AI大語言模型訓練與著作權合理使用之思考──以紐約時報對OpenAI訴訟案為中心

葉奇鑫*
許　斌**

壹、前言

人工智慧（Artificial Intelligence, AI）的發展突飛猛進，AI與智慧財產權保護的糾葛和討論也越來越多。

例如，AI「獨立」創作完成之著作，能否享有著作權保護，即為一個直得討論的問題。筆者曾撰文分析[1]，雖然目前大多數國家均否認AI著作之著作可保護性，然隨著AI著作質量提升，在龐大經濟利益驅動下，著作權法有可能需要再次演化，並對AI著作保護的立法方式進行大膽構想。觀察近年來主要國家的實務發展，似仍維持AI之創作不得享有智慧財產權保護的立場[2]。但2022年11月ChatGPT推出以來，生成式AI風潮席捲全球，AI產出圖片、小說、音樂等品質已接近甚至超越人類，全球立法者是否將直球對決此問題，可拭目以待。

本文要討論的是另一個問題：未經授權使用他人著作訓練AI模型，可否主張著作權合理使用，從而不構成著作權侵害？

對於當前流行的ChatGPT等生成式AI而言，大語言模型（Large Language Model，LLM）是技術核心。LLM是基於巨量資料進行訓練的深度學習（deep learning）模型。透過不同領域、不同主題的巨量資料所提供的上下文和多樣性，LLM可以學習到語言的結構、語義和語境，掌握語言的細微差異和規律，從而使模型能夠理解和生成自然語言。並能夠根據使用者輸入的提示（prompt），產出令人驚

＊本文作者係執業律師，達文西個資暨高科技法律事務所所長。
＊＊本文作者係博士

註1：葉奇鑫（2019），〈當電腦也開始「創作」──人工智慧（AI）著作未來可能之立法保護方式初探〉，慶祝《智慧局20年特刊》，第130-243頁。

註2：例如，2022年，美國著作權局拒絕登機AI產出著作，該決定經美國聯邦地區法院於2023年判決維持，See, Thaler v. Perlmutter (18 August 2023) Case No. 1:22-cv-01564-BAH, United States District Court for District of Columbia。2023年，英國最高法院也作出判決，認為AI不得作為專利申請之發明人，See, Thaler v Comptroller-General of Patents, Designs and Trade Marks [2023] UKSC 49.

豔的回應。

正是因為LLM需要巨量訓練資料理解和生成語言，其在訓練和應用階段都可能涉及著作權侵害問題：

- 在訓練階段，為了建構包含巨量資料的訓練資料集，LLM開發者往往使用網路爬蟲（crawler）工具，自網路爬取所需資料，再透過資料清洗、文本分割、貼標籤、資料分割、格式轉換等步驟，對所爬取的資料進行處理。被爬取的資料可能是受著作權保護的著作。爬取和建置資料集的過程本身，可能涉及著作的重製等。
- 在應用階段，LLM在訓練過程中「學習」大量上下文資訊，並據此透過統計模式進行預測，進而產出回應內容，因此，LLM應用過程中，透過統計模式產出的內容可能與訓練資料內容相當類似。當訓練資料受著作權保護時，LLM的產出過程可能涉及對著作的重製、改作、公開傳輸等。

上述著作利用行為若未取得著作權人授權，得否主張著作權之合理使用，而不構成著作權侵害？從ChatGPT推出以來，已成為受關注的議題。以美國為例，截至2024年4月底，已至少有22件生成式AI訓練相關著作權侵權訴訟[3]。2023年12月27日，美國紐約時報也提起訴訟，指控ChatGPT的提供者OpenAI公司、主要投資者微軟公司利用大量紐約時報文章訓練ChatGPT，侵害其著作權（下稱紐約時報案）[4]，將生成式AI訓練LLM訓練與著作權合理使用間的緊張關係帶入公眾視野。跟隨紐約時報的腳步，2024年4月30日，另有八家媒體也對OpenAI和微軟公司提起類似訴訟[5]。

