生成式 AI 技術應用與倫理的挑戰-國際趨勢與臺灣策略

國家科學及技術委員會
自然科學及永續研究發展處
郭廷洋 助理研究員

目錄

關鍵字：生成式人工智慧(Generative Artificial Intelligence)，學術倫理(Academic Ethics)，政策與策略(Policy and Strategy)，跨界技術應用(Cross-disciplinary Technological Applications)

摘要

本研究報告為探討生成式人工智慧(Artificial Intelligence, AI)技術的發展與應用。報告回顧了國際上的相關研究情形，概述了國外學術界在此的研究進展和學術倫理的趨勢，為本報告提供了廣泛的背景和參考框架。接下來聚焦於國科會所推動的「可信任人工智慧對話引擎」(TAIDE)的建置。TAIDE 的核心特點是其「可信任」的性質，不僅展現出卓越的智慧能力，同時也強調使用者的隱私、資料安全和技術的透明度。

報告中也提出了幾項建議：鼓勵在公部門及機構使用 AI 時，推廣 TAIDE 以確保透明和可信賴度；考慮技術發展迅速，應定期檢視「使用生成式 AI 參考指引」並設立研究倫理指南；鼓勵跨學科研究合作，特別是在特定領域，如醫療、法律及教育方面等，以確保 AI 語言模型在專業和倫理標準內發揮最大效益；及推廣公眾對 AI 的基礎教育，培養未來技術人才並深化大眾對 AI 的了解。

封面插圖：生成式 AI 對未來將造成重大影響，應同時注重相關技術及倫理問題。

壹、 緒論

(一) 背景介紹

近年來，大型語言模型(LLM, Large Language Model)如 OpenAI 的 GPT(Generative Pre-trained Transformer), Microsoft 的 Bing 及 Google 的 Bard 等系列，已經引起了全球的關注及吸引極大量的討論。這些模型的發展起源於深度學習和神經網絡技術的進步，大大提高了模型的訓練效率和預測能力。

GPT-2 是第一個引起廣泛討論的大型語言模型，由於其強大的生成能力，一度被 OpenAI 認為可能被用於不良用途而選擇不完全公開。但隨著技術的演進和社會的接受度提高，已公開的 GPT-3(後演進為 GPT-3.5)和後續版本如 GPT-4，不僅模型規模擴大，生成的語言也更加流暢且有邏輯性。

這些模型在當前科技領域的重要性不容小覷。首先，它們在自然語言處理領域創造了許多新的可能性，如自動文句生成、文章摘要、情感分析等。此外，大型語言模型也促使其他領域更為進步，如遊戲、醫療、教育和客服等，實現自動化和自我優化其模型。最為重要的是，這模型提供了一個平台，使非此專業技術背景的人也能夠與 AI 互動和獲益。

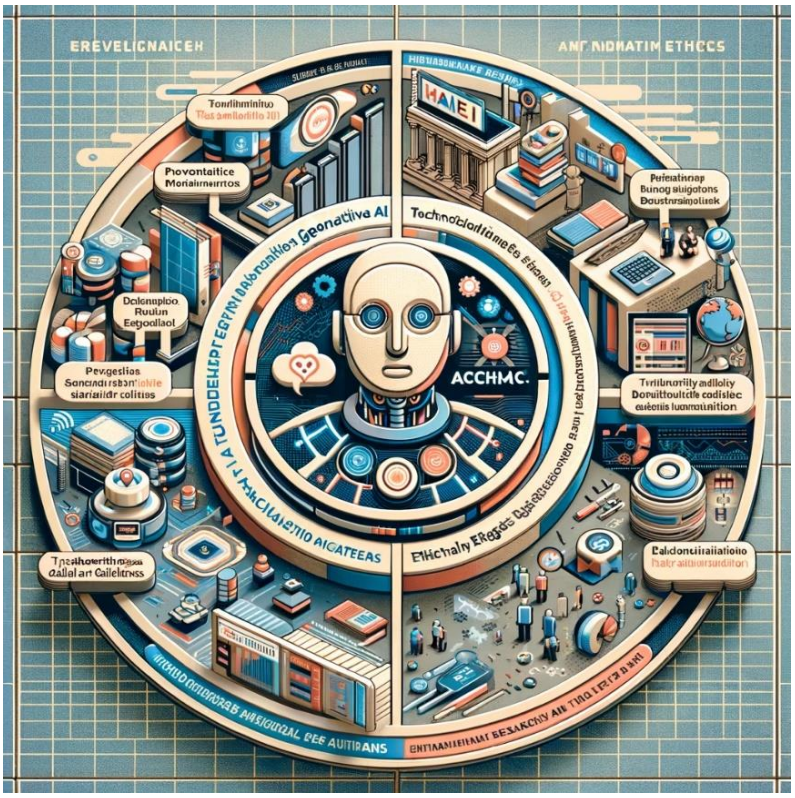

圖 1. 生成式 AI 在自然語言處理有專精技術，具有自動化、自我優化和易於接觸的特點，預期將在遊戲、醫療、教育和客戶服務領域等各社會層面帶來重大影響。

台灣在此方面也不落人後，也正如行政院政務委員兼國家科學及技術委員會主委吳政忠所指出，生成式 AI 已被證明在多個領域展現應用價值，因此對當地文化和語境的適應性成為了一個重要的議題。為此國家啟動了「可信任人工智慧對話引擎」(Trustworthy AI Dialogue Engine，簡稱 TAIDE)計畫[1]，著重於打造一款融入台灣文化、價值觀和風俗習慣的 AI 對話引擎，使其更加符合台灣用戶的需求。與此同時，透過國內教授團隊、國研院、國網中心、科政中心等多個機構部門的協作，確保模型的開發和訓練都基於豐富和高品質的資料，以及有力的計算能力支持。

圖 2. 吳主委參與 TAIDE 公部門應用工作坊之紀念合照，顯示 TAIDE 已接近實用階段。

然而，大型語言模型也帶來了相對應的挑戰。資訊真實性、潛在偏見和道德問題等成為了社會各界的焦點討論。

(二) 研究目的

隨著大型語言模型的崛起，其在學術研究的角色和影響日益凸顯。這些模型不僅具有撰寫論文摘要、文獻搜索、甚至初步研究草稿的能力，還改變了學者如何與資料互動和進行學術研究溝通的行為。評估此類模型在學術界的實際角色與潛在影響，有助於理解其如何塑型未來的研究範疇。

此主題之所以值得深入探討，是因為大型語言模型的普及化可能對學術誠信、研究品質和創新性產生深遠影響。對於像國科會的補助機構而言，了解這些影響有助於制定相關的補助策略，確保研究的品質和公正性。同時，對於審查端，能夠對這些模型生成的內容進行批判性評估，將確保學術出版的完整性和可靠性。期望透過本研究顯示大型語言模型在學術研究和出版領域中的真實價值和潛在挑戰。

(三) 研究方法與過程

本研究採取三大方法以深入了解與探討研究主題

(1) 資料蒐集：鑒於研究的主題資料相對新近，資料獲取來源主要集中於國際文獻系統及網路資料，進行廣泛的文獻蒐集與探究，確保所取得的資訊具有前瞻性和相關性。

(2) 專家交流：透過與領域內的專家學者直接交流，能夠了解其對於此大型語言模型的應用觀點，和模型可能造成的影響。

(3) 分析彙整：基於上述兩大方法所蒐集的資料進行整理和分析，最終將所得的結果統整於研究報告中。

貳、 大型語言模型的基本概念

(一) 技術背景：從人類反饋中的強化學習的大型語言模型

人工智慧的發展歷程中，特別是在大型語言模型的領域，「從人類反饋中的強化學習」被視為一種最前瞻的學習策略。這種策略不僅在傳統的機器學習模型中得到應用，更在大型語言模型中扮演著關鍵角色。當這些模型試圖生成語言或進行解釋時，人類的回饋成為調整其生成策略的重要依據。

