

審計機關導入及運用生成式人工智慧 (AI) 精進審計業務之研究

委託研究報告

計畫主持人：葉向原 教授
計畫協同主持人：吳政隆 副教授

執行單位：東吳大學巨量資料管理學院資料科學系

中華民國 113 年 8 月

目錄

壹、研究主旨與動機.....	1
貳、研究背景分析.....	4
一、大型語言模型發展.....	4
二、大型語言模型的應用.....	13
三、大型語言模型的技術.....	17
四、應用大型語言模型案例.....	22
五、繁體中文的大型語言模型.....	29
六、大型語言模型審計應用的期待與潛在風險.....	32
參、研究方法與步驟.....	35
一、外部資料庫蒐集、資料整理及文本切割.....	36
二、向量特徵萃取及相似度檢索.....	38
三、查詢指令的設計.....	40
四、基於常見的預訓練模型並微調模型.....	41
五、預訓練模型微調結果評估.....	43
六、人類標註標準答案.....	46
七、基於人類反饋的強化學習.....	46
八、實驗設計.....	50
肆、審計機關導入及運用生成式人工智慧技術策略之研析.....	55
伍、結論與建議.....	59
參考資料.....	67
附件一：人工評分表.....	71
附件二：政府審計運用生成式 AI 辦理審計業務注意事項.....	83

圖目錄

圖 1 DocLLM 運作流程步驟 來源：DocLLM: A LAYOUT-AWARE GENERATIVE LANGUAGE MODEL FOR MULTIMODAL DOCUMENT UNDERSTANDING	17
圖 2 強化學習從人類反饋 (RLHF) 運作流程步驟.....	18
圖 3 LORA 運作流程步驟[19].....	20
圖 4 研究流程與步驟	36
圖 5 原始資料集樣貌.....	37
圖 6 文本切割	38
圖 7 向量特徵萃取.....	39
圖 8 向量特徵存取於資料庫	39
圖 9 大型語言問答以及查詢文件資料庫之問答.....	40
圖 10 整理過後的資料集樣貌.....	41
圖 11 所使用到的所有台灣本土大型語言模型	42
圖 12 台灣本土大型語言模型之問答微調.....	43
圖 13 備詢之問答範例	51
圖 14 大型語言模型的 RAG 備詢之問答結果.....	52
圖 15 大型語言 GPT 模型的 RAG 備詢之問答結果.....	52

表目錄

表 1 生成式大型語言模型比較	12
表 2 使用大型語言模型的案例整理	28
表 3 繁體中文之大型語言模型整理	31
表 4 大型語言模型審計應用的考量與潛在風險.....	33
表 5 模型微調前成效的評估	45
表 6 模型微調前成效的評估	46
表 7 備詢之問答模型微調前成效的評估	52
表 8 備詢之問答模型微調後成效的評估	52
表 9 備詢之問答模型微調前後的人工評分	54

壹、研究主旨與動機

審計權作為憲法所規定的監察權之一，在審計法的明文規定下，我國的審計機關擁有多項職權，包括「監督預算之執行」、「審核財務收支，審定決算」、「稽察財物及財政上之不法或不忠於職務之行為」、「考核財務效能」、「核定財務責任」等。當前，智慧政府的審計機關正在面臨數位轉型，以協助各機關辨識未來趨勢，提醒留意即將發生的挑戰，對行政部門提出預警性意見，並協助建立完備的風險管理制度，以發揮政府審計的積極功能。根據審計法及政府良善治理觀點，政府審計的三大核心功能為監督、洞察及前瞻。在智慧政府發展策略下，政府審計面臨的挑戰主要包括以下幾點：

- 一、改善資訊處理效率：審計業務需要處理大量的資訊和檔案，包括預算執行、財務收支、決算等，這些資訊龐大而複雜。傳統的手動檢索和分析方式效率相對較低，且需耗費大量人力與成本。
- 二、專業知識要求高：審計人員需要具備豐富的專業知識，以確保對資訊的準確理解和評估。面對龐大的文獻和資訊，審計人員需要能夠快速擷取相關所需資訊。
- 三、數位治理需求增加：政府數位能力漸趨成熟，因應政府數位治理需求及財政環境變化，監督預算之執行，進而洞察及前瞻潛在風險事項，提供審計監督和預警服務日益重要。

隨著網際網路的蓬勃發展及政府資訊化業務的擴增，政府審計工作所需處理的資料來源和數量大幅增長。除了審計部資訊系統預處理的結構化資料外，各項資訊系統已累積了豐富的審核作業資訊，包括新聞媒體和被審核機關的業務資料等，這些資料型態涵蓋數據、文字、影像和聲音等。審計工作具有其知識性和技術性，現今生成式 AI 的崛起和蓬勃發展，為審計工作帶來了新的機遇。透過程式分析解釋大量數據，並運用生成式 AI

進行解讀，可以達成特定目的，例如提供審計人員知識蒐集、分享和經驗傳承的資料服務，支援審計業務改革推動。大型語言模型結合生成式檢索增強技術的應用，可以有效提升審計工作的效能與準確性，同時深化對政府財務資訊的審計分析。這一研究方向在專業性質上不僅整合先進技術以強化審計工作的能力，還提高對預算執行結果的獨立驗證程度，減少主觀分析的影響，從而確保審計的客觀性及可靠性。此外，在績效審計內涵的具體應用中，這種整合技術有望推動審計機關更全面地評估施政計畫、事業計畫或營業計畫的成效，進而提供更具洞察力的建議，自動化監測與風險評估支援功能，可提高及時辨識潛在風險和新興挑戰的能力，協助審計人員提供預警服務。

針對上述審計業務的需求，本研究以大型語言模型和生成式檢索增強技術為基礎，探討其在審計領域的應用情形，旨在加強對審計檔案的自動化處理，提高審計人員的處理效率。大型語言模型在理解文本內容方面具有優勢，其學習能力能夠迅速適應不斷變化的法規和制度。審計流程通常需要進行多重交叉比對和查找相關資訊等步驟。生成式檢索增強技術可以透過擴充資料庫，擴大審計機關的資料檢索範圍，提供有依據的相關資訊，實現便捷檢索效果。這些技術可以輔助解釋複雜的審計相關內容，提供更準確的資訊。大型語言模型的自然語言生成能力有助於產生專業的審計意見，使得審計專業人員能夠更有效地溝通風險評估結果，並採取適當的措施。這些擴增的檢索結果能夠協助審計人員更迅速、有效地檢視大量相關檔案，進行更深入的分析。這不僅能提高審計工作的效能，還能確保審計結果的準確性和可靠性，從而有助於審計機關充分發揮其監督、洞察和前瞻的角色，提升國家的良善治理。

審計部為全面評估生成式人工智慧於審計實務之應用，特委託東吳大學巨量資料管理學院資料科學系執行「審計機關導入及運用生成式人工智

慧精進審計業務之研究」之計畫，本研究重點目標如下：

- 一、蒐集文獻資料來瞭解審計機關及國內會計師事務所運用生成式 AI 的情形，以及如何使用生成式 AI，包括具體應用案例、技術選擇、實施效果等，並評估這些實務對審計機關導入生成式 AI 的影響。
- 二、探討審計機關實務操作中運用生成式 AI 的具體情況，研究提出使用生成式 AI 的模式，期提升審計人員提問技巧及探索審計業務知識管理之設計發展生成式 AI 應用於審計的架構。
- 三、提升審計人員提問技巧，俾應用生成式 AI 於查核規劃、書面審核、就地抽查及專案調查等審計工作，以提高審計效能和準確性。

本研究提出「審計部運用生成式 AI 辦理審計業務注意事項」，從審計相關的資料進行整合、處理及實際應用，研究結果具備以下重要的優勢：

- 一、使用生成式 AI 進行審計業務數據品質管理，可確保所使用的數據資料品質良好，並定期檢查及更新數據源，以確保生成式 AI 的分析基礎是可靠的。
- 二、模型訓練與驗證適當的模型訓練對於生成式 AI 的效能至關重要，確保使用的模型具有良好的準確性和泛化能力。同時，進行驗證程序以確認模型在各種情境下的穩健性。
- 三、透明度和可解釋性確保生成式 AI 的運作過程是透明且可解釋的，以讓審計人員理解模型的決策基礎，有助於建立信任，提高生成式 AI 在審計業務中的可接受度。
- 四、培訓與知識分享提供審計人員相關的生成式 AI 培訓，強化其對於技術的理解和應用，同時，建立知識分享機制，促進團隊間的經驗交流。透過上述研究和建議，可以在審計業務中更有效地導入和運用生成式 AI 智慧技術，提升審計工作的效能和準確性。

貳、研究背景分析

一、大型語言模型發展

機器學習 (Machine Learning, ML) 是人工智慧的重要分支，這項技術使電腦系統能夠從資料中學習，並將所學應用於新的資料上，從而在無需明確編程指令的情況下執行任務 (Mitchell, 1997; Samuel, 1959)。根據模型的訓練方式和應用場景，機器學習演算法通常分為監督式學習 (Supervised Learning)、非監督式學習 (Unsupervised Learning) 和強化式學習 (Reinforcement Learning)。監督式學習依賴人工標註的資料集，這些資料集包含輸入和對應的目標輸出，常應用於分類 (Classification) 和迴歸 (Regression) 問題，常見的演算法包括支持向量機 (Support Vector Machines, SVM) 和決策樹 (Decision Trees)。非監督式學習則利用大量未標註的資料來探索資料中的內在結構或模式，通常應用於分群 (Clustering) 和降維 (Dimensionality Reduction) 等任務，代表性的演算法有 K-均值分群 (K-means Clustering) 和主成分分析 (Principal Component Analysis, PCA)。強化式學習則在一個能與模型互動的環境中進行，透過獲取獎勵和懲罰訊息，讓模型學習如何最大化累積獎勵，進而制定最佳行為策略。Q-Learning 是強化式學習的一種代表性演算法，廣泛應用於需要決策和控制的任務中。此外，針對現實中經常缺乏充分標註資料或資料結構過於複雜的情況，衍生出半監督式學習 (Semi-Supervised Learning) 和自監督式學習 (Self-Supervised Learning) 方法。半監督式學習結合了監督式和非監督式學習，利用少量標註資料與大量未標註資料進行訓練，特別適用於標註成本高昂的情境。自監督式學習則介於監督式與非監督式學習之間，其訓練過程所用的資料標註並非人工添加，而是從資料本身推導出輸入與輸出的配對，旨在學習資料的內在結構和表徵 (representation)。例如，遮罩語言模型 (Masked Language Model, MLM)

技術會在訓練過程中隨機遮罩句子中的某些詞彙，並要求模型預測這些被遮罩的詞，從而理解句子中的上下文關係。

自然語言處理（Natural Language Processing, NLP）是一個跨領域的研究領域，結合了電腦科學、人工智慧與語言學，旨在讓電腦能夠理解、處理及生成自然語言，從而促進人機互動的自然性與流暢性。NLP 的目標在於模擬或增強人類語言理解能力，從而支持多樣化的應用，包括語音助理、自動翻譯、情感分析、對話生成及自動摘要等，使電腦能透過語言與人類進行交流與協作。這項技術的核心在於語言結構和語意的解析及處理，並且藉由大數據和機器學習技術的發展，NLP 在近年取得了突破性進展，為多個行業的自動化與智慧化提供了更多可能。

NLP 涵蓋多個層次的語言分析，包括以下幾個關鍵領域：

- （一）形態學分析（morphological analysis）：著重於詞語的內部結構和形態變化，例如詞根、詞綴。
- （二）句法分析（syntactic analysis）：通過解析樹（parse tree）來分析句子的語法結構，識別名詞片語、動詞片語等不同語法成分間的關係。
- （三）語義分析（semantic analysis）：致力於理解句子的語義，涵蓋詞義消歧（word sense disambiguation），即在多義詞的語境中選擇正確的詞義；以及詞彙語義學（lexical semantics），研究同義詞、反義詞及其他語義關係。
- （四）篇章分析（discourse analysis）：探討句子與段落間的邏輯關係，確保文本的結構與連貫性。常見的任務包括指稱解析（reference resolution）、篇章結構分析（discourse structure analysis）與對話解釋（dialogue interpretation）。
- （五）語用分析（pragmatic analysis）：聚焦於文本中的隱含資訊，分析言語行為背後的意圖與語境，處理語言在實際使用中的情況。

隨著人工智慧技術的革命性進步，語言模型的興起為上述問題帶來了嶄新的解決方案。透過深度學習技術，語言模型在自然語言處理領域展現出卓越的能力，突破過去僅依賴關鍵詞精確比對的限制，改以語義理解來處理複雜的查詢。這些模型利用上下文（context）來解讀詞彙的多重含義，大幅提升了對文本理解的全面性，使得文本檢索變得更加靈活。即使查詢條件不夠精確，系統仍能提供相關的查詢結果。因此，使用者不再需要費心設置精確的檢索條件，而是可以以自然的口語方式提問，由系統返回相關的查詢結果或明確的答案字串，從而實現更為友善的問答系統。

目前，自然語言處理技術主要基於深度學習方法，這些技術透過大規模資料來訓練模型，自動學習語言中的規則與模式，而無需依賴人工編寫的規則。隨著深度學習技術的不斷進步，語言模型的應用範圍持續拓展，並在多種語言任務中展現卓越性能，已成為現今自然語言處理技術的核心。

人類對知識的渴求從未停止，但資訊與知識之間的差異始終存在。作為資訊檢索（Information Retrieval, IR）技術的一環，問答系統（Question Answering Systems, QAS）旨在縮短使用者所擁有的資訊與實際需求知識之間的差距。在網路搜尋中，使用者通常更關心特定問題的答案，而非瀏覽大量搜尋結果中的各個網頁內容。問答系統的核心目標就是從大量資訊中提取精確答案，直接滿足使用者的需求。然而，自然語言的多義性、模糊性，以及問題的結構與複雜性，導致了自然語言與機器語言之間的非對稱性（asymmetry）。這些挑戰不僅影響問答系統理解問題的能力，也妨礙其檢索正確答案的能力，是現代問答系統研究的重要課題。

然而，語言模型也存在產生人工智慧幻覺（Hallucinations）的風險。這種現象是指由於資料收集和訓練過程中的問題，模型生成了與使用者輸入資訊不符、與先前生成內容矛盾，或不符合客觀事實的資訊。這些錯誤資訊的產生，對問答系統的準確性和可靠性構成挑戰。

自然語言處理領域中的一個關鍵分支是語言模型，其核心目標在於對人類語言的結構與模式進行建模，以預測詞序列中的下一個詞或填補語句中的缺失詞元（token）。以下對這些階段進行詳細說明：

（一）統計語言模型（Statistical Language Model, SLM）

統計語言模型於 1990 年代興起，基於馬可夫假設（Markov Assumption）來預測下一個詞的出現機率，主要利用若干個連續的上下文詞語來建模。具有固定上下文長度 n 的統計語言模型通常被稱為 N 元（ n -gram）模型，廣泛應用於資訊檢索、文本分類、語音辨識等領域。然而，隨著 n 值的增加，模型需要估計的轉移機率數量也會指數級增長。

（二）神經語言模型（Neural Language Model, NLM）

引入了分散式詞表示（Distributed Word Representation），也被稱為「詞嵌入」（Word Embeddings）。這種方法使用低維稠密向量來表示詞彙，能捕捉豐富的隱含語義特徵，從而有效克服統計模型中的資料稀疏問題，為後續的文本處理任務提供支持。

（三）預訓練語言模型（Pre-trained Language Model, PLM）

透過自注意力機制（Self-Attention）建模長序列中的關係，能有效捕捉長距離依賴性，並支援平行處理輸入資料，顯著提高了訓練速度和模型性能。模型確立了「預訓練—微調」的學習範式，即先利用大規模無標註文本訓練模型，然後在特定任務上進行微調。

在這個領域的發展過程中，Google 於 2017 年提出了 Transformer 模型架構。該架構透過自注意力機制，可以有效地捕捉序列中不同位置的依賴關係，相較於傳統的循環神經網絡（RNN）和長短期記憶網絡（LSTM）等結構，Transformer 在訓練和推理時對長序列數據的處理更加高效。

隨後，Google 於 2018 年推出基於 Transformer 的語言模型 BERT（Bidirectional Encoder Representations from Transformers）。BERT 透過無

監督方式處理大量未標記文本，結合多層自注意力機制和兩階段的遷移學習，因為是雙向的訓練方式，能更全面地理解語言上下文關係。BERT 的創新方法為後續大型語言模型的開發奠定了基礎，並推動了語言模型技術的不斷進步。

(四) 大型語言模型 (Large Language Model, LLM)

隨著數據量的持續增長、深度學習演算法的進步以及硬體計算資源的提升，我們見證了語言技術從早期的統計語言模型逐步演進至當前功能強大的大型語言模型。隨著模型參數量和資料規模的增加，這些模型在各類下游任務中的性能顯著提升，並且展現出令人驚訝的湧現能力。這些模型能夠高效處理並理解大量自然語言數據，在文本生成、問答系統、機器翻譯以及段落摘要等多種語言任務中表現卓越。

其中，擁有 1750 億參數的 GPT-3 具備上下文學習能力，即使只提供少量樣本數據依然能夠出色地完成各類任務，展現了優異的對話能力，引發業界對人工智慧技術的高度關注。OpenAI 近一步推出了基於大型語言模型的 ChatGPT (Chat Generative Pre-trained Transformer)，該模型採用 GPT (Generative Pre-trained Transformer) 架構，包含多達 1750 億個參數，訓練所使用的文字資料量達到 45 TB。通過大量的自然語言資料預訓練，能夠理解和生成自然流暢的對話，其主要功能涵蓋問答、文本生成及程式碼撰寫等。

近年來，大型語言模型技術蓬勃發展，吸引了包括 OpenAI、Meta AI、Mistral AI 在內的眾多科技公司積極投入研發。隨著技術的進步，各類語言模型相繼問世，並在不同領域展現出卓越的應用價值。以下將介紹幾款具有代表性且影響力顯著的模型，這些模型廣泛應用於多種場景與行業，彙整如表 1 所示：

1. GPT-4

GPT-4 是由 OpenAI 開發的第四代生成式預訓練轉換器（GPT），於 2023 年 3 月推出，提供 8,192 和 32,768 詞元的上下文窗口兩個版本。該模型支持圖像和文字輸入，能夠解讀圖像內容、總結螢幕擷取的文字，並回答包含圖表的問題。GPT-4 擅長於多種任務，包括內容創作、文本摘要、問答和機器翻譯等。

然而，儘管具備強大的新功能，GPT-4 仍面臨一些挑戰。例如，它仍然傾向於生成虛構的答案（幻覺現象），偶爾會提供不準確或混亂的回應。近來，有使用者反映 GPT-4 變得越來越「懶散」，在回答問題時的準確性下降，回應變得更加模糊甚至有時拒絕提供答案。針對這些問題，OpenAI 已展開調查，但目前尚未明確問題根源。雖然研究團隊聲稱並未對模型本身進行修改，但 GPT-4 的表現出現顯著變化，顯示其內部運作機制和決策過程仍具有高度的黑盒特性，難以完全解釋。

這些問題反映出，即便技術上有所突破，生成式預訓練轉換器在實際應用中仍存在改進空間。為提高模型的準確性、可靠性及可解釋性，持續的研究與優化至關重要，以確保其在各種應用場景中的穩定性和可信度。