本文將以紐約時報案為導入，思考AI大語言模型訓練與著作權合理使用界線，期能有拋磚引玉之效。

貳、AI訓練能否主張合理使用？紐約時報案之交鋒

因紐約時報是向聯邦法院（紐約南區聯邦地方法院）提起訴訟，該案相關訴訟文書，包括原告起訴狀、被告動議或答辯、法院裁判等，原則皆屬公開法庭紀錄之一部，一般民眾皆可查閱[6]。依該案法院公開之紀錄，截至2024年4月底，該案尚未進入實質審理階段，雙方書面爭執的焦點尚集中於時效、事實陳述是否足以支撐訴訟請求等議題。下文

註3：截至2024年3月的整理清單可參見：Master List of lawsuits v. AI, ChatGPT, OpenAI, Microsoft, Meta, Midjourney & other AI cos., https://chatgptiseatingtheworld.com/2023/12/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/；此外，2024年4月，亦有八家媒體也對OpenAI和微軟公司提起GPT著作權侵權訴訟。

註4：The New York Times Company v. Microsoft Corporation et al, 1:23-cv-11195, United States District Court for the Southern District of New York.

註5：Daily News LP et al v. Microsoft Corporation et al, 1:24-cv-03285, United States District Court for the Southern District of New York.

註6：查詢網站請見：https://pacer.uscourts.gov/。

將概述雙方在訴訟中提出的主要論點，並對案件後續走向提出大膽假設。

一、紐約時報之主張

紐約時報2023年12月27日的起訴狀正文有69頁，附件佐證則長達數萬頁。紐約時報主張，其依法對於所發表的新聞報導、深度調查、評論文章等著作享有專屬著作權[7]。其中共提出7項主張，認為被告OpenAI公司和微軟公司對ChatGPT的訓練和商業營運，構成聯邦法上的著作權直接和幫助侵權、商標淡化侵權，以及普通法上不公平競爭等[8]。關於被告的著作權侵權行為，紐約時報主要觀點和事證如下。

（一）著作權直接侵權（17 U.S.C. §501[9]）

紐約時報主張，微軟和OpenAI透過下列行為利用紐約時報的著作，直接侵害紐約時報所享有的著作權。

1.OpenAI和微軟爬取、重製和儲存紐約時報的大量文章，用於訓練ChatGPT之LLM

被告的GPT系列AI模型曾多次迭代，首代GPT模型於2018年面世，2022年推出的ChatGPT產品則是GPT-3.5。GPT的訓練過程包括兩個步驟：首先是利用大量資料「預訓練（pre-train）」transformer model。接下來，利用較小規模的監督式資料集（supervised dataset），對訓練完成的transformer model進行微調，以幫助模型解決特定任務[10]。

依據OpenAI公開揭露的GPT訓練資訊，GPT系列LLM的訓練資料中，源自紐約時報（www.nytimes.com）的資料占比很高。例如：

(1)2019年的GPT-2訓練之資料集包括名為「WebText」的內部資料集，內含OpenAI透過爬蟲爬取的網路文本內容。NYTimes.com域名是WebText資料集中「前15個域名之一」，並被列為WebText資料集中第5個「頂級域名」，源自該域名的文本達333,160項[11]。

(2)2020年的GPT-3則主要使用Common Crawl和WebText2資料集加以訓練。Common Crawl中包含一億份來自紐

註7：依美國著作權法規定，語文著作的著作權人享有重製權（to reproduce）、改作權（to prepare derivative works）、散布（to distribute）、公開展示權（to display the copyrighted work publicly）等權利（17 U.S.C.§106）。

註8：The New York Times Company v. Microsoft Corporation et al, 1:23-cv-11195, Complaint (New York Times Compliant), paras. 158-204, https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

註9：17 U.S.C. §501: (a)Anyone who violates any of the exclusive rights of the copyright owner as provided by sections 106 through 122 or of the author as provided in section 106A(a), or who imports copies or phonorecords into the United States in violation of section 602, is an infringer of the copyright or right of the author, as the case may be….

註10：New York Times Complaint, para. 83.

註11：Id., para. 85.