大型語言模型如 ChatGPT、Bard、TAIDE 等依賴於龐大的數據集進行訓練，但單獨靠數據並不足夠，語言模型可透過從人類的具體反饋中學習，這些模型可以更準確地捕捉到語言的細微差異和文化背景，最終能夠結合人類的智慧，更好地適應並應對不同的語境和應用挑戰。

(二) 技術能力與限制

大型語言模型利用龐大的數據庫，通過深度學習技術，學習語言的結構和語境，從而能夠生成人類使用的自然文句。其核心優勢在於模擬人類語言的生成能力，並在各種應用場景，如聊天機器人、文章撰寫和語言翻譯等方面展現出強大的能力。

然而，它的強大能力也存在一些明確的限制：(1)它的知識基於其訓練數據，只能知道到訓練終止日期之前的資訊，若無外部輸入無法獲得最新的資訊內容。(2)模型可能無法區分事實和偽造的資訊，因為它只是模仿在訓練數據中所見的模式。(3)儘管模型能生成流暢的文句，但它們缺乏真正的理解和情感，可能會產生偏見或不恰當的內容。

參、 大型語言模型的在學術研究的應用及影響

當開始談論大型語言模型在學術研究中的應用及影響，至目前為止已可觀察到多種類型且多元化的應用。這些應用不僅協助學者進行更為深入的研究，還使得知識的傳播和整合更為迅速和精確。以下列出幾種目前已觀察到的應用例子和現象說明：

(一) 語言翻譯及撰寫能力

學術研究經常涉及多國語言的資料和文獻，面對多國語言的文獻資料，大型語言模型展現了其出色的即時翻譯能力，使研究者得以精確地理解非母語的研究論文與報告，從而廣闊其學術視野。一項於 2023 年在《Nature》雜誌中公布的問卷調查結果亦證實此趨勢[2](如下圖 1)，其中最熱烈的回饋是它將幫助那些英語不是母語的研究者，這群研究者明確表示，透過語言生成模型，他們在學術寫作和閱讀上的障礙得到了顯著的緩解。

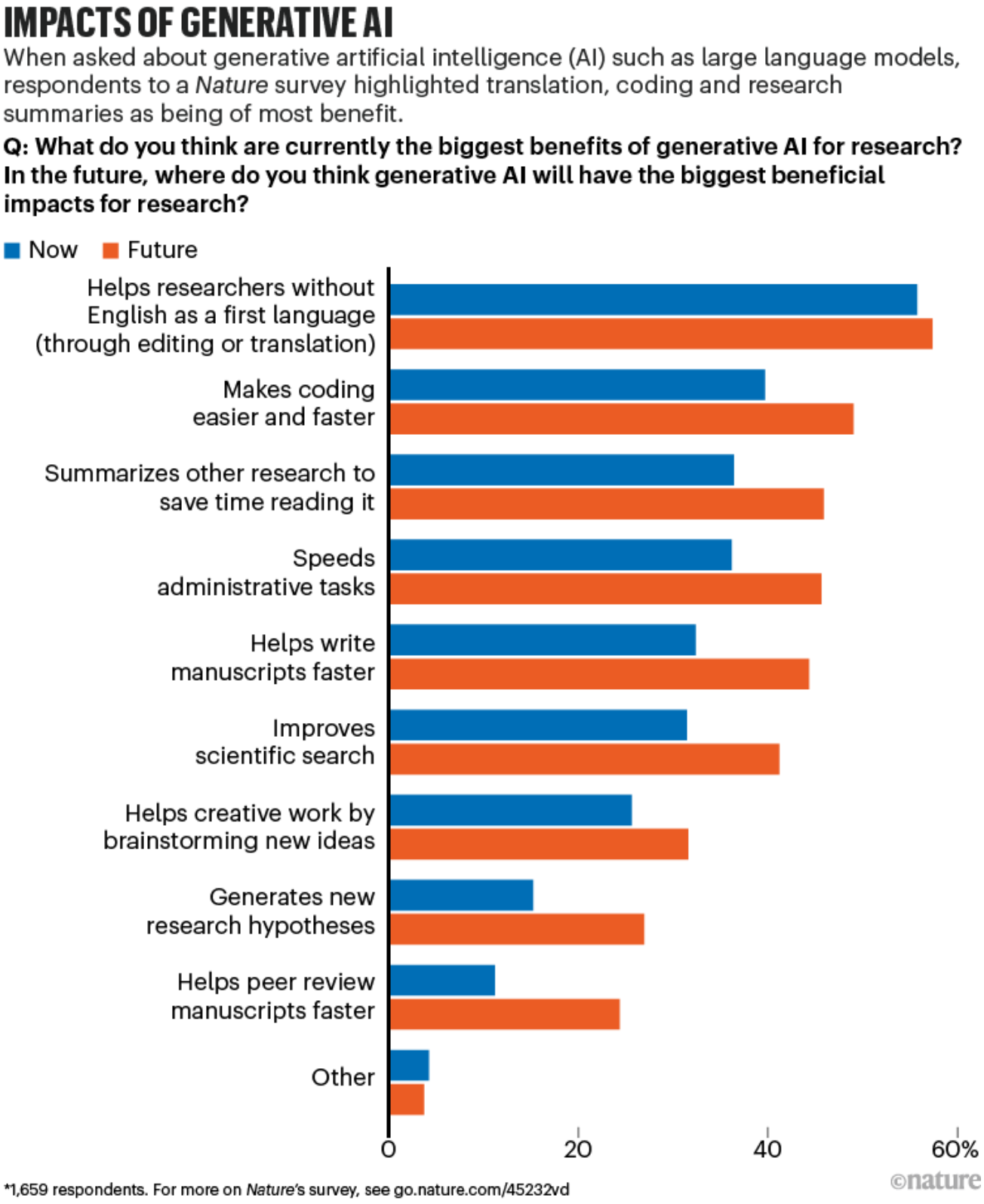

圖 3. 問卷調查顯示生成式 AI 最大的益處是幫助英語非母語者在翻譯嚴及編輯

(資料來源： *Nature* **621**, 672-675 (2023))

(二)　摘要生成能力

當學術研究者面臨海量的文獻資料時，語言模型的摘要生成技術成為了一項寶貴的工具，能夠迅速提供文章的核心思想，讓研究者無需逐篇深入閱讀。甚至許多先進的科學論文搜尋引擎，例如 Consensus[3](如下圖 4 說明)、Semantic Scholar[4]、Elicit[5]、scite[6](如下圖 5 說明)和 Iris 等等預先看到相關學術商機，已經融合此生成式 AI 技術，幫助用戶總結領域的研究成果、整理參考文獻、建議新的論文方向並生成新的研究摘要等多元功能。許多擁有大量科學摘要和參考文獻的學術出版機構與企業也正積極採納 AI 技術，將其資料庫綜合進 AI 驅動的系統中，希望在這波 AI 浪潮中不致於落後。

