀子等具穿刺功能的銳器刺入人體所造成的創

(5) 題組問題 (Grouped Questions)

例：

問 1：死者死因是什麼？

AI 答：是單一銳器穿刺傷造成的。

問 2：什麼是單一銳器穿刺傷？

AI 答：就是由刀子等具穿刺功能的銳器刺入人體所造成的創傷，我記得在解剖大體的時候，我們發現死者的右側胸部有一個單一的銳器傷。這個傷口是從右下往左上的方向，直接刺穿了橫膈膜。這就是造成心包膜囊裡面大量出血的原因。