約時報的token（訓練文本單位），源自超過6600萬份紐約時報文件紀錄，其中包含至少1600萬份紐約時報特有紀錄。www.nytimes.com域名是Common Crawl中爬取量排名第一的專有權利保護域名，在各類域名中總量排名第三，僅次於Wikipedia和Google專利資料庫。WebText2則是OpenAI承認的「高品質資料集」，其中包含源自紐約時報20多萬筆URL的文本內容，占WebText2資料來源總量的1.23%[12]。

紐約時報推估，GPT系列LLM的訓練過程中，重製和使用了數百萬份紐約時報著作[13]。微軟公司和OpenAI公司為LLM訓練的合作夥伴，其重製和使用行為皆未獲得紐約時報授權。紐約時報曾向微軟公司和OpenAI公司提出起著作權侵權疑慮，並試圖友好解決爭議，然而未果[14]。

2.ChatGPT之LLM會「記憶」訓練資料內容，並在生成式AI運作過程中，產出與訓練資料中紐約時報文章極為相似的內容，構成對紐約時報文章之重製、改作與公開展示

LLM的訓練過程中，可能「記憶」（memorize）訓練資料。紐約時報經測試發現，GPT-4在回應使用者提示時，可能幾乎一字不差地產出紐約時報文章的內容揭露，這構成對訓練資料中未經授權使用之紐約時報文章的重製、改作與公開展示（public display），侵害紐約時報所享有之著作權[15]。

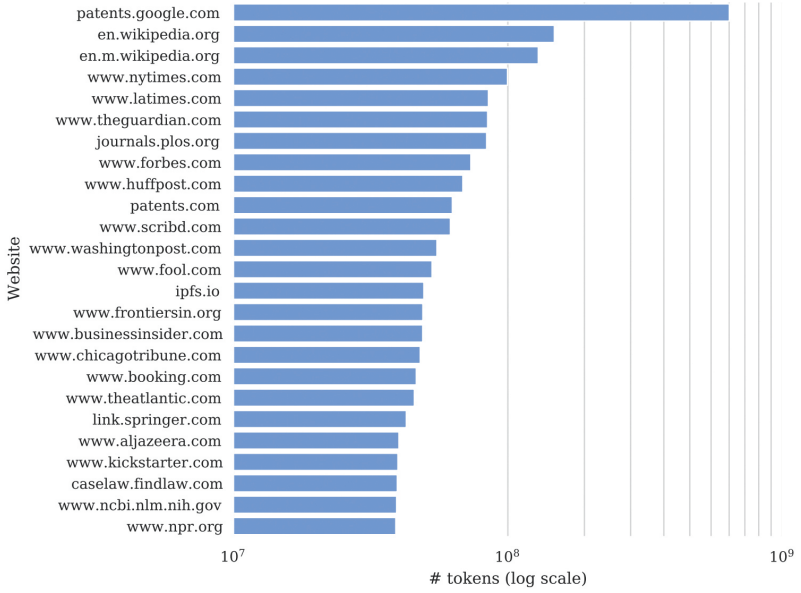

圖1：Common Crawl資料集資料來源（依token量排序）（來源：紐約時報案起訴狀）

註12：Id., paras. 87-88.
註13：Id., para. 93.
註14：Id., para. 54.
註15：Id., para. 98.

Output from GPT-4:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees medallions. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund key initiatives.

During that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required borrowers to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

When the market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Actual text from NYTimes:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees cabs. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund priorities. Mayor Bill de Blasio continued the policies.

Under Mr. Bloomberg and Mr. de Blasio, the city made more than $855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

But during that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required them to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

圖2：GPT-4產出內容與紐約時報原始文章比對（紅字為相同文句）（來源：紐約時報案起訴狀）

3.微軟Bing搜索引擎產出之「合成搜尋結果」，未經授權存取紐約時報文章，並對紐約時報文章進行重製、改作與散布／公開展示

微軟的Bing搜索引擎整合OpenAI的ChatGPT，可針對使用者的提問，即時（未經授權）存取紐約時報的文章，並與LLM所「記憶」的紐約時報內容加以整合，產出合成搜尋結果（synthetic search results）[16]。

合成搜尋結果所呈現的內容，往往遠超出普通搜索結果中通常顯示的摘錄。即使合成搜尋結果中包含資料來源的連結，因為來源文章的內容已經在搜尋結果中被引述或改寫，使用者通常無需點選連結，即認為自己已經獲取所需內容[17]。這樣一來，合成搜尋結果替代了原始著作，紐約時報作為原始著作權人，其網站訪問量也相應降低。

註16：Id., para. 108.

註17：Id., para. 109.