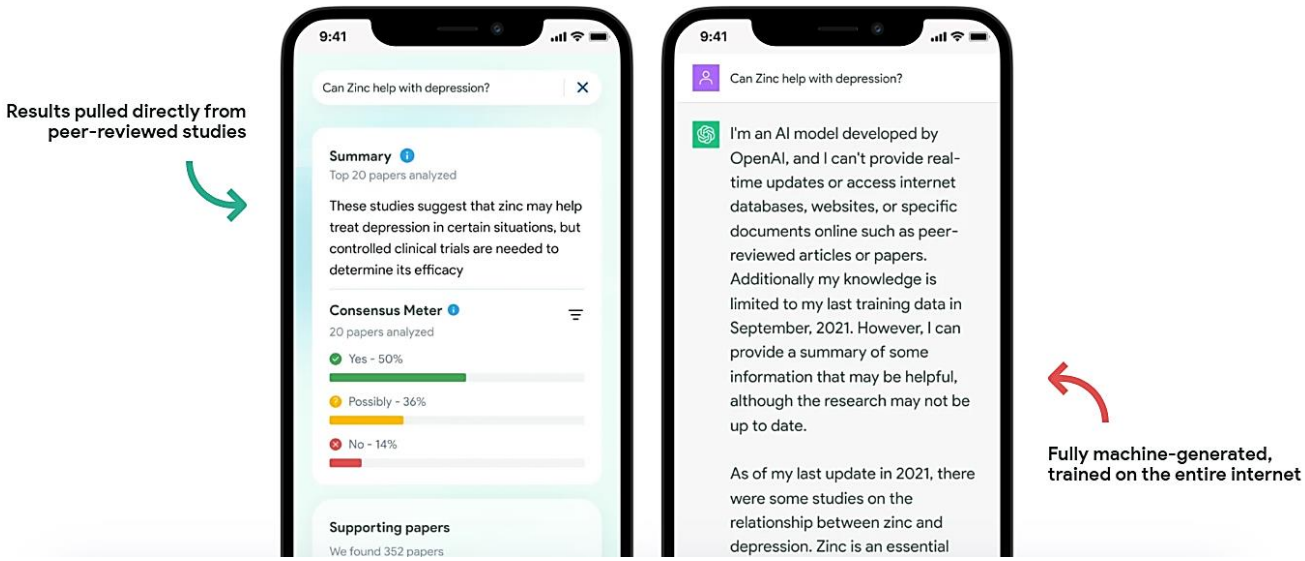

圖 4. 線上文獻摘要網站 Consensus 標榜其資料來源均是取自於經過同儕審查過的學術文獻，相比於全是由機器生成的 ChatGPT，相比之下其論文搜尋結果具有學術可信度。(圖片來源：Consensus 頁面內容)

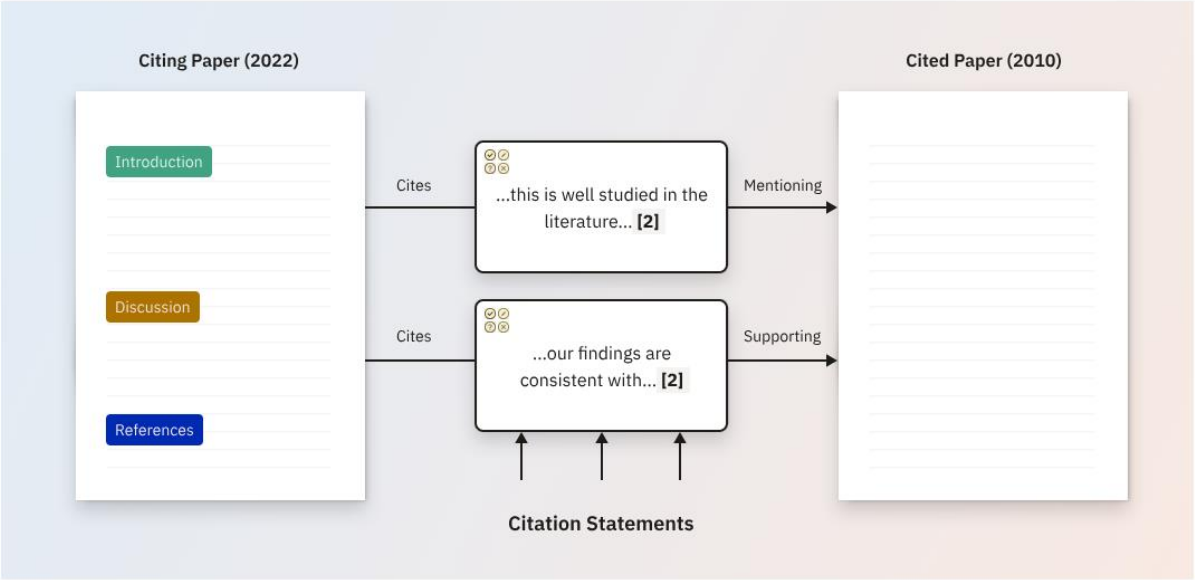

圖 5. Scite 透過 AI 的深層的語言模型的短能引用功能，協助使用者看到出版品的引用狀況，更透過閱讀文章之上下文內容，提供使用者引用的說明內容是支持正面還是反面的主張，進一步協助使用者理解論文內容。(圖片來源：Scite 頁面內容)

(三) 文獻搜尋及問答系統

透過問答系統，研究者可以向模型提問，快速地獲得具體的資訊或解答，無需浪費時間於繁瑣的搜尋和閱讀。與傳統的搜尋方式以關鍵字的搜索不同，這類型基於 AI 的系統主要採用向量比較技術；具體來說，論文文字內容會被轉化為一系列的數字向量，在「向量空間」中，向量的接近程度反映了論文之間的相似性。這種技術能揭露出隱藏或深層的文獻關聯，這些連結在傳統搜索中很可能被遺漏。

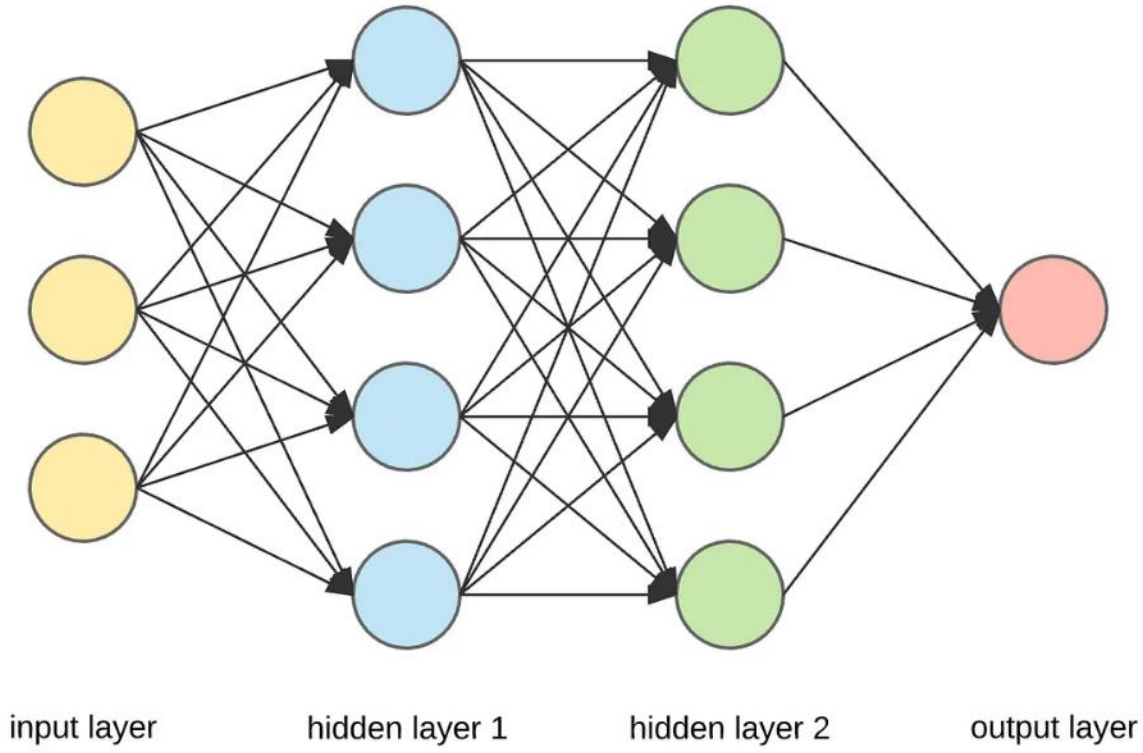

圖 6. 類神經網絡-深度學習模型為將輸入值轉化為一系列的數字向量代碼進行分析。(圖片來源：towardsdatascience.com)