2. LLaMa 2

LLaMa 2 是 Meta AI 於 2023 年 2 月推出的一款開源大型語言模型，作為 LLaMA 1 的升級版本，展現了在推理、編碼能力及知識測試等領域的顯著進步。目前該模型開源，並提供下載及微調方法，使研究學者可以自行開發應用，大幅促進 LLM 社群的研究發展。

與 LLaMA 1 相比，LLaMA 2 的參數數量增加至 700 億，是前代的四倍，使其能夠處理更複雜的文本並解決高難度問題。該模型訓練使用的數據量也增加兩倍，涵蓋超過 100 萬條新的人工標註數據。此外，LLaMa 2 支援更長的上下文長度，最多可達 4096 個詞元（token），並發表了 LLaMa 2 Long 模型的研究論文，該模型在處理長文本方面表現優異，在長

短上下文任務中優於多數開源模型，包括 GPT-3.5-turbo-16k。

模型的訓練方法採用 4,000 億個 token 的持續預訓練，並發現一種成本效益高的微調方法，即使無需人工標註資料，仍能在多項任務上展現良好效能。LLaMa 2 的釋出，不僅加強了在推理和編碼方面的能力，也為學術界和技術社群提供了強大的工具，推動了 LLM 在多種應用中的探索與實踐。

2024 年 4 月，Meta 推出最新一代大型語言模型 Llama 3，提供 80 億和 700 億參數兩種版本，並計畫推出 4000 億參數的版本，預計將成為全球首個開源的 GPT-4 級模型。Llama 3 包括預訓練和指令微調兩版本，效能超越 Claude Sonnet、Mistral Medium 和 GPT-3.5。相比前代 Llama 2，Llama 3 的 8B 和 70B 模型性能顯著提升，並可透過網頁版 Meta AI 免登錄使用。根據基準測試結果，Llama 3 8B 在 MMLU、GPQA、HumanEval 等項目中的表現優於 Google Gemma 7B 和 Mistral 7B Instruct。

為強化模型測試，Meta 開發了一套新的人類評估資料集，包含 1800 個涵蓋 12 種應用場景的提示，如建議、腦力激盪、分類、問答、編碼、創意寫作及角色塑造等，進一步強化模型的性能評估。

技術層面，LLaMA 3 採用 128K token 的詞彙表和分組查詢注意力（GQA）技術，模型訓練數據量超過 15T token，是 LLaMA 2 的七倍，並經過多重的過濾與清理，包括啟發式過濾、NSFW 篩選、語義重複資料刪除和文本分類器，確保數據品質。

未來，Meta 計畫推出具多模態、多語言支援及更長上下文處理能力的新模型。此模型規模達 4000 億參數，預期將成為推動研究創新與應用場景拓展的重要里程碑，為多樣化發展帶來更多可能性。

3. Gemini

Google 於 2024 年 2 月推出的 Gemini 模型採用了 Mixture of Experts

(MoE) 架構，相較於前一代 Gemini 1.0 Ultra，性能大幅提升。該模型在訓練效率和任務處理能力上均有所改進，能夠高效處理大量文本、影像、音頻和程式碼，並在分析、分類和摘要等任務中表現優異。

Gemini 支援更高的上下文窗口容量，達到 12.8 萬個 tokens，能處理更為複雜和多樣化的資料輸入，包括多達 30,000 行程式碼、11 小時的音訊和一小時的影片。這使得模型在面對大量訊息時，能夠保持輸出的一致性與相關性，提升了對長文本及多模態資料的處理能力。然而，隨著上下文窗口的增加，模型的運算需求也隨之提升，可能導致較長的處理等待時間。

4. Mistral AI

Mixtral 是由法國新創公司 Mistral AI 於 2023 年 12 月推出的生成式 AI 模型。Mistral AI 成立僅 7 個月便吸引了大量投資，其推出的 Mixtral 8x7B 模型旨在競爭 OpenAI 等巨頭在大型語言模型 (LLM) 領域的市場份額。Mistral AI 採取「完全開源」的方式來推出新的 LLM，以開放、透明的商業模式吸引了眾多關注。Mixtral 8x7B 的主要特點在於其高效的文本理解和生成能力，這是透過稀疏混合專家模型 (SMoE) 實現的。這種模型架構允許在需要時啟用特定的專家，從而提高模型的計算效率和推理速度。此外，Mistral AI 將 Mixtral 8x7B 完全開源，這種透明的方式使得研究人員和開發者能夠自由訪問和改進模型，並且其開放透明的商業模式也增加了模型的吸引力。

Mixtral 8x7B 不僅在文本生成方面表現出色，還在摘要、分類、文本完成和程式碼生成等多種任務中顯示出其多功能性，這使得它成為生成 AI 解決方案的理想選擇，特別是在需要高效文本處理的場景中。這一模型的高效性得益於稀疏混合專家模型的使用，這使得 Mixtral 能夠在文本理解和生成方面提供卓越的性能。此外，Mistral AI 強調其模型在提供高效能的同時，具備優異的成本效益，這對於企業和開發者而言，將大幅提升其

應用價值與競爭力。Mixtral 8x7B 的推出，展示了 Mistral AI 在大型語言模型領域的創新能力與競爭優勢，透過開源和透明化的策略，該模型為研究與應用拓展了更多可能性，不僅推動了生成式 AI 技術的發展，更為多元應用場景提供了強而有力的技術支援。

5. Claude 3

Anthropic 公司於 2024 年 3 月推出最新 AI 模型「Claude 3」，其性能被視為超越 OpenAI 的 GPT-4 和 Google 的 Gemini Ultra，成為全球最最先進的 AI 模型之一。根據相關研究報告，Claude 3 在數學運算、知識背景、撰寫程式等方面展現出卓越的能力，並能在測試中識別幽默元素，顯示出一定程度的類人特質。

模型有三個版本，分別是 Opus、Sonnet 和 Haiku，應用範圍涵蓋數據分析、客戶互動的多個領域。該模型以極快的速度和優異的推理能力解決問題，並具備處理長篇文章內容的記憶能力，即使無法生成圖像，仍在文字理解與多模態處理等方面表現優異。

這一推出表明了 Anthropic 公司在人工智慧領域的創新能力與競爭優勢，為多樣化應用場景提供了強大的技術支援，並在解決現實問題方面發揮了重要作用。

表 1 生成式大型語言模型比較

	GPT-4	LLaMa 2	LLama 3	Gemini 1.5	Mixtral AI	Claude 3
研發公司	OpenAI	Meta AI	Meta AI	Google	Mistral AI	Anthropic
推出時間	2023/03	2023/02	2024/04	2024/02	2023/12	2024/03
上下文窗口大小	8,192 or 32,768	4,096	8,000	1,000,000	32,000	200,000
是否	否	是	是	是	是	否

開源						
多模態	是	是	不是	是	否	是
應用範疇	內容創作、摘要、問答、機器翻譯等	文本生成、摘要、分類、翻譯、代碼生成等	文本生成、摘要、分類、翻譯、代碼生成等	文本分析、摘要、分類等	文本生成、摘要、分類、翻譯、代碼生成等	數據分析、客戶互動等

二、大型語言模型的應用

大型語言模型具有極為廣泛的應用潛力，其靈活性和多功能性使其在各領域均能發揮關鍵作用。例如：在跨語言溝通方面，模型不僅能進程式語言之間的轉換，亦可實現自然語言之間的精準翻譯，進一步促進了國際間的交流與理解。在創意與藝術領域，可應用於創作文學、詩歌及音樂創作，生成的文本能啟發創意思維，為藝術家提供新的靈感和創作方向。此外，模型還能根據文字內容生成影片和圖片，為視覺藝術創作提供更多可能性。

在商業應用方面，大型語言模型廣泛應用於智能客服系統、廣告生成及市場分析等領域。它們能自動回應客戶查詢、生成個性化廣告內容，並分析大量市場數據，為業務決策提供支持。上述應用案例充分展示了大型語言模型在不同領域中的多樣化應用，為各行行業提供智能化、高效率的解決方案。

(一) CodeGeeX

CodeGeeX 是一款基於華為 MindSpore 框架實現的多語言程式碼生成模型，其第二代版本 CodeGeeX2 於 2023 年基於 ChatGLM2 架構開發完成。該系統具強大的編碼能力，可支援 Python、C++、Java、JavaScript、Go 等多種程式語言，並能執行多項任務，包括程式碼註解、翻譯程式碼、錯誤

除錯及程式碼生成等。

與前代模型相比，CodeGeeX2 具有更快的推理速度和輕量化的本地部署能力，大幅提升開發效率。雖然目前 CodeGeeX 尚未全面開源，但開發者可透過提交申請，獲取模型權重與參數，經審核後用於研究和應用。

(二) Sora

Sora 是 OpenAI 於 2024 年 2 月推出的深度學習模型，利用擴散模型 (Diffusion model) 生成影片。其運作方式是將影片分割成小塊的「補丁」(Patch)，然後從雜訊中逐步還原每個補丁，直至重建完整影片。這種方法有效統一了視覺資料的不同模式，提升了生成模型的可擴展性和訓練效率。

Sora 模型的輸入可以是文字描述的提示語或靜態影像，並能生成最長達一分鐘的高畫質影片。相較於先前的文字轉影片生成模型，Sora 在影片長度、真實性和模擬仿真性方面均有顯著提升。目前，Sora 為閉源模型，OpenAI 尚未公開其模型參數或訓練資料集，但以分享部分設計概念。儘管如此，Sora 模型的問世仍被視為生成影片領域的重要進展。

(三) Stable Diffusion

Stable Diffusion (SD3) 是由 StabilityAI 團隊於 2022 年 8 月開發的文字轉圖片深度學習模型，為該系列中最先進的版本。SD3 透過從噪音中生成數據，採用迭代方法逼近並泛化訓練數據分佈，以生成高質量、符合需求的圖像。

在人類反饋的評估中，SD3 在「提示遵循」、「構圖設計」和「視覺美學」等方面表現優異，超越了其他開源及專有模型。然而，由於其迭代運算的特性，模型在推理過程中需要較高的計算成本及較長的取樣時間，這成為其主要限制。

為了應對上述挑戰，研究團隊提出改良的 directified flow 模型和可擴展的文本到圖像合成架構。透過大規模的研究，他們探索了新的噪聲取樣

方法，進一步提升模型性能。同時新的架構證明在多項指標及人類評價中具備優異表現，並展現出高度可伸縮性。

SD3 模型目前已全面開源，其實驗結果、程式碼和模型權重均已公開，為更廣泛的研究與應用提供了有力支持。

(四) CopyrightCatcher

Patronus AI 於 2024 年 3 月推出「CopyrightCatcher」API，這是業界首個用於檢測大型語言模型（LLM）輸出是否包含侵權內容的服務。該 API 能夠識別 LLM 生成與書籍等文本來源完全相同的內容，並對輸出進行評分，以判定是否涉及版權侵權，同時標示具體侵權內容的位置。

根據 Patronus AI 的研究，目前主流 LLM 在生成侵權內容的比例令人驚訝。「CopyrightCatcher」的問世，不僅反映了業界對於 LLM 生成內容可能侵權的關切，也凸顯了開發更強大解決方案以降低版權風險的迫切需求。

(五) Firewall for AI

Cloudflare 團隊於 2024 年 3 月推出了專為大型語言模型（LLM）應用設計的網頁應用程式防火牆（WAF）工具——Firewall for AI。該工具旨在檢測應用程式漏洞，並提供對模型潛在威脅的可見性。Firewall for AI 部署於 LLM 應用程式之前，通過 API 方式攔截並分析發送至 LLM 模型的請求。

此工具能夠監控每個請求，並控制允許通過的提示詞和請求，從而有效降低濫用行為及資料外流風險，防範巨量流量攻擊和敏感資料洩漏。Firewall for AI 不僅可與 Cloudflare Workers AI 平台整合使用，也可部署於其他第三方基礎設施上，提供更廣泛的安全保護。

該工具的優勢在於能夠及早偵測攻擊，為用戶和模型提供全面保護。然而，這層安全防護可能會增加系統部署與管理的複雜性，需在安全與操作成本之間取得平衡。

(六) Suno AI

Suno AI 是一個由麻省劍橋的音樂家與 Suno 公司人工智慧專家共同開發的音樂生成平台，於 2024 年 3 月推出第三代版本。使用者僅需輸入歌詞，即可根據指定風格生成完整歌曲，無需任何樂器參與。

該平台的核心運作分為兩個步驟：首先，檢索模型從龐大的資料庫中挑選出與輸入相關的文本片段；接著，生成模型根據這些片段創作出最終的歌曲文本。

(七) Scalable Instructable Multiworld Agent (SIMA)

SIMA 是由 DeepMind 團隊於 2024 年 3 月開發的系統，能夠在多個虛擬環境中透過自然語言指令進行操作。相較於以往專注於單一遊戲的研究，SIMA 展現了可擴展性、規模化、適應性及多功能性，標誌著一項革命性的進步。

DeepMind 與多家遊戲開發商合作，訓練 SIMA 在多種遊戲環境中執行任務，包括《無人深空》、《Teardown》和《模擬山羊 3》等。此研究首次證明，大型語言模型能夠理解多樣化的遊戲世界，並像人類一樣透過自然語言指令完成任務。該突破顯示通用大型語言模型的能力正從虛擬世界逐步擴展至現實應用場景。

(八) DocLLM

由於企業經常需要處理各種格式的文件，例如表格、發票、收據或合同等，這些文件通常具有固定格式，並需進行交叉比對與補充內容。DocLLM 模型可在推理過程中同時考慮文本語義與空間佈局，有效解決文件中不規則佈局及異構內容的挑戰。目前，該模型已開源供大眾使用。

DocLLM 模型包含以下四個關鍵概念，如圖 1 所表示：

1. OCRed Document：輸入文檔包含文本 tokens 及其對應的表格邊框。
2. LLM extension: 語言模型的注意機制被擴展，以捕捉文本語義與空間佈局間的依賴關係。

3. Pre-training: 填充文本塊作為預訓練目標。
4. Instruction Tuning: 在重新整理的指令數據集上進行任務適應，提升模型在特定應用場景中的表現。

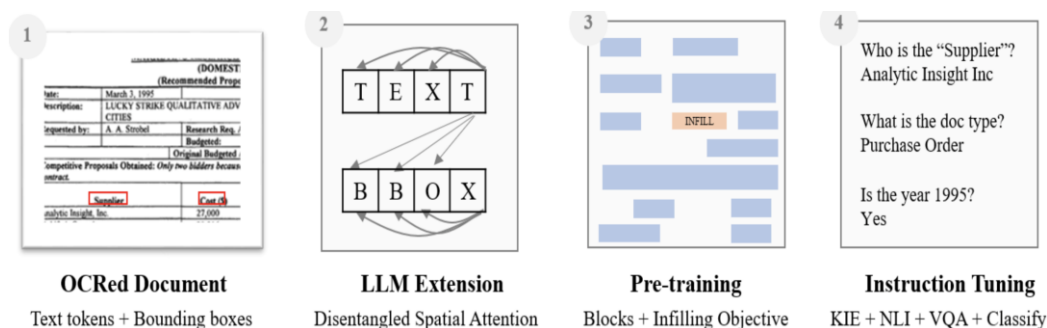


圖1 DocLLM 運作流程步驟

來源：DocLLM: A layout-aware generative language model for multimodal document understanding

三、大型語言模型的技術

隨著大型語言模型的持續發展，相關技術也在不斷演進，以加速運算效能、提高模型準確率以及改進模型評估方法。以下將介紹一些常見且重要的技術：

(一) RLHF (Reinforcement Learning from Human Feedback)

強化學習從人類反饋 (Reinforcement Learning from Human Feedback, RLHF) 是一種機器學習技術，由 Dr. Richard L.於 2019 年提出。該技術通過人類反饋調整模型，使其能夠更好地實現特定目標，是訓練大型語言模型 (LLMs) 的關鍵方法之一。RLHF 使模型能更靈活地適應不斷變化的需求，並提高生成符合人類語言特徵的回應能力。

然而，RLHF 也存在一些挑戰，例如模型可能生成敏感訊息、不真實內容、偏見或不良偏好。與傳統監督學習相比，RLHF 展現出更強的泛化能力和適應性，使模型能在更廣泛的應用場景中學習，並更好地模仿人類行為與偏好，如圖 2 所表示。

1. 人類反饋：模型與人類進行互動，這些反饋被用作監督學習的資料，

以幫助訓練獎勵模型。接收人類專家的回饋，這些回饋基於模型的行為與表現。

2. 獎勵模型：獎勵模型利用人類提供的回饋進行訓練，透過監督學習模仿人類的評價標準。
3. 策略：策略模型基於來自獎勵模型的反饋訊號進行調整，優化自身行為，從而生成更符合人類期望的回應。

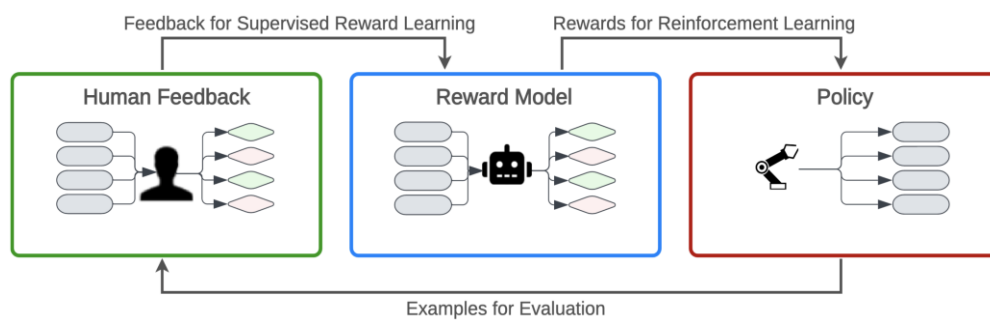


圖2 強化學習從人類反饋 (RLHF) 運作流程步驟

來源：Open problems and fundamental limitations of reinforcement learning from human feedback

(二) RAG (Retrieval-Augmented Generation)

RAG 技術由 MetaAI 團隊於 2020 年提出，其創新之處在於將檢索模型與生成模型相結合，形成一種全新的工作流程。該技術的運作分為兩個主要步驟：首先，檢索模型從龐大的資料庫中選取相關文本片段，作為後續生成過程的基礎；接著，生成模型根據這些檢索到的片段進行文本生成，將其作為參考或引導。

透過引入檢索模型，RAG 技術為生成模型提供了更豐富的上下文訊息，有效避免生成文本與上下文不符或無意義的問題。同時，這也解決了生成模型缺乏特定領域知識而導致內容不準確的問題，使生成結果更具準確性和可信度。

RAG 技術應用前景廣闊，對於需要大量資訊且需進行準確生成的任務，具備顯著價值，例如：

1. 醫療領域：可檢索醫學文獻並生成專業的醫學報告或診斷建議，為醫生進行診斷和治療提供輔助支持。
 2. 教育領域：可自動生成教案、教學大綱及練習題等教學資源，有助於提高教學效率和質量。
 3. 企業領域：在客戶服務中，透過檢索模型快速定位相關問題與解答，再由生成模型生成個性化回應，提升客戶滿意度並加快問題解決速度。
- 此技術為多個領域帶來智能化解決方案，展現出極高的實用價值。

(三) Quantization

Quantization 技術最初由 Netscape 團隊於 1990 年代初首次提出，最早應用於硬體設計。隨著深度學習技術的快速發展，Quantization 技術逐漸被引入到深度學習的部署和推理中，成為優化模型效能的重要工具。