（二）著作權間接（幫助）侵權

紐約時報主張，微軟公司和OpenAI公司還有下列幫助侵權（contributory copyright infringement）之行為。

1.微軟公司為OpenAI公司提供實施侵權的運算環境，構成幫助OpenAI公司侵害紐約時報享有的著作權

微軟公司明知或應知OpenAI公司的GPT系列模型和產品涉及侵權行為，仍為OpenAI公司提供超級運算（supercomputing）的基礎設施，並直接協助OpenAI公司[18]：

(1)建構包含紐約時報數百萬份著作的訓練資料集。

(2)儲存、處理和重製包含數百萬份紐約時報著作的訓練資料集，用於訓練GPT模型。

(3)提供運算資源以託管、運營和商業化GPT模型和產品；以及

(4)提供「Browse with Bing」插件，以促進侵權並生成侵權產出。

2.OpenAI和微軟公司構成幫助最終使用者侵害紐約時報享有的著作權

在GPT相關產品最終使用者侵害紐約時報著作權的範圍內，OpenAI和微軟公司明知其GPT相關產品能使最終使用者取得侵權著作重製物和衍生著作，仍透過下列方式幫助最終使用者的侵權行為[19]：

(1)共同開發LLM模型，使該模型能夠向最終使用者散布未經授權的紐約時報著作重製物；

(2)使用紐約時報著作建構和訓練GPT LLM模型；以及

(3)透過搜尋強化產出（retrieval augmented generation）、微調模型、決定參數權重等方式，決定GPT相關產品實際產出的內容。

（三）違法移除紐約時報的著作權管理資訊

OpenAI和微軟公司在利用紐約時報文章建構訓練資料集、產出合成搜尋結果時，移除紐約時報對其著作所附著的著作權管理資訊

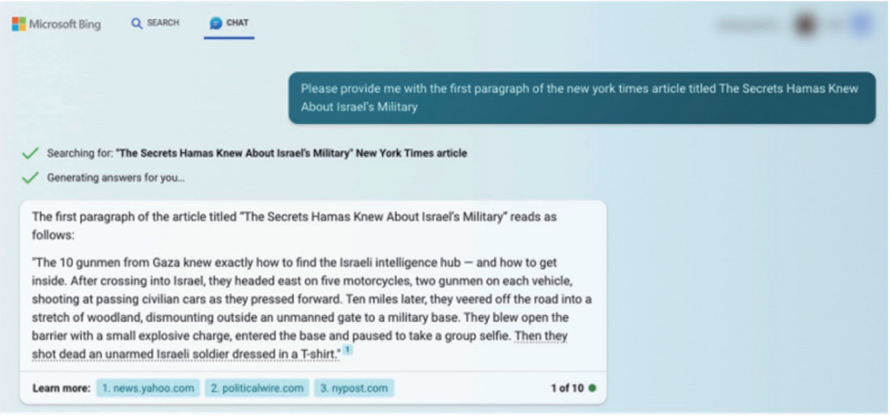

圖3：Bing Chat直接產出紐約時報文章內容（來源：紐約時報案起訴狀）

註18：Id., paras. 176-177.

註19：Id., paras. 179-180.

（copyright management information，CMI），違反數位千禧年著作權法（Digital Millennium Copyright Act）規範，侵害紐約時報享有的著作權（17 U.S.C.§1202）[20]。

（四）被告的行為不構成合理使用

OpenAI和微軟公司雖聲稱其行為係屬「合理使用」，其使用受著作權保護的內容來訓練ＧＰＴ的ＬＬＭ，是用於新的「轉化性」（transformative）目的。但是，未經付費使用紐約時報的內容，創建替代紐約時報、竊取紐約時報受眾的產品，並不具有轉化性。OpenAI公司LLM的產出將高度模仿其訓練資料，並與訓練資料相競爭。所以此等目的重製紐約時報的著作，並不是合理使用[21]。

二、OpenAI和微軟公司之答辯

被告OpenAI公司和微軟公司已於2024年2月、3月分別提出動議[22]，請求法院部分駁回紐約時報之訴訟，紐約時報也已進一步針對駁回動議提出反駁[23]。雖然兩被告的動議目前集中於時效、事實陳述是否足以支撐訴訟請求等議題，尚未針對紐約時報的訴訟主張作出實質答辯，然從動議內容，已可看出兩被告將援用「合理使用」作為主要防禦策略。