具體的應用實例為 Elsevier 的聊天機器人名為 Scopus AI [7](2023 年 11 月進入 Beta 版測試)就運用了這一技術，幫助研究人員快速獲取他們不熟悉的研究主題的摘要，用戶提問問題而後系統回應，該聊天機器人以自然語言流暢地回饋關於所問研究主題的摘要段落，並附帶引用的參考文獻和進一步的建議的後續研究方向；美國 Clarivate 公司也表示，正在努力將大型語言模型的功能帶到其 Web of Science 數據庫中，但截至現今為止尚未公布上線時間表，但已開始將部分的內容如知識產權(Intellectual Property)導入 AI 分析功能[8]。

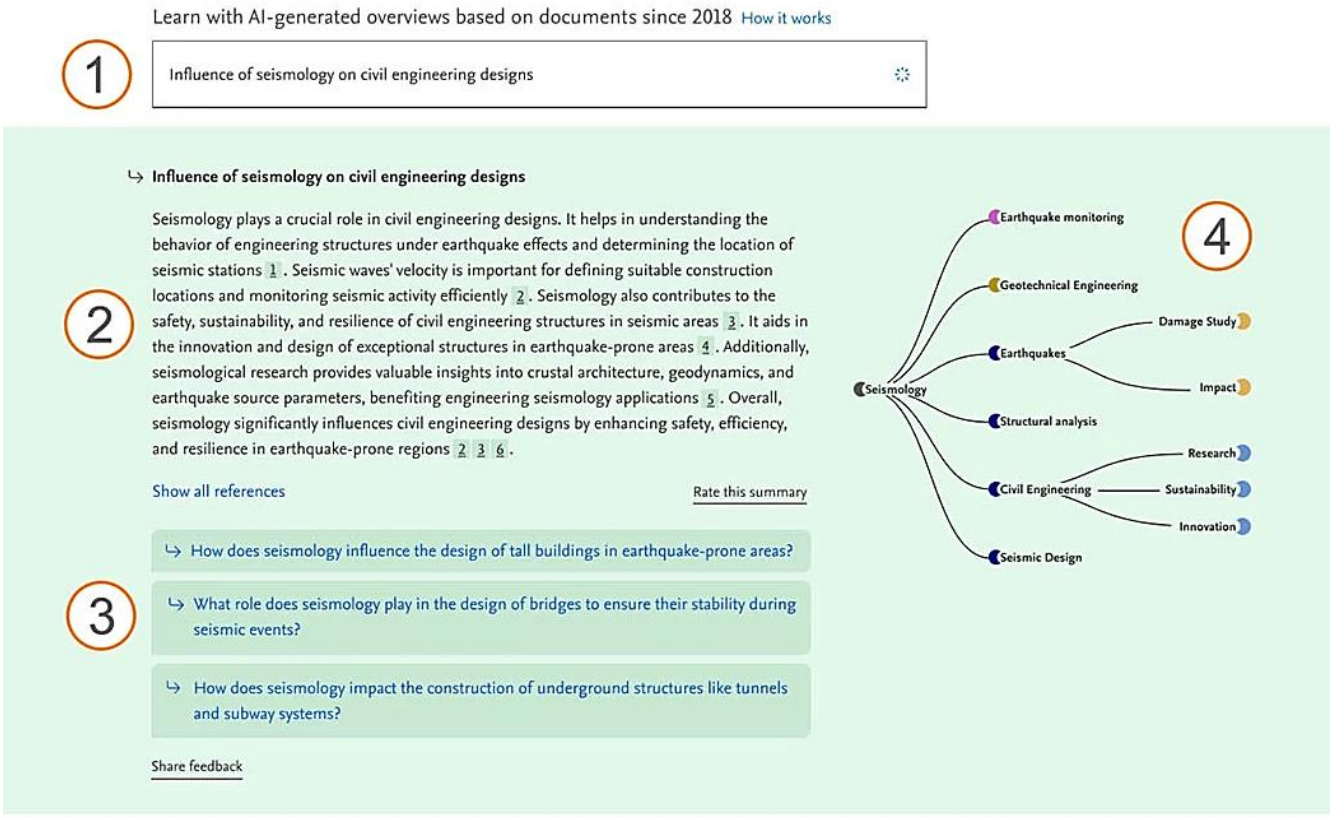

圖 7. Scopus 以生成式 AI 的敘述方式概括使用者的研究主題和相關參考文獻，進而回答使用者問題。(圖片來源：Scopus AI 頁面內容)

(四) 協助撰寫程式碼

大型語言模型已進化到可以協助程序設計師撰寫程式碼的程度，而且不僅只做基本程式語言代碼生成，而是能夠理解複雜的程式邏輯並提供合適的代碼片段或建議。例如 OpenAI 訓練了一個名為 Codex 的模型，專為此目的而生，因有這個功能，GitHub 看到了其潛力並將其整合為 Copilot 服務[9]，讓全球的程式設計師能更順暢地編寫和優化程式代碼。這不僅提升了程式開發的效率，也讓初學者更容易進入程式設計的領域，這項服務在 Gitub 的使用者回饋中，高達 88%的使用者回覆使其工作更有效率，96%的使用者認為減少了重覆性的工作，並能夠撰寫出品質更高的程式碼，更能專注在程式應用開發上[10]。

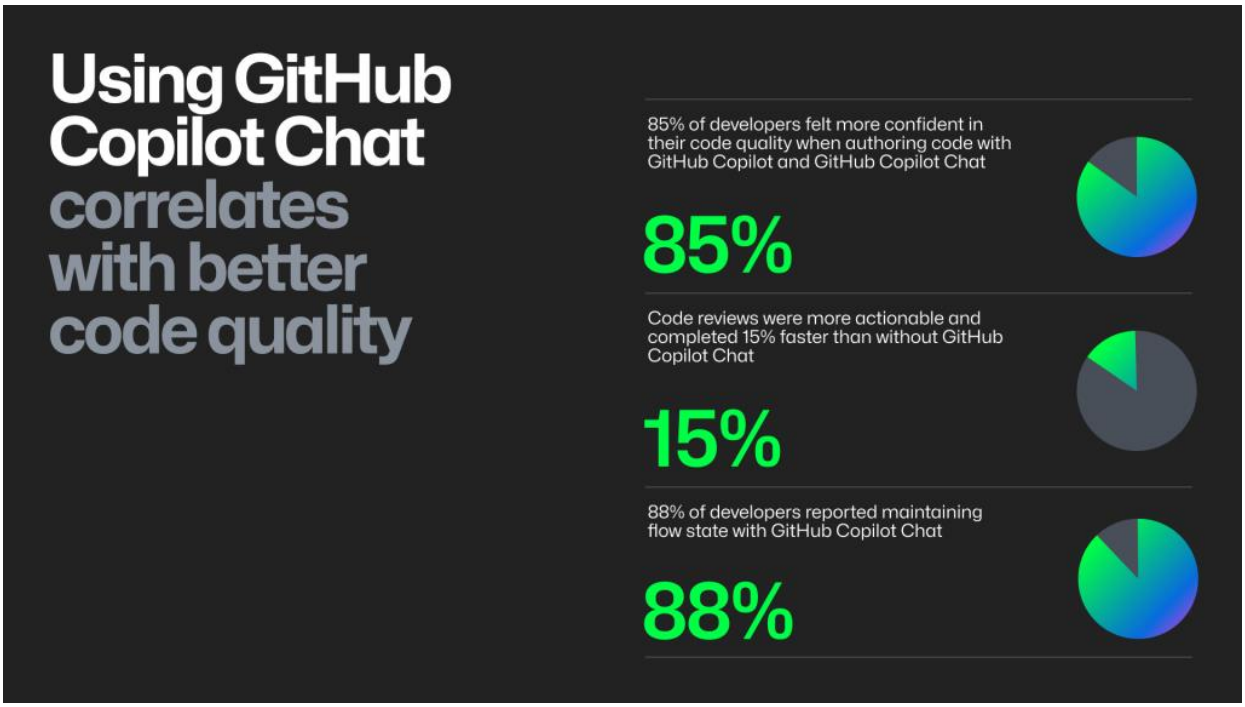

圖 8. GitHub 的研究顯示使用 Copilot 編寫程式碼時，85%的程式開發人員認為其程式編寫更有品質，速度也快了 15%，88%的開發人員認為能讓他們的編程工作更為流暢有效率。(圖片來源：GitHub blog)