Quantization 的核心原理是將深度模型的權重與參數從高精度的浮點表示轉換為低精度的定點或整數表示，藉此降低模型的計算量並減少記憶體佔用，從而提高模型在部署與推理階段的效率。其具體流程包括以下步驟：

1. 將原始的浮點表示的模型權重和參數轉換為定點或整數表示。
2. 在模型推理階段，使用這些低精度表示的權重與參數進行計算，減少計算資源的需求。

雖然 Quantization 能降低計算成本並減少記憶體使用，但也可能因精度的降低而導致模型性能下降。在不同的硬體平台或應用場景下，Quantization 的應用往往需要重新訓練或微調模型，以實現性能與精度之間的平衡。因此，如何根據具體需求進行量化策略的調整與優化，是確保 Quantization 技術有效應用的關鍵。

(四) LoRA (Low-Rank Adaptation of Large Language Models)

LoRA 技術由 Microsoft 團隊於 2021 年提出，旨在解決大型語言模型

(LLM) 微調過程中高成本的問題，提升模型效能和泛化能力。隨著預訓練模型規模的擴大，對所有模型參數進行完全微調變得越來越不可行。例如：微調 GPT-3 175B 需要極高的部屬成本，並涉及大量參數調整。

為應對此問題，LoRA 技術應運而生。如圖 3 所示，橘色模組為需訓練的 LoRA 權重，透過限制中間 Rank (r) 來大幅減少需調整的參數量。同時，LoRA 保持原始預訓練模型參數不變，顯著降低了下游任務中需訓練的參數數量，進一步減少訓練成本。

LoRA 技術已在多種模型（如 RoBERTa、DeBERTa、GPT-2 和 GPT-3）上驗證，並在多數情況下達到與標準微調（Fine-tuning）相當甚至更佳的性能。然而，該技術仍面臨一些挑戰：低秩分解過程中的超參數設定和優化可能影響模型性能，並可能限制模型的表徵能力。

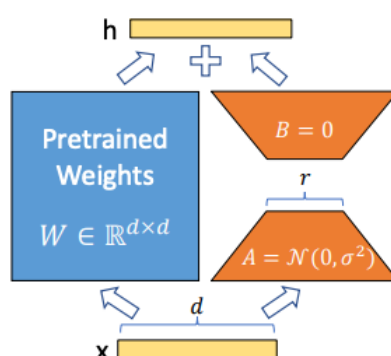


圖3 LORA 運作流程步驟[19]

(五) LangChain

Langchain 是一個專為語言模型設計的框架，旨在利用大型語言模型 (LLM) 來構建各類應用程式。傳統的 LLM 在多語言處理中常面臨支援不足、語言轉換效率低及模型訓練與部署複雜等問題。為了解決這些挑戰，Langchain 提出了一種創新的解決方案。

其核心運作方式是建立一個統一的框架，將多語言文本數據無縫整合至單一模型，提升多語言處理與理解的效率。透過統一的語言表示形式，Langchain 簡化了後續處理與分析的流程，使 LLM 在多語言場景中更具適

用性。

Langchain 技術廣泛應用於自然語言處理、跨語言資訊檢索、多語言翻譯等領域，為多語言應用的開發與部署提供了全新可能，顯著降低了多語言系統構建的複雜性。

（六）Aspire

Google 於 2024 年 1 月推出 Aspire 框架，旨在透過自我評估機制持續優化模型輸出，強化語言模型的選擇性預測（Selective Prediction）能力。該框架分為三個主要階段：

1. 特定任務調整：透過微調預訓練的語言模型，提升其在特定任務上的預測性能，確保模型更能針對特定需求進行優化。
2. 答案採樣：模型生成多個可能的答案，並對其進行篩選，從而建立一個自我評估學習的資料集。這一階段幫助模型學習如何區分正確答案與其他選項。
3. 自我評估：模型根據學習到的資料，自行評估答案的準確性，進一步提升回答問題的可靠性與準確性。

（七）Chatbot Arena

Chatbot Arena 是由 LMSYS 和加州大學柏克萊分校 SkyLab 團隊成員共同開發的一個開源研究項目，旨在為大型語言模型（LLM）提供基準測試平台。該平台透過匿名、隨機對抗評測的方式運作，並採用象棋等競技遊戲中常見的 Elo 評分系統進行模型排名。

Chatbot Arena 已發布 9 個開源 LLM 模型的 Elo 評分及其排行榜，並使用 FastChat 多模型服務系統提供多語言交互界面，從用戶投票中收集數據進行分析。該平台的主要優勢在於提供了一種有效的模型排名機制，並計劃進一步優化採樣算法、排名方法及服務系統。

然而，該平台的數據準確性與可靠性高度依賴於大量用戶的參與，評測結果也可能受到用戶偏好及投票數量的影響。

四、應用大型語言模型案例

隨著更多效能更好的 LLM 不斷被發表，這些創新加速了自然語言處理和人工智慧領域的發展，開拓了全新的應用領域。從個人用戶到企業用戶，越來越多的公司紛紛導入生成式 AI 技術，以提升客戶體驗、改善產品設計、輔助企業內部管理等。在金融領域，企業利用 LLM 技術來分析市場趨勢、進行投資組合管理，甚至開發智能交易系統；在行銷領域，許多公司利用生成式 AI 來創建個性化的行銷內容，提高用戶參與度和轉換率；在服務業，一些企業運用 LLM 技術來改進客戶服務流程，實現更高效的客戶支援和問題解決。這些創新不僅帶來了更好的用戶體驗，也為企業帶來了更高的競爭力和業務增長。生成式 AI 技術的不斷成熟和普及，可以預期這些趨勢將繼續發展，為各個產業帶來更多的創新和機遇。

（一）美國

美國國防部於 2022 年成立「首席數位人工智慧辦公室」（CDAO），旨在推動美軍數位化以確保其軍事優勢。2023 年，該辦公室與 Scale AI 簽署合作協議。Scale AI 於 2016 年成立於美國舊金山，是一家公司，專門利用人工智慧和機器學習技術幫助企業解決問題並提升工作效率，以提供高品質資料標記服務著稱。透過 Scale 的資料引擎，美軍得以管理和標註軍隊資料，這項功能將供跨層級作戰人員使用。美軍已開始積極採用大型語言模型（LLM）技術來輔助軍事策略，強化情報分析、預測威脅，以及提高戰術和戰略決策能力。此外，這項合約也滿足作戰司令部對增強資料品質以支援機器學習演算法的需求。Scale 的工具套件適用於陸軍、海軍陸戰隊、海軍、空軍、太空軍和海岸巡防隊的專案，涵蓋非機密和機密的工作流程。美國軍隊對人工智慧的龐大投資，顯示了美國政府對該技術的重視和戰略布局。未來，大型語言模型將能快速分析大量文字資料、提取相關資訊並生成關鍵摘要來支援決策過程，還能識別潛在的安全威脅、異常

或未經授權的資料存取，以保護敏感資訊並增強國家安全，進一步提升美國軍隊在全球的戰略優勢和影響力。

（二）印度

印度政府以實現「在印度製造人工智慧」和「讓人工智慧為印度服務」的願景為目標，於 2024 年 3 月推出 IndiaAI 任務，計劃投入約 12.5 億美元。這項任務由數位印度公司（DIC）旗下的「IndiaAI」獨立業務部門（IBD）負責執行，旨在透過公私合作來建立擁有 1 萬個以上 CPU 的運算架構。IndiaAI 任務的主要目標包括：滿足印度快速增長的 AI 新創公司和研究生態系統對運算能力的需求；設立開發與部署本土多模態大型語言模型及特定領域基礎模型的創新中心；建立能為新創公司和研發人員提供數據的平台；採用具影響力的 AI 解決方案以推動大規模社會經濟轉型；在教育領域擴展 AI 課程；提供融資支持 AI 新創公司發展；以及開發安全且可信賴的 AI 技術。若 IndiaAI 計劃成功，印度不僅能掌握技術主權，還能創造更多就業機會，並向全球展示如何利用這種變革性技術來造福社會及提升國際競爭力。

（三）日本

日本電氣（NEC）是一家日本跨國資訊科技公司，在台灣設有恩益禧股份有限公司。自 2018 年以來，NEC 持續與日本地方政府合作，開發人工智慧聊天機器人，其中，橫濱市政策局是其合作夥伴之一，透過聊天機器人為市民提供有關市政政策和服務的資訊，並回應常見問題。其他日本政府機構也採用 NEC 的聊天機器人，以提升居民服務的質量和效率。該機器人能回答有關市政服務和即將舉行的活動相關問題，甚至協助居民填寫線上表單，從而提升市政服務的便利性和市民滿意度。

2023 年 7 月 1 日，NEC 成立了「NEC Generative AI Hub」，該組織由人工智慧專業人員組成，專注於推動客戶的業務轉型。透過大型語言模型

(LLM)，NEC 開發的系統可將企業內部系統開發的程式碼撰寫時間減少 80%，文件編寫時間減少 50%，會議記錄時間則從平均 30 分鐘縮減至約 5 分鐘。

此外，NEC 還開發了一套結合大型語言模型與影像分析技術的災害損害評估系統，能在災難發生時迅速且準確地評估損害程度和位置。該系統未來有望為救災過程中的關鍵工具。

(四) 新加坡

在 2023 年新加坡人工智慧會議上，新加坡副總理兼財政部長黃循財正式宣布啟動「全國人工智慧策略 2.0」(NAIS 2.0)，旨在加速人工智慧 (AI) 的應用與發展，力圖使新加坡成為全球 AI 部署及創新領域的領導者。NAIS 2.0 計畫聚焦於開發具有東南亞特色的 AI 技術與應用。作為計畫的一大亮點，新加坡將與美國半導體公司輝達 (Nvidia) 合作，共同研發一款涵蓋 11 種語言的大型語言模型 (LLM)。該項目由新加坡資訊通信媒體發展局 (IMDA) 牽頭，總投資額達 7000 萬新元，所開發的「Sea-Lion」模型旨在結合新加坡及東南亞的語言與文化特徵，為區域性 AI 應用奠定技術基礎。

在醫療領域，新加坡國立大學醫院 (NUHS) 已啟用一套專用的醫療超級電腦，用於支持基因組學、大型語言模型 (LLM) 及醫學影像分析等領域的創新研究與應用。基於此計算資源，NUHS 開發了一系列針對醫療場景的專用模型。其中，RUSSELL-GPT 是一款專為醫療用途設計的語言模型，具備病歷撰寫、醫學指南生成及轉診單整理等功能，旨在減少人為疏失並輔助醫師在病患病史分析及預測方面的診斷。此外，該模型正積極拓展應用，研究其在常見疾病如尿道感染等領域中的病情預測能力，進一步提升診斷準確性。

為提升牙科診療的效率，NUHS 亦推出「SMILE AI」專案，專注於開

發生成牙位圖的人工智慧模型。該系統由兩個子模型組成：一個基於 3D 牙科掃描影像進行訓練，另一個則基於上下顎全口 X 光影像。透過整合這兩個子模型，SMILE AI 能在短短 5 分鐘內，從患者的影像掃描生成完整的 3D 牙位圖，並詳細標示牙齒狀況，顯著縮短牙科醫師從影像取得到初步診療方案制定的流程，提供更高效且直觀的診療輔助工具。

(五) 摩根大通集團 (JPMorgan Chase & Co.)

摩根大通集團 (JPMorgan Chase & Co.) 執行長 Jamie Dimon 於 2023 年採訪中，表達對人工智慧技術未來發展的高度期待。他認為，AI 是推動全球變革的重要驅動力之一。目前，該集團已將 AI 技術應用於來風險管理、自動化預測分析及投資策略優化等領域，並計劃未來進一步擴展至交易、研究和數據庫等環節，實現工作流程的全面智能化。

2023 年 5 月，摩根大通向美國專利商標局提交 IndexGPT 的商標申請。該產品類似於 ChatGPT，利用人工智慧技術為客戶量身打造投資方案，充當 AI 財務顧問，通過自動計算和分析，為客戶提供個性化的財務建議。

同年年底，摩根大通人工智慧研究小組發表了一篇關於 DocLLM 模型的技術論文。是一款專為分析商業文件設計的大型語言模型，能處理如表格、發票、收據、報告及合約等多種類型的文件。這些文檔通常結合了豐富的文本與格式信息，DocLLM 能夠精確解析文本內容與版面設計之間的關聯性，從而高效提取並分析關鍵數據。

特別是財務領域，DocLLM 在處理銀行對帳單、理賠申請及收帳明細等複雜文件時表現出色。同時，它還能將收益報告及財務報表中的數據結構化，大幅提升業務效率並降低操作成本，為金融機構提供更高效率的解決方案。

(六) 四大會計事務所

全球四大會計事務所已開始將人工智慧廣泛應用於審計流程，以提升

工作效率及精確性。

勤業眾信聯合會計師事務所（Deloitte）早在 2015 年便推出了其機器學習驅動的審計應用系統 Argus。該系統能自動從文件中擷取關鍵資訊、辨識潛在風險、趨勢和異常。Argus 的設計目的是減少手動操作，提高審計過程中的準確性，幫助審計人員更快地發現潛在風險，該系統在發表當年即榮獲「審計最佳創新獎」。

勤業眾信近期還推出了名為 DARTbot 的聊天機器人，旨在為內部約 18,000 名美國審計與保險專業人員提供專業建議與輔助支持。DARTbot 能協助員工處理日常待辦事項，並輔助決策流程。勤業眾信表示，透過結合生成式人工智慧與專業人員的經驗與判斷，能夠為客戶提供更深入的洞見與差異化服務。此外，DARTbot 在安全環境中運行，遵循嚴格的數據加密與存取控制政策，確保使用者數據的隱私與安全。

安永會計師事務所（EY）推出了 Helix 系統，專注於總帳和日記帳的檢查和分析。Helix 能快速為團隊提供即時財務數據，並對風險因素進行權重分析，涵蓋抵押貸款、庫存管理及應付帳款等領域，從而提升審計過程中的風險管理和評估能力。

安侯建業聯合會計師事務所（KPMG）則在 2020 年推出 Clara 平臺，結合預測分析、視覺化技術與認知科技，強化了基於數據的風險評估能力。Clara 可與客戶進行互動，通過智能分析引擎即時收集反饋，並與 IBM Watson 合作，進一步提升 AI 分析性能，從而使審計過程更高效與精確。

資誠聯合會計師事務所（PwC）則開發了 GL.ai 和 Halo 工具。GL.ai 利用演算法，從分類帳中檢測異常，幫助審計人員快速鎖定潛在風險區域；而 Halo 則可收集並分析大量財務數據，進行風險評估及測試，生成關鍵見解，從而提升審計的全面性與決策準確性。

（七）大成（Dentons）

大成（Dentons）法律事務所為一家全球性的法律事務所，於 2023 年推出基於 ChatGPT 技術的「fleetAI」系統，旨在提升內部法律專業人員工作效率。該系統支持法律研究、法律文件生成及相關法律論證的辨識，並可批量上傳法律文件進行分析與法條檢索。

為保障數據安全，Dentons 與 Microsoft 合作，確保所有上傳資料均受到嚴格管理，上傳的資料不會被用於模型訓練，且會在 30 天內自動刪除，確保客戶機密信息的完整性與安全性。在對資料隱私要求極高的法律行業中，Dentons 開始積極探索生成式 AI 的應用，改變律師在法律研究、訴狀草稿撰寫及客戶互動中的傳統工作方式，實現法律專業與先進技術的深度結合。

大型語言模型（LLM）的開發已不僅限於科技產業，各行各業均積極將其應用於自身領域，藉此提升業務效率與競爭力。

（八）微軟（Microsoft）

微軟（Microsoft）開發大型語言模型，並推出全新 Bing 搜尋引擎與 Edge 瀏覽器，將搜尋、瀏覽，與聊天無縫整合，以提升使用者的搜尋體驗。透過簡單的關鍵字輸入，Bing 能提供更精準的搜尋結果及更完整的答案。同時也能引導使用者進一步思考，激發新的問題與靈感。在搜尋過程中，Bing 還會附上資訊的來源連結，讓使用者能輕鬆訪問原始網站，深入了解相關內容。

（九）台灣大哥大

台灣大哥大作為台灣三大電信公司之一，近年積極拓展人工智慧市場，結合 TWCC 臺灣 AI 雲平台及 5G 技術，提供高效能的 AI 運算服務。台灣大哥大採用大型語言模型（LLM）對智能客服系統「小麥」進行優化，透過嵌入式模型和專業客服知識庫的微調，「小麥」的回應更加精確且富有人性化，顯著提升了用戶互動體驗，滿足多樣化需求。

此外，台灣大哥大推出「AI2.0 Solution Suite」平台，作為一站式的 AIaaS（AI-as-a-Service，人工智慧即服務）解決方案，並以 LLM 作為技術核心，提供包括模型工具箱、語音辨識及語音生成等功能，幫助企業快速實現多元 AI 應用。憑藉其在電信業的基礎設施與混合雲架構團隊，台灣大哥大提供機房代管、通訊網路及混合雲設計等服務，協助企業在雲端部署過程中靈活調整運算資源，同時開發 AI 協作解決方案。

「AI2.0 Solution Suite」幫助企業快速實現多樣化的 AI 應用，並藉助台灣大哥大的技術支持提升業務效率與競爭力。表 2 中總結了不同 AI 技術在實際應用中的優缺點，為企業選擇適合的 AI 技術方案提供了重要參考。

表 2 使用大型語言模型的案例整理

企業/機構	優點	可能狀況
摩根大通集團	準確地提取和分析財務文檔中的複雜內容；提高工作效率與降低成本。	對於格式複雜的文件，模型的準確性仍需進一步提升。
勤業眾信聯合會計師事務所	自動從文件中擷取重要資訊，辨識潛在風險、趨勢及異常。	依賴於模型訓練的質量和數據量，存在誤判的風險。需要定期更新知識庫和模型，確保回答的準確性和及時性。
安永會計師事務所	檢查及分析總帳與日記帳，快速提供團隊財務數據參考；分析風險因素權重，提供深入見解。	需要高質量的數據輸入，存在數據安全和隱私問題。
安侯建業聯合會計師事務所	結合預測分析、視覺化技術、認知科技、人工智慧與機器學習；即時萃取客戶互動回饋，提高互動性。	開發成本較高，且需要不斷更新和維護。
資誠聯合會計師事務所	發現分類帳異常處，改善風險評估、分析和測試；提供關鍵觀點，提升審計質量。	依賴於數據的質量和完整性，存在模型訓練成本高的問題。
大成法律事務所	進行法律研究、生成法律文件和辨識相關法律論證；資	法律行業的複雜性和多變性要求模型需不斷更

企業/機構	優點	可能狀況
	料安全管理嚴謹，改善法律研究分析效率。	新和調整。
微軟	優化使用者搜尋體驗，提供精準結果和完整答案；整合搜尋、瀏覽與聊天功能，提升用戶體驗。	搜索引擎依賴於網絡數據的質量和更新速度。
台灣大哥大	提供高效能的 AI 運算服務，優化智能客服回應；提供一站式 AIaaS 平台，支援多樣態 AI 應用。	初期開發和部署成本高，需不斷優化以適應市場需求。
新加坡國立大學醫院	撰寫病歷、摘要、轉診單和醫學指南 QA；生成 3D 牙位圖，提高醫療診斷和治療效率。	醫療數據安全和隱私問題需特別注意，需不斷更新醫療知識庫以保持準確性。