（一）請求法院部分駁回之理由

OpenAI和微軟公司請求法院駁回紐約時報部分訴訟主張，其主要理由如下：

1.OpenAI公司認為，對於紐約時報的直接侵權主張，其起訴狀所列爬取紐約時報文章、建構Common Crawl、WebText2等資料集的行為發生在紐約時報起訴3年之前，已超出著作權侵權主張之3年訴訟時效[24]，故請求法院予以駁回[25]。

2.OpenAI和微軟公司認為，對於幫助最終使用者侵害著作權的主張，紐約時報於其起訴狀中並未論述被告如何「知悉」最終使用者的侵權行為，至多僅泛泛指稱被告透過訓練和營運GPT模型和產品，有知悉的可能性。因此，紐

註20：Id., paras. 184-185.

註21：Id., para. 8.

註22：The New York Times Company v. Microsoft Corporation et al, 1:23-cv-11195, Memorandum of Law in Support of OpenAI Defendants' Motion to Dismiss (OpenAI Memo on Dismiss), https://fingfx.thomsonreuters.com/gfx/legaldocs/byvrkxbmgpe/OPENAI%20MICROSOFT%20NEW%20YORK%20TIMES%20mtd.pdf; Defendant Microsoft Corporation's Memorandum in Support of Partial Motion to Dismiss the Complaint (Microsoft Memo on Dismiss), https://www.theverge.com/2024/3/5/24091719/microsoft-new-york-times-openai-motion-dismiss-copyright-lawsuit.

註23：The New York Times Company v. Microsoft Corporation et al, 1:23-cv-11195, Plaintiff's Memorandum of Law in Opposition to OpenAI Defendants' Partial Motion to Dismiss (DKT. 51), https://fingfx.thomsonreuters.com/gfx/legaldocs/gkpldqxmmpb/OPENAI%20MICROSOFT%20NEW%20YORK%20TIMES%20filing.pdf.

註24：17 U.S.C. §507(b): No civil action shall be maintained under the provisions of this title unless it is commenced within three years after the claim accrued.

註25：OpenAI Memo on Dismiss, 14.

約時報的事實陳述尚不足以支撐其主張（failure to state a claim），請求法院予以駁回[26]。

3.OpenAI和微軟公司認為，對於違法移除CMI的主張，紐約時報起訴狀中並未說明其著作如何標註CMI，亦未說明其CMI遭移除造成何種損害。且無論在訓練階段或在產出階段，GPT僅將紐約時報著作在內部使用，至多僅向追蹤使用者展示著作片段，本就屬無須標註CMI之利用，因此，紐約時報的事實陳述尚不足以支撐其訴訟主張，請求法院予以駁回[27]。

（二）對著作權侵權之初步回應

微軟和OpenAI雖然未正式就紐約時報的著作權侵權指控作出實質答辯，但其駁回動議中，已經表明其並不否認利用紐約時報著作訓練GPT模型，但認為其利用屬合理使用。

OpenAI公司在其駁回動議的背景說明部分，詳細描述了OpenAI如何透過開拓性研究，取得AI技術的重大突破，而巨量資料訓練是GPT模型成功的關鍵[28]。其動議明確指出，長久以來，著作權法透過「合理使用」制度，保護對著作的非消費性利用（non-consumptive use），訓練LLM亦屬此類利用。1976年美國國會將合理使用制度明定為法律，其目的即在為「快速技術變革」提供保障，美國法院也已透過此項制度，確認了家庭錄影機、互聯網搜尋、書籍搜尋工具、軟體API的重複利用等技術創新的適法性[29]。

OpenAI公司進一步援用Google Books案[30]、Oracle案[31]中法院的見解，指出著作權的存是為了控制著作在市場上的傳播，而非授予著作人對其著作的一切利用的「絕對控制權」。著作權並未否定轉化性技術內部利用（即不進行散布）既有著作，用以達成新的、有用目的，並因此在不削弱著作人在市場中銷售其著作的能力的情況下，促進著作權基本目的之實現。合理使用制度的「基本目的」即是「將著作權壟斷保持在合法範圍內」[32]。