(五)　特殊專業領域的專精研究

科學技術迅速發展的時代，大型語言模型不只是一般研究的簡單工具，它也開啟了對特殊學科專精領域更深入研究的大門，這些高度專門化的領域，從醫學、生物技術到基礎但專精的物理及化學領域等，都有著極大的知識範疇和繁複的細節，有了大型語言模型的幫助，將有機會更有效、更精確地探討這些領域的細節和內涵。舉澳大利亞團隊現所發展的 DARWIN 達爾文系統為例[11,12]，是一個專為物理、化學和材料科學設計的大語言模型，並基於先進的技術和大量的科學資料進行精細調校，可能為未來的科學發現帶來重大的影響。

除了生成式 AI 外，各種機器學習 AI 已經在多個科學領域扎根並取得顯著進展，以 Google DeepMind 為代表的演算法，在圍棋界的驚人表現之後，進一步在各科學領域展現其影響力。其中 AlphaFold 作為基於人工智慧的蛋白質結構預測方法，在結構生物學中有許多重大進展。它在解析大分子的三維結構、推動藥物發現、蛋白質工程以及深化生物學理解方面發揮了關鍵作用[13]，台灣也有相關運用 AlphaFold 為工具之論文發表[14]。近期，DeepMind 又將其技術擴展到材料科學領域，開發了 GNoME（Graph Networks for Materials Exploration）[15]。這一基於圖神經網絡的模型，專注於探索和預測穩定晶體結構，預計將在材料科學和物理化學領域帶來全新的研究和探索途徑，從而使得該領域的發現和創新更加高效和精準[16]。

即便如此，實驗科學仍有其價值，近期的研究報導[17]指出儘管 AlphaFold 的預測模型在多數情況下展現出驚人的準確性，但仍存在其侷限性，尤其是在與環境因素(溶劑，pH 值等)、配體和離子相關的細節上較為貧弱，更需要實驗的驗證。因此，AI 理論模型的預測更像是有價值的假設，而非蛋白質結構的決定性特質。這突顯了在結構生物學中持續進行實驗驗證的必要性。實驗方法不僅可以確認由 AI 模型提供的預測，還能解釋那些在預測模型中可能遺漏或不足的關鍵細節，進而完整地理解蛋白質的結構和功能。

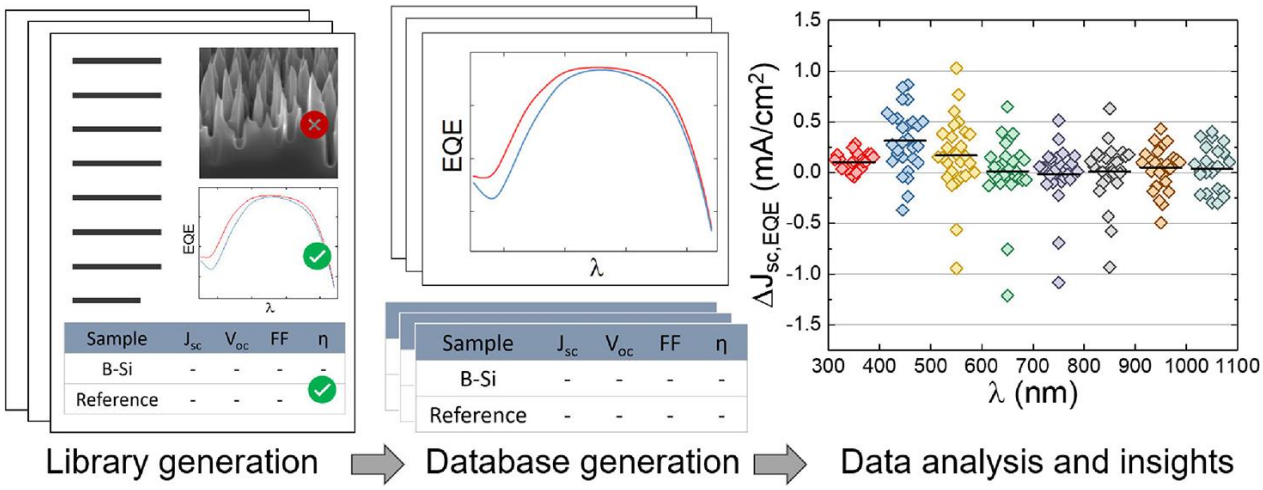

圖 9. 使用 AI 輔助研究奈米混合物運用於矽太陽以電池的應用潛力，已獲刊登於國際學術期刊 *ACS Appl. Nano Mater*. 5, 8, 11636–11647 (2022) .

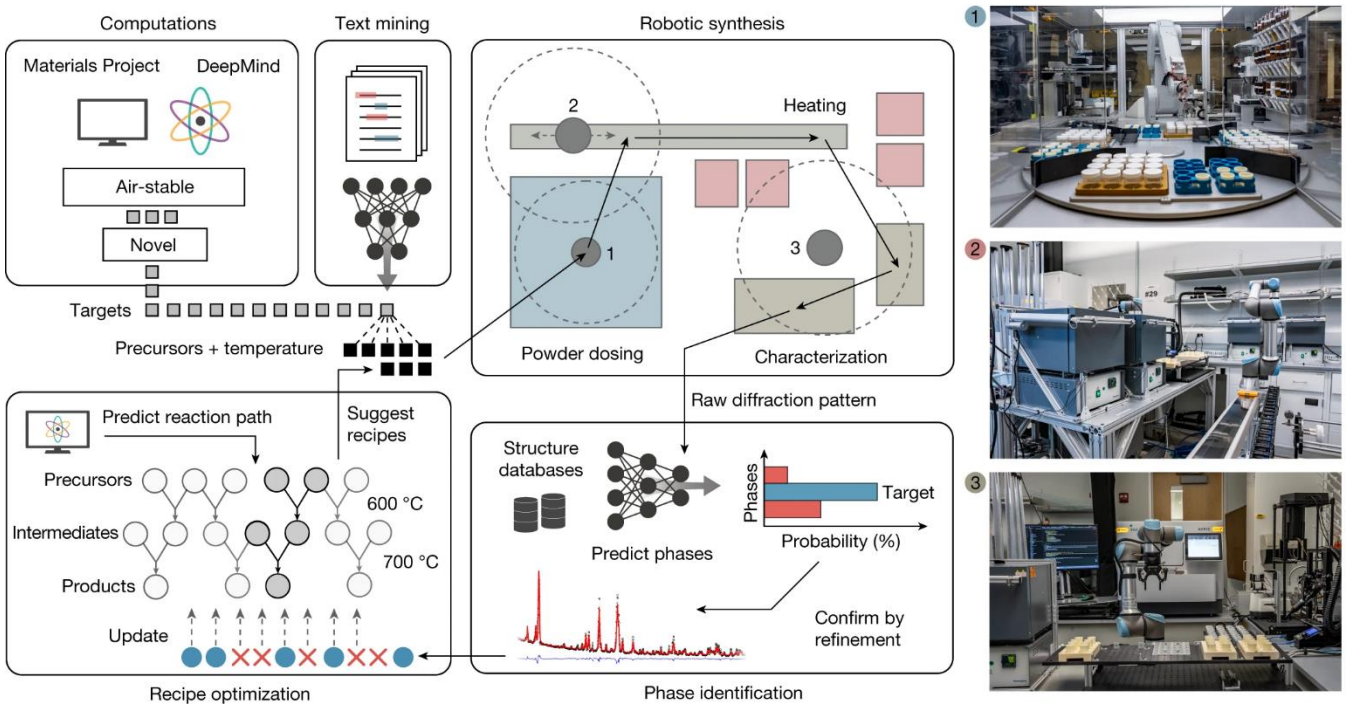

圖 10. 透過機器學習及自動化機器手臂的材料合成實驗室，從化學成份合成配方調製、樣品加熱、特性分析、XRD 評估純度、識別目標物質的反應途徑等完全自動化流程。(圖片來源：Nature (2023)[16])