五、繁體中文的大型語言模型

自大型語言模型問世以來，其應用場景不斷拓展，但多數模型仍以英文為主。為了滿足繁體中文使用者的需求，並融入臺灣文化，近年來專門針對繁體中文訓練的大型語言模型有了顯著進展。以下將介紹 2024 年後針對臺灣市場發表的相關模型，詳見表 3。

(一) 國科會 TAIDE

國科會為滿足臺灣本土化需求，並提升生成式 AI 的可信任性與適用性，推動了「可信任生成式 AI 發展先期計畫」（TAIDE）。該計畫的訓練數據來源涵蓋新聞、社交媒體文章、辭典及政府公告等數位化且公開的文本資源。

在 2023 年 6 月公布的第一階段成果中，TAIDE 展示了在自動摘要、文本翻譯、書信撰寫和文章創作等應用場景中的不錯的表現。下一階段的目標是開發任務導向的模型功能，涵蓋臺灣知識普及、事實查核及臺灣考試等應用，同時納入倫理問答等功能。計畫將邀請公部門人員參與測試，

並委託外部單位進行驗證，預計於 2024 年 3 月展示多輪問答的成果。

2024 年 4 月 15 日，國科會正式宣布釋出 TAIDE-LX-7B 模型，該模型基於 Meta Llama-2-7B 進行擴充，新增大量繁體中文字元與詞彙，使生成文本更為流暢精準，並融入臺灣文化特色，成為首款結合臺灣本土文化的大型語言模型。未來，TAIDE 模型將協助政府提供多元化的生成式 AI 服務，加速生成式 AI 技術在各領域的應用與落地。

(二) 聯發科

聯發科於 2024 年 4 月正式推出其生成式 AI 服務平台 MediaTek DaVinci，該平台旨在為集團內部開發具備資料安全性且能提升生產力的生成式 AI 工具。隨著內部多部門的廣泛使用，聯發科計劃將 MediaTek DaVinci 擴展為對外部企業開放的服務平台，進一步拓展市場應用範疇。

同時，聯發創新基地發表了最新的繁體中文大型語言模型 MediaTek Research BreeXe (MR BreeXe)，該模型擁有 450 億個參數，基於法國 AI 公司 Mixtral 所開發的 Mixtral 8x7B 模型構建，採用多專家模型技術，顯著降低運算資源需求並提升運行效率。經過預訓練後，MR BreeXe 在繁體中文測試（如 TMMLU+ 和 MT Bench TW）中表現超越 GPT-3.5。

此外，MR BreeXe 針對台灣地端應用進行了優化，改善了產業界在生成式 AI 和檢索增強生成（RAG）應用中的使用體驗。聯發創新基地負責人表示未來將結合 MediaTek Research BreeXe 與 MediaTek DaVinci，提供卓越的語言處理能力及雲端與地端的彈性解決方案，更貼近本地需求，並推動生成式 AI 的本土化應用。

這一發展不僅增強了聯發科在生成式 AI 領域的競爭力，也促進了台灣在相關技術應用與發展中的地位。

(三) Llama3

Meta 於 2024 年 4 月推出其最新一代大型語言模型 Llama3，提供 80 億

和 700 億參數的兩種版本，並計劃開發 4000 億參數的版本，若成功推出，該版本將成為全球首個開源 GPT-4 級別的模式。Llama 3 的 80 億和 700 億參數版本均有預訓練與指令微調兩種選項，根據 Meta 說法，Llama 3 在性能上優於 Claude Sonnet、Mistral Medium 和 GPT-3.5。

在基準性能方面，Llama 3 的 8B 和 70B 版本相較於 Llama 2 表現出顯著提升。使用者無需登錄即可直接在網頁版 Meta AI 平台上體驗 Llama3。基準測試結果顯示，Llama 3 8B 在 MMLU、GPQA 和 HumanEval 等多項測試中，得分均超越 Google Gemma 7B 和 Mistral 7B Instruct。

為提升模型在研究基準中的表現，Meta 開發了一套全新的人類評估資料集，包含 1800 個提示，涵蓋 12 種核心應用場景，包括建議生成、腦力激盪、分類、封閉式問答、編碼、創意寫作、信息提取、角色塑造、開放式問答、推理、文本重寫和摘要生成。

與 Llama 2 相比，Llama 3 採用了 128K token 的詞彙表 tokenizer，並引入分組查詢注意力（GQA）技術，以提升推理效率和資料處理質量。此外，Llama 3 使用了超過 15T 的 token 作為訓練數據，資料集規模是 Llama 2 的七倍。

在資料處理過程中，Meta 應用了多種技術，包括啟發式過濾器、NSFW 篩選器、語義重複資料刪除方法和文本分類器，以確保訓練資料的高品質。Meta 亦預告，8B 和 70B 版本僅為 Llama 3 系列的一部分，未來將陸續推出具備多模態、多語言支持以及更長上下文處理能力的新版本，進一步擴展 Llama 3 在各領域的應用潛力。

表 3 繁體中文之大型語言模型整理

模型名稱	TAIDE	MR BreeXe	Llama 3
研發公司/單位	國科會	聯發科	Meta AI

模型名稱	TAIDE	MR BreeXe	Llama 3
推出時間	2024/04	2024/04	2024/04
參數大小	7B	7B	8B/70B
是否開源	是	是	是
多模態	否	否	否
其他	繁體中文	繁體中文	多語言
架構類型	僅解碼器架構	僅解碼器架構	僅解碼器架構
訓練資料	414.4 億個 token 的繁體中文資料進行持續預訓練（資料量約 140GB），並使用 12 萬 8,000 筆單輪或多輪對話問答資料進行指令微調。	使用 650GB 的資料進行基礎預訓練，並使用 25GB 的中文、英文和程式碼作為額外資料，進行長上下文預訓練。	200 億個 token 的繁體中文資料進行持續預訓練，並使用數百萬筆對話資料進行指令微調。

六、大型語言模型審計應用的期待與潛在風險

大型語言模型（LLM）在審計流程中可自動化處理許多重複性任務，例如文檔分類、資料抽取和摘要撰寫。同時，LLM 也能透過分析大量財務報表、揭露資訊和行業特定數據，識別潛在風險因素與錯誤，協助審計人員進行全面的風險評估。對於非結構化的財務和交易數據，LLM 能夠提供分析結果並產出關鍵觀點，最終生成審計報告。隨著 LLM 的應用逐漸深入，未來的審計領域將迎來顯著的革新。

使用 LLM 的成本視模型的使用條件而異，開源或本地部署的模型通常免費，但需要配備相應的基礎設施，而商業模型則可能產生訂閱費用或 API 使用成本。本地化模型雖然免除了直接的使用費用，但可能在反應速

度和提示次數上存在限制。此外，傳統上部署深度學習模型需要豐富的技術知識，包括設置基礎設施、管理硬體和軟體配置等，這些技術門檻限制了 LLM 的廣泛應用。

現代 LLM 系統通過友好的視覺化和交互設計（包括圖形元素、菜單、按鈕和其他設計組件），降低了使用門檻。使用者只需在文字框輸入提示，便能快速獲得模型的回應。此外，API 提供了一種高效的方法，通過預設的規則和協議，讓軟體之間能無縫交換數據和功能，消除手動處理的繁瑣過程，提升自動化水平。

然而 LLM 在應用過程中也伴隨挑戰，各國政府對生成式 AI 技術帶來的社會影響高度關注，並陸續制定監管政策。英國國會通訊與數位委員會於 2024 年發布了《大型語言模型和生成式 AI》報告，提出應積極推動 AI 技術的發展，同時採取嚴格的監管措施。建議包括建立 AI 監管框架、加強技術研發投資、完善法律法規及提升公眾對 AI 的理解和接受度。

生成式 AI 產生內容正確性和倫理性方面仍需改進，尤其在虛假資訊、歧視性言論及隱私侵犯等問題上，各國政府需採取行動防範此類風險。此外，科技公司未經許可使用數據進行模型訓練的行為也引發版權和公平性爭議，成為監管的重點。

儘管面臨多重挑戰，各國政府認識到 LLM 的潛在價值，正加速推動其研究與應用，力求在技術創新與社會利益間達成平衡。雖然目前尚未有繁體中文 LLM 被廣泛應用於公部門審計單位，但業界普遍認為該技術將成為未來審計工作的核心工具。隨著技術成熟，審計應用中的潛在風險亦需高度關注，詳見表 4。

表 4 大型語言模型審計應用的考量與潛在風險

公平性	系統輸出結果中是否有任何不適當的偏見等?要求事先對公平性的定義有一個共同的理解。
-----	--

透明性	系統的輸出結果是否可以再現並達到解釋性等?
安全性	系統是否有可能傷害資料安全性?
保密性	如何預防或檢測對訓練數據的攻擊?是否可以防止故意引入不適當輸出的輸入數據等?
正確性	系統的輸出結果是否能正確或是有所根據的回應?

參、研究方法與步驟

大型語言模型在審計業務中的應用，可透過提示詞指令實現業務流程的自動化監測、風險評估及預測。例如，審計人員可藉由特定提示詞向模型發出查詢，引導其針對特定領域進行深入分析。使用「風險評估」或「效能評估」等指令，模型能針對相關的審計工作提供具針對性的輸出，協助審計人員快速取得與風險相關的關鍵資訊，從而提升工作效率與準確性。同時，提示詞如「法規遵從」或「會計標準」能協助模型提供法律條文或會計準則的詳細解釋，幫助優化資金使用、提升效能，並改善整體財務管理水準。

生成式檢索增強技術（RAG）進一步提升了對審計對象的全面理解，透過外部資料庫的擴增，加強對機構運作、政策執行及財務狀況的洞察。例如，RAG 技術可分析預算執行情況、支出效能、資產運用等指標，進行全面的財務效能評估，為審計人員提供更深入的審計報告及具價值的建議。同時，模型可輔助辨識相關法律條文、法規及合規性指南，確保審計對象遵守相關規範，進一步提升法規遵從性，增強監督機制。

建構基於大型語言模型與 RAG 的審計檢索架構，首要步驟為蒐集與審計業務相關的代表性文本資料，建立完善的外部資料庫以及數位審計應用。接著，將文本轉換成向量表示，透過向量相似度比對提升資訊檢索的精確性。同時，結合提示詞以提升模型對審計指令的理解與回應能力，並將相關文本片段及標準答案納入預訓練模型進行模型指令微調。最終，基於人類反饋的強化學習進一步提升模型的適應性和效能。本研究計畫結合先進的語言模型與檢索技術，致力於為審計業務提供更高效的解決方案。詳細架構如圖 4 所示。

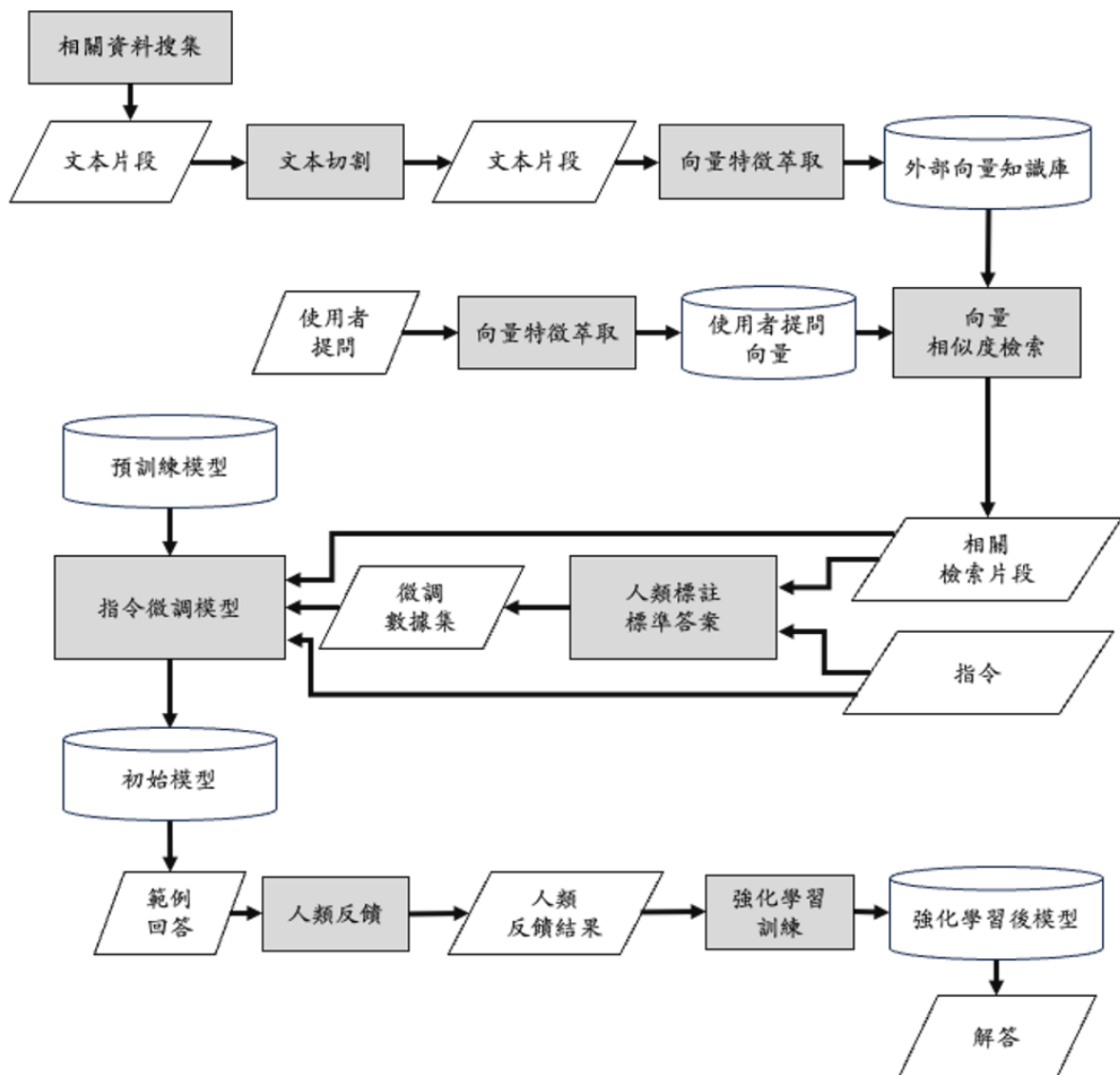


圖 4 研究流程與步驟

一、 外部資料庫蒐集、資料整理及文本切割

在前一步中，收集與審計相關的外部資料後，需對文本進行劃分，將其分割成最小的分析單元，例如，法律條文可以按照條文為單位進行切割。此過程可透過正則表達式、內建分割函數，或根據需求設定的切割邏輯來實現。分割後的文本片段能夠為後續的文本分析提供更加細緻與準確的基礎。

在構建外部資料庫時，我們收集了與審計相關的法律條文、過往案例等資料作為主要來源。在實驗中，從司法院裁判書系統中，以關鍵字「財

務」及「審計」進行查詢，共收集到 82 篇標註裁判書。這些裁判書內容被作為 RAG 補充資料來源，但未考慮判決書中的罪名標籤、日期等非核心要素。我們透過理解判決書內容，設計出問題，並從中找出正確的對應內容。

資料集包括訓練和測試集，訓練包含 400 筆數據，測試集則包含 145 筆數據，資料集儲存為 json 格式，如圖 5 所示。context 字段存放判決書內容，question 字段為設計的問題，answers 字段則為問題的正確對應答案，每個問題會有三個符合的答案，若出現只有兩個答案的狀況，則會有一個答案是重複的。

```
"data": [
  {
    "context": "犯罪事實及證據並所犯法條分敘如下：犯罪事實一、林寬祥為林平形之弟，渠等之父林旅生於民國103年11月23日過世時，遺有屏東縣○○○段○○○○段○○○○段○○○○段○○○○段○○○○地號土地5筆，及坐落在屏東縣○○○段○○○○段○○○○段○○○○地號上之000號（門牌號碼係屏東縣○○○路○○○號）建物1筆（下稱上開土地及建物）。林平形因遠居新竹，遂將印章及印鑑證明交付林寬祥，委託林寬祥辦理繼承土地、土地貸款償還及相關繼承事務，詎林寬祥明知林平形已明確表明其為法定繼承人之一，對遺產有其合法之應繼分，無欲放棄繼承或拋棄繼承，應依法公平辦理登記，竟意圖為自己不法利益，基於行使偽造文書、使公務員登載不實及背信之犯意，於103年12月16日前某日，違背其任務，未經林平形之同意，利用不知情之代書林素霞繕打遺產分割協議書、林旅生繼承系統表及土地登記申請書後，將林平形交付之印鑑章，蓋印在上開遺產分割協議書上，而偽造用以表示林平形同意將上開不動產依遺產分割協議書上所載分配登記。復於103年12月16日，由林寬祥持上開協議書等文件及林平形所交付之印鑑證明、印鑑章，至屏東縣恆春地政事務所，辦理上揭遺產分割協議書上所載之土地、建物所有權應有部分之移轉登記，使各該承辦案件之公務人員經形式審查後，將關於上開土地及建物繼承分割之不實事項，登載於其職務上所掌管之相關土地、建物繼承登記之公文書，進而將上開土地及建物之所有權移轉登記至林寬祥等人名下，以此方式為違背林平形之委託，致生損害於林平形之遺產繼承權利及地政機關對不動產登記管理之正確性。嗣林平形於104年底某日，經其胞妹林秀菊告知後，始查悉上情。二、案經林平形委由孫嘉佑律師（法律扶助律師）告訴偵辦。",
    "question": "行為人受雇於誰或任職於哪?",
    "id": "1-1",
    "answers": {
      "answer_start": [
        23,
        166,
        166
      ],
      "text": [
        "林寬祥為林平形之弟",
        "林平形因遠居新竹，遂將印章及印鑑證明交付林寬祥，委託林寬祥辦理繼承土地、土地貸款償還及相關繼承事務",
        "林平形因遠居新竹，遂將印章及印鑑證明交付林寬祥，委託林寬祥辦理繼承土地、土地貸款償還及相關繼承事務"
      ]
    }
  },
],
```

圖 5 原始資料集樣貌

在文本分割過程中，我們使用 LangChain 框架的文本切割器。以 langchain 架構中的 RecursiveCharacterTextSplitter 將文本做切割，外部資料庫的文本切割，其分割成最小的單位，文本分割的單元不一定是一句話，原則上是基於模型自訂的切割方法來切割。輸入文本的長度需考慮大型語言模型的輸入限制，使用者可自行設定切割單元的大小及片段之間的重疊度，如圖 6 程式碼所表示。

```

### 用langchain框架的文本分割器將文本分割
from langchain.text_splitter import RecursiveCharacterTextSplitter

text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=300,
    chunk_overlap=100,
    length_function=len,
    add_start_index=True,
)
all_splits = text_splitter.split_documents(data)

```

圖 6 文本切割

二、 向量特徵萃取及相似度檢索

利用詞嵌入技術進行向量特徵萃取，能夠有效支持向量資料庫中的資訊儲存與檢索。每段文本首先被轉換為向量，並存入向量資料庫。查詢過程中，查詢內容同樣會被轉換為向量，例如，「根據去年財務報表推斷今年度的監督政府預算機關、基金單位數統計表」這類查詢，會被轉換成對應的向量形式。接著，通過計算查詢向量與資料庫中各文本向量的相似度，我們能迅速找到與查詢內容最相關的文本，並根據相似度高低進行排序，將最相關的文檔排在前列。