(六) 未來可能發展的應用展望

生成式 AI 具有進一步改變科學研究的巨大潛力，預計將深刻影響多個領域，包括但不限於以下可能的發展：

(1)個性差異化：目前語言生成模型已經能夠根據使用者的數據生成個性化回應，未來這種技術將進一步發展，可以設定有更加量身訂做和客製化的用戶體驗和語言理解，例如 ChatGPT 4 發布的 Beta 版引入了針對特定專案需求的語言生成選項及生成平台。

(2)處理情感的能力：預期生成式 AI 可能發展出識別和回應情感的能力，從而創造出更加人性化和具有同理心的互動體驗。

(3)與現有其他技術整合：生成式 AI 可以與其他技術整合，如聊天機器人和虛擬實境等，將提供更加無縫和整合的體驗，跨越不同多個平台和設備。此一趨勢也已被國科會工程處之相關學門認知，已著手規劃徵求 113 年之學門主題式計畫「生成式人工智慧機器人控制之整合與應用」，聚焦於應用情境及技術實現及落地的可能性。

(4)面向特定行業領域的模型：隨著在各個行業中對專業知識和專長的需求增長，針對特定領域的 AI 語言模型已陸續有多項應用已經產生，如醫療保健、金融、法律和科學等。這些專門的模型更加貼近行業實際需求，可以為領域內的用戶提供更準確、更相關和更深入的訊息。

這些應用展現了大型語言模型在學術研究中的多面性和實用性，而隨著技術的進步，我們可以預期未來還會有更多創新的應用方式出現。然而，技術的影響是雙面的，它也帶來新的挑戰：資料的隱私問題，過度倚賴生成資料而未經查證，著作者權的問題，這些新的倫理和實務問題，需要所有人共同面對和思考解決對策。

肆、 大型語言模型所帶來的挑戰及問題

隨著大型語言模型的興起和廣泛應用，它們不僅帶來了巨大的價值，同時也伴隨著一些重要的挑戰和問題。這些問題涉及道德、偽造、以及社會層面，需要各方面的關注和應對。

(一) 研究倫理問題

大型語言模型在學術研究領域中的應用已經引起了深入且廣泛的討論。這種模型具有高效撰寫學術論文、整理研究報告的能力，其產出的研究摘要品質之出色，常令人難以分辨其是否由非人類 AI 所生成。但這同時也帶來了一系列的倫理挑戰。例如，這些模型有時可能會參考那些表述流暢但數據不精確的低品質研究，導致讀者誤解。令人更加擔憂的是，當前的學術審核系統，無論是期刊出版商、同行評審還是普通讀者，均缺乏有效手段來鑑別這些 AI 生成的資訊。

有鑑於此，多家知名的學術期刊已對 AI 在研究中的角色提出清楚指引。《科學 Science》期刊堅決認為，利用 AI 工具產生的稿件應被視為學術不誠信的行為。他們明確指出，任何由 AI 所生成的內容都將被視為剽竊，除非經過特許。而《自然 Nature》期刊雖然不接受將 ChatGPT 或其他 AI 工具列為論文作者，但它並未禁止這些工具的使用，只是要求在特定部分詳細說明其應用。Elsevier 出版集團也已明確其對 AI 的政策，要求作者必須透明地聲明其使用情況。Wiley 出版集團的立場亦強調，AI 工具本身無法擔任原創研究的發起者，且不能為論文或研究設計承擔責任。《ACS Publications》則更進一步指出 AI 如何可能影響學術寫作的真實性和品質，呼籲制定更加嚴格的行為守則。而《Physical Review Journals》的策略是允許 AI 工具用於輕度編輯，但明確指出責任仍由人類作者和評審承擔。學術界對於 AI 在研究寫作中的角色持褒貶不一的態度，但都確實強調了作者身份、著作權和學術真實性的重要性。這些指引的詳情整理於下表 1。

表 1、主要科學期刊出版商使用人工智慧工具的政策

期刊出版方	*政策聲明*	*出處*
AAAS (American Association for the Advancement of Science; publisher of Science)	We would not allow AI to be listed as an author on a paper we published,and use of AI-generated text without proper citation could be considered plagiarism.不允許 AI 在發表的論文上被列為作者，且未正確引用 AI 生成的文本可能被視為剽竊。	Science 2023,379, 313
Springer Nature (publisher of Nature)	ChatGPT doesn't meet the standard for authorship. Authors using LLMs (large language models) in any way while developinga paper should document their use in the methods or acknowledgements sections. ChatGPT 並不符合作者資格的標準。使用 LLMs（大型語言模型）進行論文開發的作者，應在方法或鳴謝部分記錄其使用情況。	Nature 2023,613, 612
Elsevier	The use of AI tools can improve the readability and language of the research article but cannot replace key tasks that should be done by the authors, such as interpreting data or drawing scientific conclusions. AI and AI-assisted tools cannot be credited as an author on published work. AI 工具可增進文章可讀性，但不能替代作者解釋數據或結論，且 AI 不能被認為是文章作品作者。	*1
Wiley	Artificial Intelligence Generated Content (AIGC) tools—such as ChatGPT and others based on large language models (LLMs)—cannot be considered capable of initiating an original piece of research without direction by human authors. They also cannot be accountable for a published work or for research design, which is a generally held requirement of authorship, nor do they have legal standing or the ability to hold or assign copyright. AI 生成工具如 ChatGPT 不能獨立進行原創研究，且無法對作品負責或擁有版權。	*2
ACS (American Chemical Society) Publications	AI tools do not qualify for authorship and that any such tools used to produce text or images should be disclosed within the manuscript. AI 工具不符合作者身份的資格，使用這些工具生成的文章或圖像應在初稿中進行披露。	*3
Physical Review Journals (APS, American Physical Society)	●Authors and Referees may use ChatGPT and similar AI-based writing tools exclusively to polish, condense, or otherwise lightly edit their writing. As always, authors must take full responsibility for the contents of their manuscripts; similarly, referees must take full responsibility for the contents of their reports. 作者和評審可用 AI 工具修飾文章，但仍須對內容承擔全責。 ●An AI-based writing tool does not meet the criteria for authorship because it is neither accountable nor can it take responsibility for a research paper's contents. A writing tool should, therefore, not be listed as an author but could be listed in the Acknowledgments. Authors should disclose the use of AI tools to editors in their Cover Letter and (if desired) within the paper itself. Referees should disclose the use of AI tools to editors when submitting a report. These disclosures will help editors understand how researchers use the tools in preparing manuscripts or other aspects of the peer review process. AI 寫作工具因不能負責於研究內容而不應列為作者，可列於致謝。	*4

	作者與審稿者應告知編輯使用 AI 工具的情況，幫助其理解工具在稿件準備或評審中的用途。 ●To protect the confidentiality of peer-reviewed materials, referees should not upload the contents of submitted manuscripts into external AI-assistance tools. 為確保評審機密性，審稿者不應上傳手稿至外部 AI 工具。

*1：https://www.elsevier.com/about/policies/publishingethics
*2：https://authorservices.wiley.com/ethics-guidelines/index.html
*3：https://axial.acs.org/publishing/ai-in-publishing-the-ghost-writer-in-the-machine
*4：https://journals.aps.org/authors/ai-based-writing-tools

(二) 信賴與品質議題

在大型語言模型的背景下，信賴與品質的議題顯得十分重要。語言模型往往是基於大量的資料進行訓練，該資料涵蓋了各種文句和知識來源。然而資料的品質和來源多樣性對模型輸出的可靠性和正確性有直接的影響，例如前日產生爭議的新聞報導中的個別研究人員實驗使用的測試模型即是錯置使用非台灣習慣之中文語料庫。語言模型會吸收存在於其訓練資料中的偏見和不正確的信息。如果訓練數據中有誤導性或是具有偏見的資訊，模型在回應相關問題時可能也會反映這些誤導或偏見，此外因為數據資料庫的時間限制，語言模型所提供的內容不一定是最新的確切資訊，因此使用者必須對模型的回應持批判性思考，並在必要時進行額外的核實。

使用大型語言模型進行科學文獻搜尋時，存在可靠性和準確性的問題，因為不是真正理解其輸出的內容，而可能輸出含有事實錯誤和偏見的內容，甚至可能虛構不存在的參考文獻，進而對研究造成嚴重的誤導。例如 Elsevier 所發展的 Scopus AI 便採取了較保守的策略，採取了多種措施來限制和指導其搜索和回答的方式，從而提高其回答的可靠性，它基於五或十個真實存在的研究摘要來產生答案，且只搜索自 2018 年以來的論文以確保資訊的新鮮度，另外發展的研究人員還設定了產生文句的低閾值，減少其選擇偏離標準答案的機率[18]。

國科會於今年所發展的「可信任人工智慧對話引擎」(TAIDE) 模型代表了對人工智慧信賴度的進一步追求。一般的大型語言模型在回應查詢時，可能會受到其訓練資料的影響，而反映出某些偏見或誤導性資訊。TAIDE 的設計和目標是為了提高對話模型的可靠性和透明性，其訓練資料語庫以臺灣文化為基底，融入在地特有的語言、價值觀、風俗習慣等元素，使其在回答問題時能

更為精確且少有偏差，期望能夠達到更高的答覆品質及穩定準確的回應。

然而技術層面要未到十分完美無瑕是非常困難的，因此即使是如 TAIDE 這樣的先進模型，使用者在互動時仍需要保持警覺，並對任何機器生成的回答持以批判性思考，並在必要時進行額外的查證。此 TAIDE 模型的出現，不僅顯示了 AI 技術的發展方向，也強調了為了增加公眾的信任，模型的透明度和可靠性是不可或缺的關鍵因素。

(三) 少數族群之偏見議題

大型語言模型通常針對廣泛的問題提供答案，但在面對少數族群、特定文化或背景的特殊問題時，其答案可能變得不夠明確或失去精準性。當訓練資料缺乏對某一少數族群的充足資訊，模型對該族群的文化、價值觀和需求的理解可能會變得表面化或不足。

這種現象可能使少數族群在利用這些 AI 工具時，感受到被邊緣化或誤解。當使用者不了解模型的運作機制，他們可能會誤認為模型提供的答案是絕對中立和客觀的，而未察覺到其潛在偏見，這樣的情況有可能無意中助長了社會中已存在的刻板印象和偏見。

舉例來說，2023 年 11 月 Elon Musk 領導的 X 公司(原 Twitter)規劃推出了聊天 AI 稱為 Grok[19]。這款語言生成模型以其獨特性格聞名，包括「叛逆傾向」、「黑色幽默」、「避免過度政治正確」和「喜歡諷刺」等特點。此前，Musk 於同年 4 月在社群中發布了命名為 TruthGPT 的計畫，旨在探索宇宙真理和社群真實。雖然相關新聞內容說明至今進度只限於內部封閉測試而尚未公開，若是其為真，如果用戶未能充分理解這些聊天機器人的本質而去濫用，可能會產生誤解，進而引發更大規模的偏見。

(四) 保密議題：

大型語言模型所涉及的隱私和保密性議題也引發了公眾的廣泛擔憂，在學術界及業界均有不同的考量環節，但確保語言模型的精準和安全使用至關重要，所涉及方都必須對其如何儲存、保護和數據使用有徹底的了解。

在學術界，同行評審是確保研究品質的重要環節。大型語言模型有助於提供審查人的評審效率，但因上傳數據至雲端資料庫可能帶來潛在保密風險。由

於保密性的擔憂，幾個重要的研究或補助機構，例如美國國家衛生研究院(NIH)[20]及澳大利亞研究委員會(Australian Research Council，ARC)[21]等已禁止使用 ChatGPT 和其他生成式 AI 工具進行研究補助經費的同行評審，可以想像其他國家的經費補助機構也會開始制定相關的措施。重要的出版社如 Elsevier[22]、Taylor & Francis、Physical Review Journals 和 IOP Publishing 等[23]，也禁止審稿人上傳手稿到生成型 AI 平台以產生審稿報告，因為這些論文可能會被回饋到訓練數據庫中，這將違反保密條款。部分出版商和機構(例如 Wiley)已開始考慮使用私有主機的大型語言模型協助篩選稿件，以確保數據或需保密的訊息不會被外洩。

在商業領域中，業界公司往往擁有大量敏感商業資訊，從業務策略到客戶資料等。如果使用語言模型，必須確保這些數據不會不小心被迴饋至資料庫或洩露。特別是在高度競爭的行業中，任何形式的資料外洩都可能對公司的競爭地位造成嚴重影響。

隨著大型語言模型的廣泛應用已為現代人類帶來了不少便利性，但也引發了多方面的問題和挑戰。學術界面臨如何定義 AI 在研究論文中的角色的倫理問題，以及如何識別 AI 生成的內容。模型的資料來源和訓練方法使其回應可能存在偏見或不正確的資訊，尤其針對少數族群，可能出現答案不夠精確或偏頗的情況。此外模型涉及的隱私和保密性問題也被廣泛討論，特別是在學術研究和審核過程中需要特別謹慎。雖然大型語言模型帶來了眾多優勢，但同時也提醒我們在使用過程中需持批判性思考，並關注其可能帶來的社會、道德和技術挑戰。

伍、 現階段國家的政策

隨著人工智慧技術的快速進展，其研究手段、應用策略與場景也持續轉變。國科會身為國家科研的領航者和主要補助機關，深刻洞悉大型語言模型發展所帶來的可能發展與伴隨的挑戰，並擔綱領導之責。於 112 年 10 月 3 日，國科會頒布了「行政院及所屬機關（構）使用生成式 AI 參考指引」。該指引明確訂出了策略大綱，著重於生成式 AI 的雙面性：其所帶來的巨大機會與風險。雖然這類語言模型具有多重功能，我們更應留意其潛在帶來的問題，如智慧財產、人權及業務機密的潛在風險。

此參考指引之內容全文引用如下：

一、為使行政院及所屬機關（構）（以下簡稱各機關）使用生成式 AI 提升行政效率，並避免其可能帶來之國家安全、資訊安全、人權、隱私、倫理及法律等風險，特就各機關使用生成式 AI 應注意之事項，訂定本參考指引。

二、生成式 AI 產出之資訊，須由業務承辦人就其風險進行客觀且專業之最終判斷，不得取代業務承辦人之自主思維、創造力及人際互動。

三、製作機密文書應由業務承辦人親自撰寫，禁止使用生成式 AI。前項所稱機密文書，指行政院「文書處理手冊」所定之國家機密文書及一般公務機密文書。

四、業務承辦人不得向生成式 AI 提供涉及公務應保密、個人及未經機關（構）同意公開之資訊，亦不得向生成式 AI 詢問可能涉及機密業務或個人資料之問題。但封閉式地端部署之生成式 AI 模型， 於確認系統環境安全性後，得依文書或資訊機密等級分級使用。

五、各機關不可完全信任生成式 AI 產出之資訊，亦不得以未經確認之產出內容直接作成行政行為或作為公務決策之唯一依據。

六、各機關使用生成式 AI 作為執行業務或提供服務輔助工具時，應適當揭露。

七、使用生成式 AI 應遵守資通安全、個人資料保護、著作權 及 相關資訊使用規定，並注意其侵害智慧財產權與人格權之可能性。各機關得依使用生成式 AI 之設備及業務性質，訂定使用生成式 AI 之規範或內控管理措施。