在這個過程中，可動態調整變數 k ，檢索出前 k 個相關文檔，以適應不同場景的需求。這種靈活性確保檢索結果能滿足特定情境下的需求，最終，將篩選出的相關資訊匯總並返回至結果列表。

在實驗中，我們使用 langchain 框架讀取 rag.json 檔案作為補充資料，該檔案來源於司法院裁判書系統中的判決書，並通過 LangChain 的 JSONLoader 進行讀取。嵌入模型採用 sentence-transformers/all-MiniLM-L6-v2 的 embedding，未來亦可根據需求更換模型，程式碼如下圖 7 所示。我們使用 chroma 向量資料庫 (vector database) 將處以過的判決書內容切割為最小單位後存入資料庫，並轉換為向量形式，構建與審計相關的外部向量資料庫，以支持後續的檢索與相似度計算，如圖 8 所表示。

在檢索過程中，利用相似度計算，分析使用者查詢向量與資料庫中

文本向量的相似性，從而提取相關文本段落，如圖 9 所示。

嵌入模型的核心功能在於將輸入文本轉換為浮點數值的嵌入向量 (Embeddings Vector)，這些向量能捕捉文本之間的語義關聯。當兩段文本的向量距離較近時，表明它們在語義上具有高度相似性，這一特性能夠在多種應用場景中發揮重要作用，例如：

- (一) 搜尋功能：透過計算文本向量之間的距離，可以快速定位某段文字在文章中的位置，或從大量文章中找出與查詢最相關的內容。
- (二) 分類與聚類：基於文本的嵌入向量，可將相似性高的文章分為同一類別，便於資料整理和管理。
- (三) 資訊檢索與內容分析：嵌入模型能識別文本中的關鍵詞，協助用戶聚焦主題相關內容，提升資訊檢索與分析效率。

```
### 使用embedding模型解析
from torch import cuda
from langchain.embeddings.huggingface import HuggingFaceEmbeddings

embed_model_id = 'sentence-transformers/all-MiniLM-L6-v2'

device = f'cuda:{cuda.current_device()}' if cuda.is_available() else 'cpu'

embed_model = HuggingFaceEmbeddings(
    model_name=embed_model_id,
    model_kwargs={'device': device},
    encode_kwargs={'device': device, 'batch_size': 32}
)
```

圖 7 向量特徵萃取

```
### 使用chroma向量資料庫(vector database)
__import__('pysqlite3')
import pysqlite3
import sys
sys.modules['sqlite3'] = sys.modules["pysqlite3"]
from langchain.vectorstores.chroma import Chroma

vectorstore = Chroma.from_documents(all_splits, embed_model, persist_directory="./chroma/")
```

圖 8 向量特徵存取於資料庫

```

### 選用想要使用的llm，若為台灣本土大型語言，包含Breeze, Taiwan-Llama, Taide等皆可使用，須為gguf格式
from langchain.embeddings import LlamaCppEmbeddings
from langchain.llms import LlamaCpp
from langchain.callbacks.manager import CallbackManager
from langchain.callbacks.streaming_stdout import StreamingStdOutCallbackHandler

n_gpu_layers = -1
n_batch = 512
callback_manager = CallbackManager([StreamingStdOutCallbackHandler()])

llm = LlamaCpp(
    model_path="./Breeze-7B-Base-v0.1-Q8_0.gguf",
    n_gpu_layers=n_gpu_layers,
    n_batch=n_batch,
    n_ctx=10240,
    temperature=0.5,
    max_tokens=300,
    top_p=1,
    f16_kv=True,
    callback_manager=callback_manager,
    verbose=True,
)
### 直接只用大型語言模型進行問答
llm("審計部的業務內容包含什麼？")

```

```

### 增加RAG
from langchain.chains import RetrievalQA

rag_pipeline = RetrievalQA.from_chain_type(
    llm=llm, chain_type='stuff',
    retriever=vectorstore.as_retriever()
)
### 使用大型語言模型+RAG進行問答
rag_pipeline("劉俊南發生了什麼事？")

```

圖 9 大型語言問答以及查詢文件資料庫之問答

三、查詢指令的設計

指令（prompt）是用於引導模型微調的重要提示，結合相關檢索片段，為模型生成相應回應提供關鍵指引。指令的核心目的是在檢索結果的基礎上，協助模型生成符合預期的回應。因此，指令設計需具備明確且具體的指示，不僅涵蓋微調任務的要點，還能引導模型準確理解檢索結果的內容。

本計畫根據不同的審計任務，設計相應的指令，達成數位審計的目標

的。此一標記的目的在於建立一微調資料集，其回覆結果為正確，並用以進行模型微調。人工標記者需能夠提供一致而準確的標準答案，確保模型微調過程中具有高品質的參考依據。

提示詞設計旨在利用檢索結果引導模型根據指令生成答案，並基於檢索到的相關上下文，提升模型回應的準確性。由於檢索結果涵蓋了先前獲取的內容，有助於模型更全面地理解審計業務的相關資訊，進而提高生成回應的相關性與適用性。

我們將原始資料集轉換為符合 llama.cpp 可以 finetune 的格式，如圖 10 所示。每筆資料開頭設計為 `<s>Below is an... the request.`，`###Instruction` 部分為判決書的內容加上設計的問題，`###Response` 部分為對應答案，將所有答案以句號串起來，若出現重複答案，則刪除重複的答案。

```
<s>Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
犯罪事實及證據並所犯法條分敘如下：犯罪事實一、林寬祥為林芋彤之弟，渠等之父林旅生於民國103年11月23日過世時，遺有屏東縣○○○段○○○○段00c0000c0000c00c00地號土地5筆，及坐落在屏東縣○○○段○○○○段00地號上之00建號（門牌號碼係屏東縣○○○路000號）建物1筆（下稱上開土地及建物）。林芋彤因遠居新竹，遂將印章及印鑑證明交付林寬祥，委託林寬祥辦理繼承土地、土地貸款償還及相關繼承事務，詎林寬祥明知林芋彤已明確表明其為法定繼承人之一，對遺產有其合法之應繼分，無欲放棄繼承或拋棄繼承，應依法公平辦理登記，竟意圖為自己不法利益，基於行使偽造文書、使公務員登載不實及背信之犯意，於103年12月16日前某日，違背其任務，未經林芋彤之同意，利用不知情之代書林義壽繕打遺產分割協議書、林旅生繼承系統表及土地登記申請書後，將林芋彤交付之印鑑章，蓋印在上開遺產分割協議書上，而偽造用以表示林芋彤同意將上開不動產依遺產分割協議書上所載分配登記。復於103年12月16日，由林寬祥持上開協議書等文件及林芋彤所交付之印鑑證明、印鑑章，至屏東縣恆春地政事務所，辦理上揭遺產分割協議書上所載之土地、建物所有權應有部分之移轉登記，使該承辦案件之公務人員經形式審查後，將關於上開土地及建物繼承分割之不實事項，登載於其職務上所掌管之相關土地、建物繼承登記之公文書，進而將上開土地及建物之所有權移轉登記至林寬祥等人名下，以此方式為違背林芋彤之委託，致生損害於林芋彤之遺產繼承權利及地政機關對不動產登記管理之正確性。嗣林芋彤於104年底某日，經其胞姊林秀菊告知後，始查悉上情。二、案經林芋彤委由孫嘉佑律師（法律扶助律師）告訴偵辦。

行為人受雇於誰或在職於哪？

### Response:
林寬祥為林芋彤之弟。林芋彤因遠居新竹，遂將印章及印鑑證明交付林寬祥，委託林寬祥辦理繼承土地、土地貸款償還及相關繼承事務。
```

圖 10 整理過後的資料集樣貌

四、基於常見的預訓練模型並微調模型

為了提升預訓練模型在審計業務任務中的性能，我們採用指令微調的方法。原始的預訓練模型通常在大規模數據上進行通用性訓練，以應對多種任務。然而，這些模型在特定領域的表現可能有所差異。例如，模型可能在某些領域（如法律）表現欠佳，而在其他領域則表現優異。為使預訓練模型更好地適應審計業務需求，我們將使用指令微調的方式，包括設計相關的審計指令，並提供這些指令的標準答案，作為微調資料集的一部分。模型可以透過微調過程，學習如何更好地適應審計任務，

提高其在這個特定領域的性能。這種個性化微調有助於模型更好地理解和解讀審計相關的語境和指令，從而提升其在該領域的效能。

由於我們後續要解讀的大多是繁體中文的文件，在此需選用 gguf 格式的繁體中文大型語言模型，並具有台灣本土文化的模型，包含 Breeze、Taiwan-Llama、Taide 和 Taiwan-LLM 皆可使用，如圖 11 所示。

📁 Breeze-7B-Base-v0.1-Q8_0.gguf	18 days ago
📁 Breeze-7B-Instruct-64k-v0.1-Q8_0.gguf	18 days ago
📁 Breeze-7B-Instruct-v0.1-Q8_0.gguf	18 days ago
📁 taide-7b-a.2-q4_k_m.gguf	last month
📁 taide-8b-a.3-q4_k_m.gguf	18 days ago
📁 Taiwan-LLaMa-13b-1.0.Q8_0.gguf	18 days ago
📁 Taiwan-LLM-13B-v2.0-chat-Q8_0.gguf	18 days ago
📁 Taiwan-LLM-7B-v2.0.1-chat-Q8_0.gguf	18 days ago
📁 Taiwan-LLM-7B-v2.1-chat-Q8_0.gguf	18 days ago

圖 11 所使用到的所有台灣本土大型語言模型

透過設計精確的指令，模型能夠準確執行對應任務。這些指令結合查詢所得的文本段落與人類撰寫的標準答案，構成微調資料集，用於模型微調。

我們的微調過程如圖 12，`--model-base` 參數指示欲進行微調的基礎模型，`--train-data` 為 finetune 的訓練資料集，`--STR` 表示資料集的每筆開始符號。

```

### 使用llama.cpp內附的finetune程式碼進行模型finetune

# 安裝llama.cpp所需的程式，並且完成設定
!git clone https://github.com/ggerganov/llama.cpp
!cd llama.cpp
!cmake -B build
!cmake --build build --config Release
!cmake -B build -DLLAMA_CUDA=ON
# 進行finetune
!./bin/finetune --model-base ../../taide-7b-a.2-q4_k_m.gguf \
                --lora-out ../../taide-7b-a.2-q4_k_m-FINETUNE.gguf \
                --train-data "../../langchain/training_combine_answer.txt" \
                --save-every 10 --threads 6 --adam-iter 30 --batch 4 \
                --ctx 64 --use-checkpointing --STR "<s>"

```

圖 12 台灣本土大型語言模型之問答微調

五、預訓練模型微調結果評估

在將大型語言模型的問答結果與資料集中的標準答案進行比對時，我們採用了 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) 和 BLEU (Bilingual Evaluation Understudy) 作為評估指標。這兩種指標是文本生成系統 (如機器翻譯與文本生成) 中常用的自動化評估方法。

ROUGE 的評分主要包括召回率 (Recall)、精確度 (Precision) 和 F1 分數 (F1 score)。召回率是指生成文本中覆蓋參考文本的比例，而精確度則是參考文本覆蓋生成文本的比例。F1 分數是召回率與精確度的調和平均數，用以綜合評估生成文本的質量。

ROUGE 方法是一種有效的自動化文本摘要評估工具，特別適用於生成型問答系統等領域。傳統上，文本的摘要品質評估通常依賴人工判斷，例如連貫性、簡潔性、語法正確性和內容完整性等。然而，這種方法耗時且昂貴，且具有一定的主觀性。ROUGE 提供了一種自動化的評估方式，通過比較電腦生成摘要與人工參考摘要的相似性來衡量摘要質量。

ROUGE 通過計算候選摘要和參考摘要之間共同出現的 n-gram、詞序列或最長公共子序列 (Longest Common Subsequence, LCS)，來衡量摘

要的質量。例如，ROUGE-1 衡量單詞級別的重疊，ROUGE-2 衡量雙詞級別的重疊，而 ROUGE-L 則基於最長公共子序列的重疊度來衡量生成文本和參考文本之間的最長公共子序列，考量文本生成過程中詞序與長度的匹配程度。

(一) ROUGE-N

是一種基於 n-gram 的評估方法，透過統計候選摘要與參考摘要文本中共同出現的 n-gram 個數來衡量相似性。其計算方式中，分母是參考摘要中的 n-gram 總數，分子則為候選摘要與參考摘要中共同出現的 n-gram 個數。

ROUGE-N 可應用於不同粒度的 n-gram，例如 ROUGE-1 衡量單詞級別的重疊 (unigram)，而 ROUGE-2 則衡量雙詞級別的重疊 (bigram)。該方法主要評估候選摘要對參考摘要的覆蓋率，即召回率，但亦可通過計算 F1-Score，綜合考量召回率與精確度，從而提供更全面的評估結果。

(二) ROUGE-L

是一種基於最長共同子序列 (Longest Common Subsequence, LCS) 概念的一種評估方法，這一變體不要求詞語連續匹配，重點在於反映整體文本或句子級別的詞序結構。ROUGE-L 通過計算候選摘要與參考摘要之間的 LCS，進一步評估文本間的相似性。

評估指標包括召回率、精確率和 F1 分數，以綜合衡量生成文本在句子結構上的匹配程度。由於 ROUGE-L 能捕捉語言中的長距依賴關係，因此能更好地評估句子結構上的相似性。

(三) ROUGE-Lsum

這是 ROUGE-L 的一個變體，專為句子級別的摘要任務設計。它透過將文本拆分為多個句子，逐句計算 LCS，並取各個句的平均值以進行更細粒度的摘要評估。這種方法適合在句子層面上進行更精確的比較，有助於提升摘要提取的評估效果。

儘管 ROUGE 在語義相似性與推理能力評估方面存在一定局限性，且對詞序高度敏感，但其透過計算生成文本與參考文本的重疊度，能有效評估候選摘要與標準答案之間的相似性，特別適用於生成型問答系統。ROUGE 的優勢在於減少人工評估成本與主觀性，並且適合於大規模的系統表現評估。實際上，許多研究也證明 ROUGE 可應用於更複雜的資訊擷取與問答任務，將答案視為整體文本的摘要來進行評估（Engelbach et al., 2023）。

此外，BLEU 指標是一種衡量自動生成文本與參考文本之間相似度的工具，通過計算 N-gram 匹配來評估翻譯質量。BLEU 分數範圍在 0 至 1 之間，分數越高表示生成文本與參考文本越相似。

模型微調前的評估結果如表 5 所示，模型微調後的評估結果如表 6 所示，可以發現 Breeze 模型整體表現最好，並且遠高於其他模型的評估結果，接著依序為 Taiwan-LLM 7B、TAIDE、Taiwan-Llama，最後為 Taiwan-LLM 13B。模型微調前後的成效差異並不大，但還是有小幅度的進步，推斷有可能是訓練資料集筆數不多的關係所導致。然而，低分數並不一定意味著內容錯誤，仍需進行後續的人工驗證。

表 5 模型微調前成效的評估

	rouge_score_1	rouge_score_2	rouge_score_L	rouge_score_Lsm	bleuscore
Breeze-7B-Base-v0.1-Q8_0	0.1815	0.1031	0.1702	0.1685	0.0118
Breeze-7B-Instruct-64k-v0.1-Q8_0	0.1495	0.0948	0.1439	0.1439	0.0000
Breeze-7B-Instruct-v0.1-Q8_0	0.1461	0.0789	0.1431	0.1425	0.0092
taide-7b-a.2-q4_k_m	0.0806	0.0452	0.0788	0.0788	0.0000
taide-8b-a.3-q4_k_m	0.0739	0.0333	0.0695	0.0694	0.0000
Taiwan-LLaMa-13b-1.0.Q8_0	0.0700	0.0304	0.0692	0.0661	0.0000
Taiwan-LLM-13B-v2.0-chat-Q8_0	0.0284	0.0173	0.0269	0.0269	0.0000
Taiwan-LLM-7B-v2.0.1-chat-Q8_0	0.1326	0.0674	0.1300	0.1300	0.0000
Taiwan-LLM-7B-v2.1-chat-Q8_0	0.1557	0.0754	0.1462	0.1462	0.0000

表 6 模型微調前成效的評估

	rouge_score_1	rouge_score_2	rouge_score_L	rouge_score_Lsm	bleuscore
Breeze-7B-Base-v0.1-Q8_0	0.2125	0.1437	0.2016	0.2020	0.0100
Breeze-7B-Instruct-64k-v0.1-Q8_0	0.1676	0.0959	0.1573	0.1573	0.0074
Breeze-7B-Instruct-v0.1-Q8_0	0.1896	0.1165	0.1781	0.1784	0.0032
taide-7b-a.2-q4_k_m	0.0848	0.0425	0.0820	0.0812	0.0000
taide-8b-a.3-q4_k_m	0.0752	0.0195	0.0665	0.0657	0.0000
Taiwan-LLaMa-13b-1.0.Q8_0	0.0766	0.0328	0.0742	0.0734	0.0018
Taiwan-LLM-13B-v2.0-chat-Q8_0	0.0487	0.0291	0.0481	0.0481	0.0000
Taiwan-LLM-7B-v2.0.1-chat-Q8_0	0.1346	0.0616	0.1263	0.1263	0.0000
Taiwan-LLM-7B-v2.1-chat-Q8_0	0.1402	0.0678	0.1324	0.1324	0.0000

六、 人類標註標準答案

標記的目的在於建立一微調資料集，其回覆結果為正確，並用以進行模型微調。人工標記者需能夠提供一致而準確的標準答案，本計畫中由具備審計專業知識的專業人員編寫標準答案，以確保微調模型的目標擁有高品質的參考標準。計畫的提示詞設計基於利用檢索結果引導模型根據指令生成答案，同時根據檢索到的相應上下文，提升模型對指令的回答準確性。由於檢索結果涵蓋了先前的檢索內容，有助於模型更全面地理解審計業務的相關內容，進而提高生成回應的相應性。

此外，Gradio 是一款能快速建立使用者互動界面的工具，在此以使用者使用不同頁籤的大型語言模型來進行問答，並可以儲存聊天記錄，未來使用者在與模型對話結束後，可對模型的回應進行好壞評分，以便為後續人類反饋強化學習提供數據支持。

七、 基於人類反饋的強化學習

為了提升大型語言模型的輸出品質，本計畫目的在於透過人類回饋來優化模型，縮小預訓練模型與下游任務之間的性能差距，同改善原有基礎模型過度泛化的情況，並克服其缺乏與審計業務相關專業知識的不足，這種方法有助於提升模型在審計領域的專業性和適用性。

具體而言，計畫將先前用於指令微調的「提示詞範例」與「人類反饋的結果」一同輸入至前一階段經過微調後的初始模型（supervised

finetuning model, SFT)。同時，將其與當前用於強化學習訓練中的模型（Tune model (RL policy)）進行對比，獲得兩個模型的輸出文本。接著，使用獎勵模型（RM）對兩個模型的輸出文本進行評分，判斷其彼此之優劣程度，同時計算「獎勵/懲罰」的程度。最後，以此結果對當前用於強化學習訓練的模型進行參數更新，不斷改進模型的生成效果。此步驟的目的在於透過比較微調後模型的生成文本表現，引導模型進行進一步的優化，確保其生成的內容更符合所需。