八、各機關應就所辦採購事項，要求得標之法人、團體或個人注意本參考指引，並遵守各機關依前點所訂定之規範或內控管理措施。

九、公營事業機構、公立學校、行政法人及政府捐助之財團法人使用生成式 AI，

得準用本參考指引。

十、行政院及所屬機關（構）以外之機關得參照本參考指引，訂定使用生成式 AI 之規範。

綜合上指引內容，已具體指出不應使用生成式 AI 製作機密文件，並禁止向其提供需保密的公務或個資，此舉有助於維護國家與公眾的敏感資料安全。指引也強調生成式 AI 的輸出需經由業務承辦人進行專業且客觀的最終判斷，不僅確保 AI 不替代人的專業與創意，還明確表示生成內容不應成為行政或公務決策的唯一依據。若以生成式 AI 提供服務或執行業務，應適時進行透明的揭露，確保公眾的信賴與理解。最後，該指引明確強調了在使用過程中需嚴格遵循的資通安全、個資保護、著作權等法律規範。

國科會在人工智慧領域持續投入，特別對於大型語言模型的生成式 AI 技術展現高度關注和支持。為此國科會特別推動了「TAIDE」—「可信任人工智慧對話引擎」的建置。TAIDE 的獨特之處在於它的「可信任」特性，意味著其不僅具備高度的智慧能力，還重視使用者的隱私、資料安全和技術的透明度。這樣的設計旨在保障為台灣設計的用戶利益，避免不必要的風險，並建立公眾對 AI 技術的信心。同時也透過 TAIDE，期望能夠為國內的研究者、開發者和相關業者提供一個集中、專業且高效的資源和工具，並促使生成式 AI 在各領域的應用更具深度和廣度，將技術帶來的好處更安全地推廣到各個層面。

陸、 個人研究建議

綜合上述論述，對於現今發展仍有可推展的空間，個人有幾個建議供參考：

(一)**TAIDE 的應用與推廣**：有鑑於「可信任人工智慧對話引擎」(TAIDE) 具有確保生成式 AI 可信賴性的特點，建議架設完成後應積極支持並推廣其應用，確保公部門及機構等使用 AI 時能達到透明與可信賴的效果。

(二)**TAIDE 之技術面向應隨時注重其建置目的及特質**：(1)**透明度與可解釋性**，將有助於使用者理解其生成內涵，進一步提高對 AI 系統的信任和接受度。(2)**隱私和數據保護**，系統應就傳輸數據加密、訪問控制、後台資料儲存使用等面向進行資安防護措施，以免相關數據被濫用。(3)**公平性與無偏見**，這可能是最困難的部分，需要在設計和訓練素材中持續監控並且使用多樣化的數據內容，避免演算法或是訓練素材的偏差，而造成對特定群體產生不公平的系統。

(三)**持續更新與完善指引**：考量到技術發展迅速，建議定期檢視與更新「使用生成式 AI 參考指引」，以確保其內容與時俱進。另外對於學術研究發展，也建議設立明確的研究倫理指南，將研究和使用 AI 的倫理和規範制定出來，這樣可確保推動 AI 技術發展的同時，也能夠兼顧社會價值和人權。

(四)**國科會相關計畫的倫理、審查議題調適**：考慮到生成式 AI 是一種新興技術，國科會的相關倫理和審查標準也需與時俱進，即使有生成式 AI 的參考指引，學術科學的倫理和審查程序同樣需要更新，以避免在新技術下產生模糊的灰色地帶。例如前述美國國家衛生研究院和澳大利亞研究委員會等已經規定審查人員不得將資料上傳至生成式 AI，我國「國家科學及技術委員會審查獎勵及補助案件迴避及保密作業要點」中的保密原則指出，未經授權的相關人員不得洩露審查資料予他人。但在生成式 AI 的部分黑盒子尚未透明化的情況下，若是審查人員將計畫書機密資料上傳是否會進入機器學習之迴圈，成為值得進一步討論的議題，另在相關學術倫理的規則訂定也需要一併考慮討論。

(五)**推動鼓勵、培育跨領域研究合作**：鼓勵計算機科學家、社會學家、心理學家等不同領域背景的專家學者合作，以跨學科的方式共同研究 AI 技術在多元領域中的應用和影響。例如在醫療、法律或教育等特定領域下的應用挑戰，AI 語言模型如何在符合專業標準和倫理規範的前提下發揮最大效益？特別是語言模型建議或是生成的內容能和現有的專業智識、技能、和判斷能力相互協助而融合，而不是單純取而代之。

(六)**推廣公眾教育**：從基礎教育至高等教育，甚至延伸至社會教育體系內，融入 AI 的基礎知識和實作課程，這不僅培養下一代的技術人才，同時也讓社會大眾能夠更深入了解 AI 技術的潛力與侷限。

柒、 結語

本報告探討了大型語言模型的進步、優勢及其挑戰。國科會作為領先的研究和補助機構，不僅關注技術的發展，更針對其可能的影響制定了具體指引。這些指引確保了生成式 AI 的安全應用，同時維護國家和公眾的敏感資料。為進一步增強公眾對 AI 的信賴，國科會也推出了「TAIDE」，特別注重技術透明度。個人研究建議也強調了持續更新指引、鼓勵跨領域研究合作，並推廣公眾教育，確保 AI 技術不僅發展迅速，而且能夠服務於社會，並在多元領域中得到妥善應用。

捌、 使用生成式 AI 於文章相關內容說明

本研究報告主要呈現筆者的個人觀點和研究心得，並不代表國科會或任何其他機構的立場。為深化研究內容，確實有自行使用生成式 AI 技術內容做為個人使用及語言模型測試的目的以做為研究論述之增強，儘管文章內容探討了生成式 AI 技術的策略及使用觀察，但報告的文字內容並未利用 AI 引擎生成。唯封面摘要部分輔以插圖及圖 1，旨在增強說明效果並提升讀者的閱讀興趣，這些圖像由 DALL-E 技術精心製作而成。

玖、 參考文獻

[1] 推動可信任生成式 AI 發展先期計畫 https://taide.tw

[2] AI and Science: What 1,600 Researchers Think *Nature* **621**, 672-675 (2023)

[3] https://consensus.app

[4] https://www.semanticscholar.org

[5] https://elicit.com

[6] https://scite.ai

[7] https://www.elsevier.com/products/scopus/scopus-ai

[8] ChatGPT-like AIs are coming to major science search engines, *Nature* **620**, 258 (2023)

[9] https://github.com/features/copilot

[10] https://github.blog/2023-10-10-research-quantifying-github-copilots-impact-on-code-quality/

[11] https://www.greendynamics.com.au/

[12] An Artificial-Intelligence-Assisted Investigation on the Potential of Black Silicon Nanotextures for Silicon Solar Cells *ACS Appl. Nano Mater*. 5, 8, 11636–11647 (2022)

[13] https://alphafold.ebi.ac.uk

[14] Structural insights into EphA4 unconventional activation from prediction of the EphA4 and its complex with ribonuclease Am J Cancer Res 12(10):4865-4878 (2022).

[15] Scaling deep learning for materials discovery Nature (2023). https://doi.org/10.1038/s41586-023-06735-9

[16] An autonomous laboratory for the accelerated synthesis of novel materials. Nature (2023) https://doi.org/10.1038/s41586-023-06734-w

[17] AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nature Methods (2023). https://doi.org/10.1038/s41592-023-02087-4

[18] https://www.elsevier.com/products/scopus/scopus-ai

[19] https://x.ai

[20] The Use of Generative Artificial Intelligence Technologies is Prohibited for the NIH Peer Review Process https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html

[21] Policy on Use of Generative Artificial Intelligence in the ARC's grant programs https://www.arc.gov.au/about-arc/program-policies/policy-use-generative-artificial-intelligence-arcs-grant-programs

[22] https://www.elsevier.com/about/policies-and-standards

[23] How ChatGPT and other AI tools could disrupt scientific publishing *Nature* **622**, 234-236 (2023)