使用大型語言模型的過程中，即便給定相同的指令，也可能生成多樣化的回應。由於模型無法自動理解人類偏好，需透過偏好訓練（Preference Learning）進行模型微調，目前大多數做法為透過監督微調（Supervised Fine-Tuning，簡稱 SFT）加上基於人類回饋的強化學習（Reinforcement Learning from Human Feedback，簡稱 RLHF）來進行微調，但透過 RLHF 需處理最佳化問題，首先需要事先收集人類偏好回答訓練集（Preference Dataset），訓練一個獎勵模型（Reward Model），再透過強化學習（Reinforcement Learning）演算法讓大型語言模型學習正確的輸出，在 RLHF 階段中，高度依賴 Reward Model 的反饋，訓練過程較為複雜且不穩定。相較 RLHF，使用 Direct Preference Optimization（DPO）[24]可以不用建立 Reward Model，直接裡用 Preference Dataset 訓練模型，此外，DPO 的訓練過程，模型採用監督式學習（Supervised Learning），而非強化學習（Reinforcement Learning），可提升模型訓練時的穩定度，也可降低訓練過程中所需的運算資源。

Direct Preference Optimization（DPO）是一種用於大規模語言模型（LLM）訓練的技術，其主要目標是直接優化模型的回應品質，以符合用戶的偏好。DPO 的核心思想是透過對模型生成的不同回應進行比較和排序，來直接優化模型的輸出，使其更符合人類偏好。這種方法避免了

傳統強化學習方法中的一些複雜性，例如需要用代理模型（proxy model）來估計回應的品質。以下是 DPO 在 LLM 中的運作原理：

（一）收集人類偏好數據

首先，DPO 需要一個標註數據集，其中包含不同模型回應的比較標註。這些標註可以來自人類評審員，他們根據特定的標準（如回應的相關性、自然性、精確性等）來比較不同回應，並標記出哪一個回應更符合用戶的偏好。

（二）構建偏好模型

基於收集到的標註數據，訓練一個偏好模型。該模型旨在學習用戶的偏好，並能夠評估或預測給定的回應是否符合這些偏好。訓練過程通常使用對比學習（contrastive learning），即模型學習從一對回應中選出更符合偏好的回應。

對比學習是一種自監督學習方法，旨在學習資料表示（representation），使相似的資料點在表示空間中更接近，而不相似的資料點則更遠。這一方法廣泛應用於各種任務中，包括圖像分類、語言模型訓練、推薦系統等。在大規模語言模型（LLM）和人工智慧系統的訓練中，對比學習提供了一種有效的方式來學習高質量的特徵表示，有助於模型更好地理解 and 區分不同輸入。

對比學習的核心原理是透過比較成對的資料樣本來學習有意義的表示：

1. 正樣本對（Positive Pair）：具有相似特徵或關聯性的資料對。例如，相同物體的不同圖片，或同一文本的不同改寫版本。
2. 負樣本對（Negative Pair）：沒有顯著關聯性的資料對。例如，不同物體的圖片，或不同主題的文本。

（三）直接優化語言模型

在 DPO 中，不再需要一個代理模型來估計回應的品質，而是直接在原始語言模型上進行優化。利用偏好模型提供的回饋信號，透過優化目標函數使得模型生成的回應更符合人類偏好。

(四) 迭代訓練

DPO 的訓練過程採用迭代方式進行。隨著更多數據被標註，偏好模型和語言模型會不斷改進，從而提高生成回應品質。每一輪訓練基於新的標註數據和偏好模型的反饋，對語言模型的參數進行調整，以逐步優化模型的表現。

(五) 評估和微調

最終模型需要經過大量的測試，以驗證其在各種情境下的性能。為了修正潛在的偏差或問題，模型可能需要進一步微調，確保其回應的品質和準確性持續提升。

DPO 的優勢在於它能夠直接對語言模型的輸出進行優化，而無需依賴複雜的強化學習框架，從而提高訓練效率。利用對比學習，DPO 有效提升模型在未標註資料上的泛化能力。在缺乏大規模標註數據的情況下，降低了對大量標註數據的依賴。透過強調正樣本對的相似性和負樣本對的差異性，對比學習能夠獲得更加區分性和語義豐富的特徵表示。此外，DPO 直接根據人類偏好進行優化，能顯著提高模型生成回應的自然性、相關性和精確性。為開發智能、用戶友好的大型語言模型提供了高效途徑。

對比學習作為一種強大且靈活的自監督學習方法，可學習資料的內在結構與表示。其在大規模語言模型及其他 AI 系統中的應用，顯著增強了模型的理解能力，從而更有效地處理各種複雜任務。

DPO 的核心方法是基於成對比較從偏好數據中推導出損失函數，旨在根據每對輸入的相對優勢進行分類，而非為每個輸入分配具體分數。

與傳統基於獎勵模型的強化學習（如 RHF，Reinforcement Learning from Human Feedback）有所不同，DPO 專注於比較輸入之間的相對優劣，而非單一評分模型的結果。DPO 的特點在於使用了基於對數（log）概率的比例來表示獎勵，而非具體的數值評分，這使得它在計算上更加冗長。然而，這種設計允許 DPO 能夠表達多個不同的獎勵等級，相比傳統的標量獎勵模型具有更多靈活性和潛力。在應用生成型大型語言模型（LLM）時，這一特點尤為重要，因為隨著這些模型的基礎設施不斷優化，生成式獎勵模型可能最終能夠達到甚至超越標量獎勵模型的性能。

此外，DPO 在減少主觀偏見方面展現了潛力。透過精準理解並學習偏好數據，它能生成更清晰的回答，長期應用中可提升模型的可靠性和客觀性。同時，DPO 提供了一個新的視角來研究如何根據人類偏好進行合理的微調與優化。

獎勵模型在語言模型的訓練中扮演著至關重要的角色。其不僅幫助模型學習生成高質量回應，還能融入人類偏好，讓模型回應更契合使用者需求。然而，設計一個準確且通用的獎勵模型存在挑戰。不同應用場景下偏好各異，因此需要靈活的獎勵機制來適應多元需求。同時，如何避免在訓練中引入偏見，確保生成回應的公正性，也是亟待解決的問題。

隨著越來越多開源模型的發布，在透明度、研究自主性及性能方面，研究者將有更多選擇，減少對少數大型商業模型（如 GPT-4）的依賴。獎勵模型不僅有助於提升語言模型的表現力，還為探索新問題與應用場景開闢了廣闊空間，推動 LLM 更具適應性與創新力。

八、 實驗設計

審計長在立法院備詢時，通常會就上一年度的總決算審核報告進行說明，並回答立法委員對於審核結果和發現的各種問題。然而，涉及範圍相當廣泛，文件相當多，若能夠輔助審計人員依據立法委員的提問，

從中快速的找出審計過程中的重要發現，具體審計意見進行解釋，來說明對審計建議的回應與改進措施，如圖 13。

內容	Q	A short	A detail
<p>(三) 政府為防止國內新型冠狀病毒肺炎 (COVID-19) 疫情擴散，便捷疫調工作，實地簡訊實聯制並建置疫調輔助平臺，惟部分蒐集之個人資料檔案恐落於非公務個人電腦或行動裝置，又相關機關未執行外部稽核及辦理資訊系統防護控制措施等，均存有個人資料洩漏或不當使用等風險，允應督促履責機關研謀因應，完善民眾個人資料保護，2020年新型冠狀病毒肺炎 (COVID-19) 疫情蔓延全球，嚴重特殊傳染性肺炎中央流行疫情指揮中心 (下稱指揮中心) 為維護民眾生活及產業經濟於一定安全條件下逐步恢復正常運作，並兼顧個人資料 (下稱個資) 保護及疫情調查 (下稱疫調) 需求，於109年5月29日訂定「COVID-19 (武漢肺炎) 防疫新生活運動：實聯制措施指引」，規範各公務及非公務機關應依實聯制措施，且對於蒐集之個資僅可保存28日，之後必須刪除或銷毀。110年5月因國內疫情升溫，為減少實聯制紙本填寫之接觸，指揮中心實地「簡訊實聯制」，並請國家通訊傳播委員會 (下稱通傳會) 協助辦理疫調輔助平臺服務租賃採購案 (契約總金額1,999萬餘元，中華電信股份有限公司企業客戶分公司得標)，衛生福利部疾病管制署 (下稱疾管署) 配合驗收及經費支付作業。「疫調輔助平臺」於110年7月26日上線，介接五大電信 (中華電信、台灣大、台灣之星、亞太及遠傳) 實聯制簡訊資料庫，提供熱區地圖、確診者活動軌跡等查詢功能 (圖1)，可協助地方政府疫調人員掌握個案相關活動史及匡列接觸者等，達到防疫之目的。111年4月27日指揮中心考量COVID-19疫情進入社區流行階段，宣布取消營業場所及公共場所簡訊實聯制措施，截至同年5月26日簡訊實聯制實地期間累積發送約47.8億則簡訊已全數刪除完畢，經查行政院資通安全處 (下稱資安處)、國家發展委員會、通傳會及疾管署辦理簡訊實聯制資料管理作業情形，核有：1. 疫調輔助平臺111年1月1日至5月25日止，計有1萬7,513筆依號使用紀錄，其中849筆為使用者於非公務環境下，以其網路服務登入平臺，相關個資檔案恐落於非公務個人電腦或行動裝置，不易管控而發生個人資料被竊取或洩漏等風險；2. 指揮中心自111年4月27日起取消營業場所及公共場所簡訊實聯制措施，惟該平臺自111年4月28日至5月25日止，仍提供簡訊實聯制檔案下載功能，並有43筆下載紀錄，存有個人資料洩漏或不當使用等風險；3. 通傳會於查訪電信業者辦理實聯制資料刪除過程，未檢核原始檔案、備份伺服器及電子郵件等資料貯存點，未能確保確診者相關個資已完全刪除；4. 疫調輔助平臺包含確診者疫調軌跡並介接簡訊實聯制等相關資料，各使用機關、平臺廠商及五大電信業者於資料匯入、透過電子郵件等營運傳遞、備份或下載過程均可能另行留存防疫個資，惟資安處、通傳會及疾管署尚未辦理外部稽核，無法確保防疫資料已執行刪除或銷毀；5. 通傳會外置之疫調輔助平臺可下載簡訊實聯制及確診者軌跡等個資，惟委外契約未訂定廠商應遵之資安管理責任及機制，該會及疾管署亦未依資通安全管理法相關子法規定辦理資訊系統防護控制措施等情事，經品請行政院院督所屬履責機關研謀因應。據後：1. 指揮中心已請防疫相關機關落實個資管理及刪除作業，後續將宣導防疫個資匯點及刪除作業；2. 指揮中心已請通傳會提前終止疫調輔助平臺契約，該平臺目前僅供靜態統計資料查詢；3. 指揮中心已於111年6月6日請通傳會督導委外廠商執行疫調輔助平臺儲存實聯制個資銷毀作業；4. 指揮中心已於111年6月6日函請市縣政府依行政程序銷毀疫調輔助平臺下載之離線資料，並轉知轄下單位配合辦理，另復檢核結束後，資安處將會同國家發展委員會，擇重點機關檢視確認防疫個資處理情形；5. 爾後將持續宣導請機關務必注意整體資安防護，並落實法遵應辦事項。</p>			<p>政府實地簡訊實聯制並建置查核情形</p> <p>一、 執行概況：通傳會建、介接五大電信實聯制簡訊指定日期區間簡訊實聯制之診者活動軌跡疫調、熱區地、查核發現：(一)通傳、辦理資料刪除過程，未檢核、分防疫機關透過非公務環境、或於實聯制措施退場、下載實聯制個資刪除或銷毀、機關改善情形：(一)調輔助平臺契約，平臺已關、並督導委外廠商完成平臺、介安處已將電子防疫個資之輸入等必檢核項目。</p>

圖 13 備詢之問答範例

我們透過 RAG 的方法讓大型語言模型去查詢知識內容進而能回答出結果，如圖 14 與 15。大型語言模型能夠快速處理大量文本資料，根據政府政策文件、法律條文和相關報告，自動生成清晰、精確的問答草稿，且能夠從大量的文章中挑選出是從哪份資料而來，大型語言模型可以作為一個訊息檢索工具，快速查找與備詢議題相關的背景訊息。如圖 14 中顯示回答從第 23 篇資料中得來。這些草稿可以作為立法委員在質詢政府官員時的參考，確保問答的內容具有邏輯性和針對性當立法委員需要了解某個議題的歷史背景、數據支持或國際比較時，大型語言模型可以即時提供相關的數據分析和背景報告。此外，大型語言模型還可以模擬政府官員的回答，幫助委員進行預備演練，備詢問答通常會涉及即時應答的環節，大型語言模型可以根據過往政府官員的公開答辯紀錄和政策文件，預測可能的回應方式。這種模擬對話的方式可以幫助委員提前準備應對策略，提高問答環節的應變能力。



圖 14 大型語言模型的 RAG 備詢之問答結果



圖 15 大型語言 GPT 模型的 RAG 備詢之問答結果

表 7 備詢之問答模型微調前成效的評估

	GPT-4	Breeze-7B-Base-v0.1-Q8_0	taide-7b-a.2-q4_k_m	Taiwan-LLaMa-13b-1.0.Q8_0
rouge1	0.0572471	0.156455059	0.095157457	0.093548387
rouge2	0	0.030061892	0.045078577	0.022988506
rougeL	0.05620544	0.156541219	0.095157457	0.093548387
rougeLsum	0.05147059	0.146326165	0.093137255	0.095400239

表 8 備詢之問答模型微調後成效的評估

	GPT-4	Breeze-7B-Base-v0.1-Q8_0	taide-7b-a.2-q4_k_m	Taiwan-LLaMa-13b-1.0.Q8_0
rouge1	-	0.116559829	0.135743464	0.130189466
rouge2	-	0.0375	0.047916667	0.022988506
rougeL	-	0.120138889	0.133986928	0.127371917
rougeLsum	-	0.118990385	0.135416667	0.131085523

表 7 及表 8 分別展示了在微調前後，不同問答模型在 Rouge 指標下的表現差異，說明各模型（包括 GPT-4、Breeze-7B、taide-7b、Taiwan-LLaMa 和 Taiwan-LLM 等模型）在微調前後的表現變化。Rouge 指標包括 Rouge-1、Rouge-2、Rouge-L 和 Rouge-Lsum，分別用來衡量模型在詞彙、詞組和長文本上的相似度。在微調前，從 Rouge-1 指標來看，Breeze-7B-Base-v0.1-Q8_0 是表現最好的模型，高於其他模型，表明它在詞彙層面的匹配上表現最佳。相比之下，GPT-4 的 Rouge-1 得分表現相對較差，這也顯示其對於審計內容無法完整的抓取其語意。然而對於其他模型中，Taiwan-LLaMa-13b-1.0.Q8 和 taide-7b-a.2-q4_k_m 表現則與 GPT-4 類似。在 Rouge-2 指標上，這個指標主要衡量詞組匹配，Breeze-7B-Base-v0.1-Q8 再次領先，這表明 Breeze-7B 的詞組匹配能力依然保持領先。在

Rouge-L 指標中，該指標著重於長句子或文本的結構匹配，Breeze-7B-Base-v0.1-Q8 的得分仍然是最高的，這意味著 Breeze-7B 在生成與參考文本結構相似的句子時，保持了領先的表現。GPT-4 的 Rouge-L 得分為 0.0562，與其 Rouge-1 的表現相似，仍然處於中低水平。與此同時，Taiwan-LLaMa 和 taide-7b 的 Rouge-L，顯示出它們在句子結構上的匹配度較為接近。最後，在 Rouge-Lsum（對整體文本進行的評估）指標上，Breeze-7B-Base-v0.1-Q8 依然表現最為突出，這說明它在生成與參考答案整體語義一致的文本上具有優勢。總結來看，Breeze-7B-Base-v0.1-Q8 在各個 Rouge 指標上的表現都明顯優於其他模型，特別是在 Rouge-1 和 Rouge-L 指標上，顯示出其優異的文本匹配能力，在處理詞彙和句子結構匹配時的能力非常強。相比之下，GPT-4 在這些指標上的得分較低略顯弱勢，taide-7b 和 Taiwan-LLaMa 的表現相當。這些結果說明了在微調前，Breeze-7B 基於其強大的詞彙匹配和句子結構能力，能夠生成更符合參考文本的答案，而其他模型則可能需要進一步的微調來提高其在詞組和長文本結構上的表現。微調後幾乎所有模型的 Rouge 指標都有不同程度的提升，但 Breeze-7B-Base-v0.1-Q8 略有下降，但其整體表現仍然保持在較高水準。taide-7b-a.2-q4 和 Taiwan-LLaMa-13b-1.0.Q8 的 Rouge 指標在微調後也有所提升，甚至高過於 Breeze-7B-Base-v0.1-Q8，表明其在句子結構上的匹配度有了大幅改善。這表明透過微調，部分模型能夠顯著提升其在問答生成中的表現，對於生成與參考文本更匹配的答案展現了更強的能力。由於 GPT4 乃是商業大型語言模型，微調需要額外付費，因此在本研究並未進行微調的動作。

由於使用 ROUGE 指標主要是針對生成回應與標準答案的重疊程度進行評估，該評估方式相對嚴格，尤其在生成的文句不一定完全符合正確答案的字詞情況下。因此，我們引入了人工評分，以評估生成文本的流

暢性和語意完整性。我們選取了 12 個問題（如附件所示），並邀請 32 名審計人員進行評分，每題的滿意度範圍為 1 至 5 分，1 分表示非常不滿意，5 分則表示非常滿意。為確保評分結果的準確性和一致性，要求審計人員根據生成文句的自然流暢度、語意的一致性、以及回應的相關性進行打分。這種評估方式，不僅補足了 ROUGE 指標在某些語義細節上的不足，也能更全面地反映生成模型在實際應用中的表現。

表 9 備詢之問答模型微調前後的人工評分

	題號	各模型平均分數		
		Breeze_base	Taide_7B	Llama_13b
LORA	1	3.4	2.8	2
	2	3.2	2.7	2.5
	3	2.4	3.5	3.1
	4	3.6	2.8	2.5
	5	3.4	2.6	2.1
	6	3.2	2.6	2
Base	7	3	2.7	2.2
	8	2.2	2.8	3.1
	9	3.5	3.6	3.2
	10	3.2	2.5	2
	11	3.3	3.1	2.7
	12	2	3.3	2.5

表 9 顯示了三種模型（Breeze_base、Taide_7B 和 Llama_13b）的平均分數，針對不同題目以及兩種狀態（Lora 微調後 和 Base 微調前）進行比較。從結果來看，Breeze_base 模型在 12 個題目中有 8 題為第一名的成果，表現優異，尤其是在 Lora 模式下獲得了最高分數 3.6。而 Taide_7B 則在第 3 題表現突出，得到了 3.5 的高分，顯示其在某些題目上具有一定的優勢。不過，Taide_7B 在其他題目的得分相對平均，沒有顯著的高低波動。相比之下，Llama_13b 的表現略遜一籌，在多數題目中的得分較低。從整體來看，Breeze_base 在各題目中的表現相對穩定，無論是在 Lora 還是 Base 模式下，都能維持較高的分數，而 Taide_7B 儘管在某些題目上具有優勢，但整體表現略遜於 Breeze_base，Llama_13b 在大部分題目上的表現則較為一般，分數偏低。

肆、審計機關導入及運用生成式人工智慧技術策略之研析

從表 7 可以觀察到 Breeze-7B-Instruct、taide-7b、Taiwan-LLM-7B 及 OpenAI GPT-4 等四個不同模型在評估指標上的審計備詢之問答表現差異。在對繁體中文問答任務進行分析時，模型架構的選擇至關重要，在法律與審計相關文件上 Breeze-7B 在 LoRA 微調前後的表現均優於 TAIDE-7B 和 TAIWAN-LLM-7B，特別是在 ROUGE-1 和 ROUGE-L 的指標上，這顯示 Breeze-7B 對詞序和整體文本結構的理解能力較強，其是從 Mistral-7B 演進而來的專為繁體中文所設計，或許是由於 Mistral-7B 的基座模型的能力相較於 llama 就已經具有高效能的優勢。TAIDE-7B 的表現相對較差，這可能表明它在捕捉文本內的複雜關聯性方面存在一定局限。TAIWAN-LLM-7B 則介於兩者之間，顯示它在某些指標上具備競爭力。由於繁體中文的大型語言模型皆有其優勢，因此綜合各模型的特點和表現，能夠更有效地解決不同應用場景中的需求，透過結合這些模型的長處，可以在具體任務中提升文本生成的準確性和一致性，並且針對不同的文本層次進行優化，使得繁體中文的大型語言模型能更好地服務於高精度領域的應用。

繁體中文語言模型都能夠有效利用外部知識庫來提供上下文背景知識，這使得模型在回答問題時，能夠更好地理解 and 生成相關內容。透過引入額外的知識源，RAG 有助於減少人工智慧幻覺，即模型生成不準確或與事實不符的信息。實驗結果表明，RAG 不僅能夠加強模型的問答能力，還能在未進行 LoRA 微調的情況下顯著提升性能。在當前人工智慧領域中，LLM 結合 RAG 技術展現了巨大的應用潛力，特別是在需要從大量數據中檢索相關資訊並生成回應的場景。然而，在 RAG 系統的優化過程中，仍面臨一些挑戰，例如檢索器可能返回不準確或不相關的資料、LLM 生成基於錯誤或過時資訊的答案等。以下是 RAG 系統的挑戰及其相應優化策略：

一、檢索過程缺失內容導致回答不全面

在檢索過程中，由於大型語言模型存在輸入字符數量的限制，需對文本進行切割。然而，這可能導致某些重要資訊未被檢索到，或生成的答案未能全面涵蓋問題的核心內容。例如：系統考慮了排名最高的前 K 個文檔，而實際上關鍵答案可能隱藏在後續的文檔中。過大的文本切割可能導致模型無法準確檢索到所需資訊，而過小的文本切割則可能破壞上下文，我們需要適當的擴展上下文來提高檢索完整性，確保回應能覆蓋所有重要的資訊；亦或是透過加強檢索算法，例如當模型接收大量文本時，前後段的內容會比中間段落更容易被注意到，這與自然語言處理算法中的「注意力衰減」有相似之處。因此，將最相關的文檔安排在開頭或結尾，不僅能幫助模型更快定位正確資訊，還能避免模型陷入中間部分不重要的細節中，或是引入更多訓練數據以改善模型對相關性的判斷。

二、語意理解不確定

預訓練的嵌入模型雖然能夠涵蓋廣泛的語義空間，但未必能完全適應特定任務需求。透過對嵌入模型進行微調，可以更精確地映射特定領域的語義關聯，加強模型對關鍵字或概念的理解與匹配能力，從而提高檢索到的文檔與查詢之間的匹配度。

三、生成的答案過於模糊

可以透過調整模型的細節生成機制，或依據不同的應用場景設計回答的精細度，並明確指示模型在不確定答案時直接回應「我不知道」。這種設定能有效減少錯誤資訊的傳播，如果知識庫包含偏見或不準確的信息，則可能會對生成的句子產生負面影響。

大型語言模型經過 LoRA 微調後，ROUGE 指標上的表現普遍顯著提升。這些表明 LoRA 微調對大型語言模型的影響顯著，這可能表明 LoRA 微調改進了模型對詞組合的捕捉能力，使生成的文本更接近參考答案，從而提升模型在生成過程中的精確度和相關性，進一步增強問答任務中的表

現。然而，模型架構不同，對 LoRA 微調的適應性也存在差異。因此，這為不同模型設計針對性的微調策略尤為重要，需要根據模型的初始性能和特性制定相應的優化方案。

在模型評估方面，除了依賴 ROUGE 和 BLEU 等技術指標外，還應觀察模型生成的語句的質量，以評估其對領域知識的掌握程度。此外，也可引入用戶體驗評估，收集實際審計人員的反饋，幫助識別模型在具體應用中的表現是否達到預期目標，並通過人機協同優化，實現根據用戶需求進行即時個性化調整。這些措施有助於提升大型語言模型與生成式檢索增強技術在審計業務中的實用性與效能，進一步增強系統的準確性與用戶體驗。

當前大型語言模型架構種類繁多，從 OpenAI 發布的 GPT，到 Meta AI 推出的開源模型 LLaMA，以及基於 LLaMA 發展的其他語言模型。在繁體中文領域，也出現了多種模型架構。這些模型在問答任務中的適應性及其作為問答系統的基礎模型的表現，是本研究優先探討的核心問題。本研究針對不同任務測試多種大型語言模型，未來可考慮融合多模型的優勢，也可以透過進一步調整模型參數和優化訓練數據集，針對具體應用場景進行微調。特別是對繁體中文語言特性的精細訓練與測試，將有助於改進模型在實際應用中的表現。

LoRA 微調的優勢在於不需要龐大的訓練數據集，且訓練時間和成本相對較低。建議先進行小規模實驗，例如使用 100 到 1,000 筆數據進行短期訓練，以快速驗證模型是否有顯著改善。這樣的實驗可以幫助減少不必要的時間與資源投入，並觀察模型是否具備進一步調整的潛力，使其在特定領域中更為有效與精確。然而，微調面臨以下挑戰：

一、**是否擁有足夠的運算資源**：微調成本高，且需要大量的硬體資源。

租用雲端服務，如 GCP、Azure、AWS 等，是一種可行的方式，但將資料上傳到雲端存在一定風險。微調通常是一個長期的試錯過程，

這類工作通常只有大公司或政府機構具備足夠的計算資源和專業知識來承擔，因此，對他們來說這是一個可行的選項。

二、是否擁有足夠的訓練數據：擁有足夠且高品質的任務相關訓練數據時，微調可以顯著提升模型性能；若訓練數據不足或質量不高，微調則可能無法達到預期效果。即使經過微調，大型語言模型的表現有時仍只能算是中等偏下，甚至某些情況下，微調後的表現還不如原始模型。（可見表 7 與表 8）。在微調時能設計有效的提示詞，對非專業使用者來說將大有幫助。微調使解決方案更加個性化，RAG 提升了問題處理的準確度和相關性，而提示詞工程則增強了系統的易用性和使用者體驗。

在微調過程中，我們需要確保微調數據的高品質，避免過度擬合，並兼顧模型通用性與專業性的平衡，成功的微調不僅能提升模型的實用性，還能夠為使用者提供針對性的解決方案，這在涉及複雜專業領域時尤為重要。

伍、結論與建議

過去，使用者通常需要手動瀏覽知識庫或依賴關鍵字搜尋，這不僅耗時，還可能因關鍵字不同而漏掉重要資訊。然而，透過對話式系統，使用者可以直接提出問題，系統會根據查詢自動整合多個資料來源，提供即時且準確的答案。

這種方式的優勢在於：

- 一、**即時與便利性**：使用者無需再花費大量時間搜尋知識庫，只需提出具體問題，系統即可迅速提供相關資訊。
- 二、**整合性**：系統能整合來自不同來源的資料，如文檔、報告和數據庫，生成全面的解答，而不僅依賴於單一資料來源。
- 三、**準確性**：透過自然語言處理技術，系統能理解問題的意圖，並提供更符合使用者需求的答案，減少因關鍵字搜尋帶來的潛在誤解或遺漏。
- 四、**互動性**：使用者能與系統進行更自然的交流，根據系統的回答進行追問，進一步探索相關知識，提升知識查詢的深度與廣度。
- 五、**增強的學習體驗**：系統不僅提供查詢與回答功能，還能成為使用者學習和獲取新知識的工具，促進知識的傳遞與擴展。

創新的知識管理方式不僅提升了查詢的效率和準確性，還讓知識獲取過程變得更流暢、人性化，為使用者帶來全新的體驗。這類產品開發仰賴於大型語言模型的輔助，傳統的 AI 產品往往需要投入大量資源來收集數據和進行模型訓練，過程中充滿未知變數，最終訓練出的模型也未必達到預期性能，然而，使用如 GPT-4 等經過大規模訓練並具備強大表現的通用語言模型，不僅能即時提供高效能，還能避免不必要的迭代嘗試，節省大量開發時間和資源，使產品能更快推向市場，並提供穩定的技術基礎。未來若有性能更高的模型，亦可靈活替換和升級，這種策略不僅能保持技術

領先性，還能減少早期技術選擇所帶來的潛在限制。因此，先使用當前模型進行產品開發，能有效降低技術風險，並專注於探索需求端的不確定性，從而打造出更具競爭力的解決方案。

儘管大型語言模型具備強大的通用能力，但在特定領域的應用中仍面臨挑戰。當模型對某些專業領域的文本不熟悉，或對推理任務掌握不足時，能否區分模型的回答是因知識匱乏還是由於「幻覺現象」（即模型生成錯誤或虛構資訊）尤為關鍵。因此，本研究探討了 RAG 和 LoRA 之間的細微差別，以深入了解高效微調在大型語言模型應用中的重要性。

RAG 技術結合外部檢索系統與生成模型，有效解決模型知識不足的問題。透過 RAG，模型能在不依賴大量計算資源、時間或數據的情況下，適應特定任務需求。具體來說，當用戶提出查詢時，RAG 會首先從龐大的外部知識庫中檢索相關資訊，這些知識庫可能包括文檔、文章或其他相關文本來源。然後，系統將檢索到的資訊輸入大型語言模型中，根據查詢和檢索到的上下文生成精確的回應。這種方式不僅解決了模型知識的局限，還使用戶能靈活切換不同的知識庫，使模型能適應各種領域與任務。因此，RAG 有效平衡了模型的通用性與領域專精需求，為特定領域應用提供了更精準且靈活的解決方案。

如果 RAG 無法完全解決模型對特定領域文本的不熟悉問題，我們可以透過提示詞進一步輔助。例如，當我們希望獲得繁體中文的回答時，使用「正體中文」作為提示詞通常能提高回應的精準度。然而，提示詞的效能取決於語言模型本身的能力，並非所有模型都能有效運用提示詞來達成理想結果，前提是該大型語言模型具備強大的通用能力。例如，針對風險管理，我們可以用「預測未來風險事件的影響」或「評估當前資產負債表的風險」這類提示詞來引導模型生成更準確的回應。然而，當應用涉及特定

領域數據，且需要高度的準確性與對專業術語的理解時，模型的通用能力可能不足，此時微調就變得非常必要。

若我們希望模型以特定的風格、語氣或格式進行回應，或展現特定行為特徵，LoRA 微調可以幫助實現這些需求。微調並非重新訓練整個模型，而是調整模型部分參數的權重，通常只需要少量的標註數據，讓模型能快速學習特定任務的知識，並在特定推理場景中正確回答問題，從而提升訓練效率，快速適應不同應用場景。

在審計機關導入及運用生成式人工智慧來精進審計業務的過程中，本研究提出以下建議：

一、建議審計工作應用 AI 技術，可優先參採使用以繁體中文預訓練、可落地調校之開放本土型語言模型，以尋求低成本、高安全性、高接受度等利基，並支持國家 AI 科技發展。

導入 AI 技術於審計工作之前，應先確立使用 AI 技術的方式及目標，並就何種審計流程和具體業務內容適合藉由 AI 技術來提升作品質或效率，予以詳細評估。此外，應避免以大型語言模型連續、大量產製文字並直接引用，以免偏離 AI 作為輔助工具的定位。AI 技術應主要用於減少審計業務重複性工作，並優化知識檢索的品質與效率。例如，使用大型語言模型梳理、分析巨量文本數據，透過機器學習來識別潛在風險或異常，或建立智能化數據分析模型以提升分析精確度和效率，從而使審計人員能將精力集中於高階分析及決策層面。

審計業務具有高度專業性、敏感性及獨特性，審計人員日常工作需處理業務數據資料鉅量且龐雜，資料來源廣泛、種類繁多，審計機關導入 AI 技術處理前述資料前，宜全面性客觀比較各種大型言語模型特色，覈實評估其成本效益，並衡酌審計業務應用情境重點，擇優採用。如採用開源大型語言模型，須考量機關自行購置相關軟、硬體

設備，及維護營運之相關成本。如採用非開源付費大型語言模型，則應注意應用生成式 AI 處理機敏資料時，可能衍生法律、道德及資訊安全等問題，必須審慎考量資料保密性，並恪遵隱私及資訊保護相關法令規範，除一般語言模型使用費以外，實際推廣時往往另須考量資訊安全保障額外增加之成本。審計機關應確保 AI 相關應用技術於提高工作品質、改善工作效率、簡化工作流程或是降低成本等環節，可提供足夠價值，唯有 AI 相關應用技術預期能明顯提升審計工作運作效能，並且達到令人滿意之投資回報，始能合理地推動導入。

當前大型語言模型多以簡體中文文本訓練，欠缺繁體中文資料，因此回應中常出現偏誤。有鑑於此，行政院國家科學發展委員會日前已完成台灣繁體中文大型語言模型「TAIDE」（Trustworthy AI Dialog Engine），開放政府機構及民間企業自由應用，以呼應外界對臺灣本土化語言模型需求，並確保本土型生成式 AI 之可信任性及適用性，「TAIDE」具有繁體中文文本生成能力、具本土文化特色、高度可信任性、低成本等多項特色。基於政府一體、跨機關資源整合、跨機關合作樣態多元化等原則，建議審計機關導入 AI 技術，優先參採使用以繁體中文預訓練、可落地調校之開放本土型語言模型（如 TAIDE、Breeze 等），尋求低成本、高安全性、使用者高接受度等 AI 技術應用之利基，有助於未來各種審計工作應用情境得以靈活運用並持續優化，一方面可體現審計業務創新應用，另一方面可支持國家科技發展。

二、建議布署允當 AI 運算力資源，支援落地大型語言模型之調校、檢索擴增生成等需求，俾運用 AI 技術創新審計作業模式，因應變革業務環境。

政府機關若倉促全面導入 AI 技術，極可能因為技術尚未成熟、使用者抗拒及數據隱私安全等問題，導致適得其反的結果。因此，生成式 AI 技術的導入宜採取循序漸進策略，將 AI 技術分為多階段引入，每個階段逐步測試和優化，以確保技術穩定性和可行性。首先，可選擇具有明確業務需求的審計流程進行試點應用，測試技術的穩定性和效果，然後逐步擴展其應用範圍。試點項目應著重測試生成式 AI 在性能、準確性以及對審計業務的影響，特別是生成內容是否符合審計專業需求、語言模型之檢索速度和準確性、系統於高負載的表現，測試結果將回饋作為後續技術優化之參考數據。

在測試過程中，除賡續蒐集使用者回饋意見外，尤應關注 AI 生成內容之質量是否符合實際要求，透過反饋系統進行優化，特別在提示詞設計、檢索算法的微調、人機交互介面等方面，以確保 AI 技術應用之實效性及良好的使用者體驗。這種反覆調整過程，可確保生成式 AI 技術在實際審計工作中的價值，並逐步擴大其應用範圍。

針對審計機關初期 AI 技術應用範疇，並配合前項建議採用落地安全 AI 語言模型之各種調校作業實際需求，建議審計機關可建置以下 GPU 主機規格：2U 機架式雙路人工智慧運算伺服器（GPU Server），配備 L40S 二個高效能 GPU（含同等級或以上）、CPU 處理器（Intel i7 含同等級或以上）、全閃存硬碟（All Flash Disk）之硬體設施。前開規格配置可有效提升運算效能，滿足審計機關在大型語言模型知識管理應用中的實際需求，確保系統運行穩定性與安全性。

值此 AI 世代，人工智能技術發展一日千里，未來大型語言模型勢將越發龐大、複雜，展望未來，審計機關導入生成式 AI 技術不應僅依賴機關內部技術團隊，還需建立跨部門多元合作夥伴關係，引入外部活力與資源，舉如：可與國家高速網路中心等機構合作，以獲取

專業支援和創新能力，共同推動 AI 技術實效應用。其次，導入 AI 技術後，仍需持續監控系統效能與準確性，尤其是初期運行階段，應建立妥適之反饋機制，使審計員可隨時回報系統不足之處，反饋對於持續優化模型和系統功能至關重要，透過蒐集、分析使用者反饋，可及時識別風險並尋求解決，確保系統不斷提升性能並滿足使用者需求。此外，持續監控及優化過程可增強審計人員對於創新技術之信心，促進審計人員更有效率地應用生成式 AI 技術於實際工作，從而發揮最大潛力。

三、建議針對本案研究成果規劃 AI 應用技術課程，包括詳細解說落地大型語言模型調校工作之基礎理論、調校資料格式、調校參數、操作步驟等，並演示模型調校成果。

檢索擴增生成技術 RAG (Retrieval-Augmented Generation) 表現之良窳，高度依賴於資料質量、準確性、來源可信度及更新時效性，因此，導入 RAG 技術時，首先需先對現有知識庫進行全面評估及管理，以確保資料完整、準確性，並得及時更新。良好的資料基礎不僅能夠提升系統表現，還能有效減少風險和錯誤的發生。因此，在實施 RAG 技術時，推動數據標準化與清理流程乃是關鍵的第一步，此舉將有助於確保審計資料品質，並支持系統在後續應用中表現穩定。此外，應建立完善的資料治理框架，以支持高效的資料管理，包括：定期更新資料庫的機制，確保資料及時性與相關性；進行資料清理，移除雜訊、重複資訊、修正並標註錯誤數據等工作，可確保系統準確性及可靠性，上開措施將有助於 RAG 技術順利運行，使其在審計流程中能夠發揮更大的價值。

審計機關在導入生成式 AI 技術後，審計人員需要學習如何使用和解釋這些技術，並理解如何將生成結果與實際審計業務結合，為確

保導入 AI 技術計畫得以成功執行，應先培養審計人員數據素養與技術技能。其範疇應含括理解 AI 生成結果的原理與應用，還需定期提供培訓與專業發展課程，幫助審計人員提升數據素養與技術應用能力，保持與技術的持續發展接軌。具體來說，審計人員需要掌握正確使用生成式檢索增強技術的能力，培訓內容應涵蓋系統的基本操作、生成結果的解釋與應用，以及根據具體需求調整提示詞的技巧，這些技能幫助審計人員靈活應對不同業務情境，並提升其在實際業務中的應用能力，從而實現技術與業務的深度融合。

針對本研究結果有關 RAG (Retrieval-Augmented Generation) 及 Fine-tuning 調校技術與程式運行，建議就本案研究過程相關實例，分階段規劃相關課程 (按: Fine-tuning 課程部分應配合機關運算力布署進度進行)，俾將本案針對落地語言模型相關調校，順利技術移轉給具有發展潛質的審計人員。課程可涵蓋 RAG 及 Fine-tuning 的基本原理、實作技巧、調校流程，以及在審計工作中具體可應用的案例，旨在提升審計人員對 AI 技術的理解與實務操作能力。此外，課程中可設計實作環節，使參與者能在實際環境中應用所學，強化他們在數據處理、分析及報告生成方面的能力，應對未來挑戰。

以上各項建議，旨在協助審計機關順利導入生成式 AI 技術，並在審計業務中最大化其利用價值，透過科學管理、靈活的技術框架以及穩固的數據基礎，協助審計機關實現技術與業務流程之深度融合，進而提升審計工作品質與效率。導入生成式 AI 技術有機會顯著提升審計效率及準確性，惟同時亦面臨數據管理、合規性、技術整合及人力資源等諸多問題與挑戰，未來，AI 技術在審計業務的應用，勢必仰賴技術發展成熟度及審計人員操作技能，透過妥適規劃，逐步實施知識庫建構、微調模型動化等措

施，審計機關可達到更高的工作效能與精準度，此外，跨部門合作亦可確保系統長期運行穩定可靠，靈活應對變化的業務需求。

參考資料

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
2. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023). GPT understands, too. *AI Open*.
3. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
4. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
5. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J. B., ... & Mustafa, B. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
6. Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
7. <https://www.anthropic.com/news/claude-3-family>
8. Zheng, Q., Xia, X., Zou, X., Dong, Y., Wang, S., Xue, Y., ... & Tang, J. (2023). Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.
9. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... & Sun, L. (2024). Sora: A Review on Background, Technology, Limitations, and Opportunities

- of Large Vision Models. arXiv preprint arXiv:2402.17177.
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., ... & Rombach, R. (2024). Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206.
 11. <https://www.patronus.ai/blog/introducing-copyright-catcher>
 12. <https://blog.cloudflare.com/firewall-for-ai>
 13. <https://www.suno.ai/blog/v3>
 14. <https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/sima-generalist-ai-agent-for-3d-virtual-environments/Scaling%20Instructable%20Agents%20Across%20Many%20Simulated%20Worlds.pdf>
 15. Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., ... & Liu, X. (2023). DocLLM: A layout-aware generative language model for multimodal document understanding. arXiv preprint arXiv:2401.00908.
 16. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>, 13, 1.
 17. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
 18. Zhao, Y., Lin, C. Y., Zhu, K., Ye, Z., Chen, L., Zheng, S., ... & Kasikci, B. (2023). Atom: Low-bit quantization for efficient and accurate llm serving. arXiv preprint arXiv:2310.19102.
 19. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
 20. Pandya, K., & Holia, M. (2023). Automating Customer Service using

- LangChain: Building custom open-source GPT Chatbot for organizations. arXiv preprint arXiv:2310.05421.
21. <https://blog.research.google/2024/01/introducing-aspire-for-selective.html>
 22. <https://chat.lmsys.org/>
 23. Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
 24. <https://www.cnbc.com/2023/05/25/jpmorgan-develops-ai-investment-advisor.html>
 25. Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., ... & Liu, X. (2023). DocLLM: A layout-aware generative language model for multimodal document understanding. arXiv preprint arXiv:2401.00908.
 26. <https://www.dentons.com/en/about-dentons/news-events-and-awards/news/2023/august/dentons-to-launch-client-secure-version-of-chatgpt>
 27. <https://blog.google/products/search/generative-ai-search/>
 28. <https://blog.google/products/shopping/shopping-graph-explained/>
 29. <https://twc.ai/2024/01/17/%E5%AF%A6%E8%B8%90ai-2-0-%E5%8F%B0%E7%81%A3%E5%A4%A7%E5%93%A5%E5%A4%A7%E4%BC%81%E6%A5%AD%E6%9C%8D%E5%8B%99%E6%94%9C%E6%89%8B%E5%8F%B0%E6%99%BA%E9%9B%B2%EF%BC%8C%E5%8D%87%E7%B4%9A%E6%99%BA%E8%83%BD/>
 30. Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.
 31. <https://nuhsplus.edu.sg/article/ai-healthcare-in-nuhs-receives-boost-from-supercomputer>
 32. <https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5402.htm>

33. <https://pib.gov.in/PressReleaseDetail.aspx?PRID=2012355>
34. <https://www.smartnation.gov.sg/nais/>
35. <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-da-advanced-audit-analytics.pdf>
36. https://www.ey.com/en_gl/services/audit/technology/helix
37. <https://kpmg.com/xx/en/home/services/audit/kpmg-clara.html>
38. <https://www.pwc.com/m1/en/events/socpa-2020/documents/gl-ai-brochure.pdf>
39. <https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-launches-innovative-dartbot-internal-chatbot.html>

附件一：人工評分表

審計部生成式AI回答評分表

請各位同學們動動發財的小手，一起替各選項的回答準確度進行評分吧、(๑˘▽˘)ノ

1. 1.對於企業發生重大個資外洩案件，審計部有沒有瞭解機關辦理個資外洩的問題？

A.是的，審計部有對各機關辦理個資外洩的狀況進行查核，並提出審核意見。在2021年的報告中，審計部指出了三個問題：第一個問題是個資法對於企業個資外洩通報規範不足；第二個問題是部分機關對企業個資外洩案件未落實行政檢查；第三個問題是數位部未能引導企業善用該部監管之台灣電腦網路危機處理暨協調中心資源，致情資無法即時傳遞。

B.是的，審計部對企業發生重大個資外洩案件有進行相關的查核和研究。在審核意見中，審計部指出了三個主要問題：(一)個資法對企業個資外洩通報規範不足；(二)部分機關對企業個資外洩案件未落實行政檢查；(三)數位部未能引導企業善用該部監管之台灣電腦網路危機處理暨協調中心資源，致情資無法即時傳遞與處理。

C.根據所提供的三份審核意見，並沒有直接討論到審計部是否瞭解機關辦理個資外洩的問題。然而，可以推測的是，審計部的職責之一是監督政府部門的運作，包括個人資料保護法的遵循情況。因此，審計部可能已經對相關機關進行了績效評估或稽核工作，以確定他們是否符合法律規定和要求。

D.No. 審計部沒有瞭解機關辦理個資外洩的問題。Yes. 審計部瞭解機關辦理個資外洩的問題。Not applicable. 不適用。

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2. 2.電商個資外洩事件頻傳，審計部查核情形? 對於企業發生重大個資外洩案件，審計部有沒有解機關辦理個資外洩的問題?

A.目前有很多電商個資外洩的事件，審計部已經開始調查了，他們正在加強查核各機關資安相關事項的辦理情形，以確保個人資料的安全。目前還沒有具體的結果，但審計部會持續進行調查和查核，以確保個資的安全。More Helpful Answer: 目前有很多電商個資外洩的事件，審計部已經開始調查了，他們正在加強查核各機關資安相關事項的辦理情形，以確保個人資料

B.審計部將持續加強查核各機關的資安相關事項辦理情形，以確保個資安全與隱私保護。

C.根據所提供關於審計部的訊息，我無法直接回答有關審計部針對電商個資外洩事件的具體查核情況的問題。我的知識截至2021年9月，當時並沒有關於審計部具體處理此事件的細節。此外，審計部可能根據當時的政策和資源分配進行不同類型的稽核，但具體到每一起個別事件的查核結果我不能提供。

D.電商個資外洩事件頻傳，審計部已經展開查核。從108年起，審計部分別針對經濟部、衛福部等相關機關進行了解與檢討

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3。 3.中央政府各機關開發App成效與資訊安全情形為何？

A. This question is not yet answered, be the first to answer. Other questions in the same exam 1 依「國家發展委員會組織法草案」第2條規定，該會之任務何者不包括？(A)規劃及協調經濟、科技、社會、文化、教育等領域發展政策 (B)推動國家永續發展 (C)統籌跨部會

B. 基於所提供之審計作為及相關函文內容，審計部曾針對中央政府各機關開發之App進行審計並提出改善建議。在112年之後，數位發展部已要求民眾下載的App需通過TAF認可實驗室的資安檢測；此外，分年將逐步推動政府機關App納入無障礙檢測。然而，遊買集App的使用效能過低一事已在111年8月被監察院糾正並通知

C. 經函請相關權責機關研謀妥處。據復：數位發展部已於112年12月8日函請各機關，提供民眾下載之App應通過TAF認可

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

4。 4.科技發展日新月異，請問審計部有沒有針對新興科技進行查核？

A.是的，審計部會針對新興科技進行查核。例如，最近我們查核了區塊鏈的應用與管理情況，發現部分機關在管理上仍有改進空間。另外，我們也查核了人工智慧的相關應用，發現部分機關在資料蒐集、處理和使用方面沒有遵循適當的程序和準則，可能會有隱私權和個資保護等問題。因此，審計部會持續關注新興科技的發展趨勢

B.我們不掌握的信息。因為我們不知道審計部是否專門針對新興科技進行查核，且審計部的職能範圍和重點可能隨其政策和實際情況而變化。因此，我無法根據所提供的文本資訊來回答這個問題。請提供更多背景或相關資訊以便進一步協助。

C.審計部已經開始針對新興科技進行查核。

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5。 5.健保、華航、iRent等公私領域個資外洩後續處理情形及強化機制為何？成立個資專責機關! 監督機制？數位部角色？

A.個資外洩後的因應作為：(1)針對個資外洩的事件，由主管機關進行調查並提出相關建議。(2)對於企業個資維護作業缺失，提供相關建議以加強個資保護。(3)對行政機關監管企業個資維護作業缺失，研提建議意見以改善監管機制。

B.針對健保、華航、和雲(iRent)等公私領域的個資外洩事件，本部已進行相關專案調查並提出建議。在強化機制方面，他們研究了健保署健保醫療個資維護作業缺失、個資法修正草案對企業個資外洩通報規範不足、以及行政機關監管企業個資維護作業缺失，並通知相關機關進行檢討與改善。

C.由於本模型並非即時資訊來源，且未獲得最新資料，無法提供確切答案。然而，一般來說，政府部門在處理類似事件時可能會採取以下措施：釐清責任：調查洩密原因及相關單位的責任

D.健保、華航、iRent等公私領域個資外洩後續處理情形及強化機制為何？成立個資專責機關與監督機制？數位部角色？

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

6。 6.i郵購靠員工消費，會員資料資安堪虞，審計部查核情形？

A.執行概況：數位部訂定網際網路零售業及網際網路零售服務平台業個人資料檔案安全維護計畫及業務終止後個人資料處理作業辦法（下稱電商業者個資安維辦法），以強化業者個人資料保護措施。

B.由於題主所提的背景文本中並未提及與i郵購相關的具體事件或資料外洩事件，且未明確指出審計部的相關調查結果。因此，我無法根據所提供的資訊來回答此問題。若您能提供更多相關背景資訊和特定問題，我會盡力協助解答。

C.i郵購靠員工消費，會員資料資安堪虞，審計部查核情形？請詳細回答。i郵購靠員工消費，會員資料資安堪虞

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

7. 7.對於企業發生重大個資外洩案件，審計部有沒有瞭解機關辦理個資外洩的問題？

A.是的，審計部於105年及106年針對各主管機關辦理個資外洩案件之行政檢查結果，發現部分機關未落實對企業個資外洩案件之行政檢查。

B.是的，審計部在審核意見中指出了三個主要問題：(一)個資法對企業個資外洩通報規範不足；(二)部分機關對企業個資外洩案件未落實行政檢查；(三)數位部未能引導企業善用該部監管之台灣電腦網路危機處理暨協調中心資源，致情資無法即時傳遞與處理。

C.根據所提供的資訊，我無法確定審計部是否知道機關在處理企業個資外洩問題上是否有不足之處。然而，可以推測的是，審計部可能會對相關機關的行政檢查和資源利用情況進行監督和評估。此外，由於數位部的監管下設有台灣電腦網路危機處理暨協調中心，審計部可能也會關注到該中心的運作與情資傳遞及處理效率的關係。

D.根據審核報告，審計部並未深入了解機關辦理個人資料外洩的情況。

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8。 8.電商個資外洩事件頻傳，審計部查核情形？

A.「當廣續」是動詞，意思是繼續、持續。在句子中，它表示「要繼續」的意思。「當」是一個介詞，可以用來表示時間或地點。在這裡，「當」後面接的是一個名詞組，表示「各機關」。「廣續」是一個形容詞，可以用來描述事情或狀態的持續性。在句子中，它表示「要繼續」的意思。

B.根據所提供的文字，我無法直接得知審計部針對電商個資外洩事件的具體查核情形。然而，一般來說，審計部的職責之一是監督政府及所屬機關（構）單位在資訊安全管理和保護民眾個人資料方面的表現。雖然文中提到將廣續加強資安相關事項的查核，但具體到電商個資外洩事件的處理上，可能需參考其查核範圍與頻率等細節才能提供更準確。

C.審計部為了加強保護人民的個人資料，防止電子商務企業等機構外洩個資，近期對各機關的查核情形進行了深入瞭解。

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

9。 9.中央政府各機關開發App成效與資訊安全情形為何？

A.依據審計處的報告，各機關開發App的成效因機關性質、使用者需求等因素而異，但整體而言，以交通部、內政部及教育部的App使用人次較多。此外，部分App可能未符合民眾期待或功能不足，因此建議相關單位應定期檢討改進。在資訊安全方面，審計處指出，各機關開發App時，應加強資安檢測

B.中央政府各機關開發App的成效與資訊安全情形是：數位發展部已要求各機關提供民眾下載之App應通過TAF認可實驗室之資安檢測，並分年分階段推動政府機關App納入無障礙檢測。此外，遊買集App使用效能過低已報請監察院處理，經監察院於111年8月糾正交通部。

C.根據審計部針對中央政府各機關開發的App的審計作為，已要求相關權責機關研謀妥處並函復結果。數位發展部已於112年12月8日發函，要求所有提供給民眾下載的應用程式（App）都必須通過TAF認可實驗室的資安檢測。此外，數位發展部也分年將推動政府機關App納入無障礙檢測，以提升整體政府的數位服務品質。

D.經函請相關權責機關研謀妥處。據復：數位發展部已於112年12月8日函請各機關，提供民眾下載之App應通過TAF認可

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

10。 10.請各機關，提供民眾下載之App應通過TAF認可

A.科技發展日新月異，審計部每年都會針對新興科技進行查核，以確保其符合相關法規要求並避免違法情事。舉例來說，去年審計部針對「人工智慧」進行了查核，發現部分政府機關在使用人工智慧技術時，並未遵循相關的法律規範和倫理準則，導致一些不當或違法的行為發生。因此，審計部提出了相關建議，以確保政府機關在使用新興科技時能夠遵守法規要求並保障公民。

B.不確定

C.根據所提供的文字，我們無法確定審計部是否專門針對新興科技進行查核。審計部的職責是監督政府和公營企業的財務和績效，以及確保他們遵守法律和規定。雖然不排除審計部可能對新興科技的財務影響和相關風險進行評估，但從所提供文本中來看，我們無法得出明確的結論。建議向審計部查詢以獲得更準確的資訊。

D.我不知道。

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

11。 11.健保、華航、iRent等公私領域個資外洩後續處理情形及強化機制為何？成立個資專責機與監督機制？數位部角色？

A.本部因應作為：有關健保、華航、和雲(iRent)及格上公司個資外洩情形，本部已辦理相關專案調查，並就健保署健保醫療個資維護作業缺失、個資法修正草案對企業個資外洩通報規範不足、行政機關監管企業個資維護作業缺失，研提建議意見，通知相關機關檢討改善。

B.針對健保、華航、和雲(iRent)等公私領域的個資外洩事件，本部已進行相關專案調查並提出建議意見。在強化機制方面，本部將研究修訂「個資法」以加強企業個資外洩通報規範。此外，也要求行政機關檢討改善對企業個資維護作業缺失。至於成立專責機關與監督機制，目前尚未有明確的規劃。數位部則扮演著重要角色，負責督導和協調相關單位的工作，

C.本部已在調查相關專案，針對各單位於個資保護上所存在的問題提出建議意見，並要求相關機關檢討改善。對於公私領域個資外洩的後續處理及強化機制，將視情況研議成立個資專責機關與監督機制。數位部的角色在於協調、推動及監督相關部會落實個人資料保護工作，確保民眾的隱私安全。

D.本部因應作為：有關健保、華航、和雲(iRent)及格上公司個資外洩情形，本部已辦理相關專案調查。

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

12。 12. i郵購靠員工消費，會員資料資安堪虞，審計部查核情形？

A.i郵購靠員工消費，會員資料資安堪虞，審計部查核發現。

B.有關i郵購員工消費及會員資料安全問題，根據審計部於2021年所發布的調查報告指出，該公司在員工消費及會員資料管理上存在一些缺失：一、員工消費管理：未建立明確的員工消費政策，使員工對消費規範認知不足。未落實員工消費單據審核機制。

C.執行概況：數位部訂定網際網路零售業及網際網路零售服務平台業個人資料檔案安全維護計畫

每列請僅選取一個答案。

	1	2	3	4	5
A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Google 並未認可或建立這項內容。

Google 表單

附件二：政府審計運用生成式 AI 辦理審計業務注意事項

隨著生成式 AI 技術的快速發展，然而由於生成式 AI 具有創造性，若未謹慎使用，可能會引發準確性與數據隱私等問題。因此，政府審計部門在運用生成式 AI 辦理業務時，應考慮幾項注意事項：

- 一、 政府審計單位在使用外部生成式 AI 技術時，應確保僅使用公開或經授權的資料，避免涉及任何未經公開或應保密的資訊，以防止敏感資料洩露的風險。對於封閉式部署的 AI 系統，使用前也應徹底確認系統的安全性，並確保其符合相關安全規範，減少潛在的資訊安全風險。
- 二、 使用的數據需經過充分的匿名化處理，以防止任何可識別個人或機構的敏感資訊被洩露，並確保數據的存取權限受到嚴格控管，藉此降低敏感資料遭不當存取或濫用的可能性。
- 三、 生成式 AI 模型的準確性和可靠性取決於訓練數據的真實性和多樣性，故在模型訓練過程中應嚴格審查數據的來源和品質，以避免因數據偏差導致模型輸出結果的偏誤。同時，定期對模型進行性能測試與優化，確保模型保持高度的準確性和公平性，降低決策誤導的風險。
- 四、 生成式 AI 的使用應完全符合政府審計相關法規，並針對技術人員提供必要的 AI 風險管理和操作培訓，增強技術人員在突發情況下的應對能力，進一步確保審計業務的穩定運行。
- 五、 生成式 AI 應作為輔助工具，而非決策主體。審計人員應在人機協作中保留專業判斷力，不可依賴 AI 生成結果作為唯一決策依據，確保審計結論的合理性和準確性，以維護審計報告的專業性和可信度。
- 六、 審計人員在生成式 AI 的應用過程中需符合透明與公正原則，保持高道德標準，嚴禁利用 AI 結果進行不實或偏頗的報告，並應向利害關係人清晰說明 AI 的應用範疇和局限性，以防止生成式 AI 的誤用或濫用對公眾信任造成影響。

七、 政府審計部門應制定全面的風險管理與合規制度，以因應生成式 AI 應用中可能出現的各類風險，確保 AI 技術在審計業務中的應用符合公正、透明和責任原則。

在生成式 AI 的運用日益深入的背景下，政府審計部門應確立系統化的風險管理及合規策略，並不斷提升技術人員的專業素養，以確保生成式 AI 技術在輔助審計業務執行時，同時維持公正、透明及問責的原則。