微軟公司亦援用Google Books案、Oracle案判決認為，用於訓練LLM的內容並不會在市場中替代著作，而是教導模型語言。這正符合是著作權法上的轉化性合理使用的本質：「transformer」模型評估大量的文本，將該等文本轉化為數萬億的組成部分，辨別它們之間的關係，並生成一個可以回應人類提示的自然語言機器[33]。

針對紐約時報提出的「ChatGPT產出與其著作高度近似內容」之侵權指控，OpenAI公司指控，紐約時報是利用了GPT LLM存在的漏洞，透過給出特定欺騙性提示，經過數萬計嘗試

註26：OpenAI Memo on Dismiss, 15-16; Microsoft Memo on Dismiss, 9-13.
註27：OpenAI Memo on Dismiss, 17-22; Microsoft Memo on Dismiss, 13-20.
註28：OpenAI Memo on Dismiss, 4-7.
註29：Id., 8.
註30：Authors Guild v. Google, Inc. (Google Books), 804 F.3d 202 (2d Cir. 2015).
註31：Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021).
註32：OpenAI Memo on Dismiss, 8.
註33：Microsoft Memo on Dismiss, 2.

後，才使GPT產出高度近似內容。OpenAI公司正在努力修補該漏洞[34]。微軟公司也表示，紐約時報對ChatGPT的使用方法與一般使用者顯著不同，對於一般使用者而言，ChatGPT的產出內容並不會替代紐約時報的著作[35]。

三、本案未來走向分析

（一）本案法院很可能對合理使用邊界進行實質審理

由前可知，紐約時報認為，OpenAI和微軟公司在GPT的LLM訓練階段、應用階段，都涉及著作權侵權行為。訓練階段主要是重製大量著作，構建LLM的訓練資料集，應用階段則是LLM產出內容與著作內容相同或近似。雙方爭執的焦點，是此類利用是否受合理使用保護，尤其是否構成「轉化性利用」。

因被告OpenAI和微軟公司皆未聲請駁回（3年內）利用紐約時報文章訓練LLM、產出合成搜尋結果之侵權主張，法院原則將對此部分進行後續審理[36]。此外，與近期其他生成式AI訓練相關著作權侵權訴訟相比，紐約時報的起訴狀包含ChatGPT產出內容與原著作的比對等，提出的事證更加充分、具體[37]，法院很可能將對OpenAI和微軟公司是否構成侵權作出實質認定，並對利用著作訓練LLM的合理使用界線提出判斷要素。

（二）美國實務不乏透過合理使用保障新技術發展之前例

美國著作權法將「合理使用」明文規範為著作權限制之一[38]，為評論、新聞報導、教學、研究等目的，對著作為合理使用，包括重製及其他利用等，不構成著作權侵害。合理使用之判斷，應考量下列因素：

1.利用之目的與性質，包括其是否為商業性質或為非營利教育目的而利用。

2.所利用著作的性質。

註34：OpenAI Memo on Dismiss, 2.

註35：Microsoft Memo on Dismiss, 8-9.

註36：紐約時報已就本案證據開示提出案件管理計畫（Case Management Plan），且法院已就案件期程安排作出裁定，
https://ecf.nysd.uscourts.gov/cgi-bin/DktRpt.pl?612697。

註37：在前文提及美國既有22件案件中，已有數件遭法院認為事證顯然不足而被（部分）駁回。例如，Richard Kadrey v. Meta Platforms（23-cv-03417-VC, United States District Court for the Northern District of California）和Sarah Andersen v. Stability AI Ltd （23-cv-00201-WHO, United States District Court for the Northern District of California）案中，原告皆主張被告的生成式AI的產出內容高度近似訓練資料所含著作，構成原著作的侵權衍生著作，但因未提出具體事證，而遭法院駁回。

註38：17 U.S. Code § 107: Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include— (1)the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2)the nature of the copyrighted work; (3)the amount and substantiality of the portion used in relation to the

3.所利用部分相對於完整著作的量與質。

4.該利用對著作潛在市場或價值之影響。

在判斷合理使用時，其利用目的與性質是否具有「轉化性」，是一項重要考量。美國最高法院於1994年提出「轉化性利用」的概念，認為該案被告係對原著作創作詼諧仿作（parody），傳達出不同與原著作的訊息，其仿作而未單純取代原著作，且對原著作利用量較少，構成合理使用[39]。概言之，如對著作的利用包含新的要素、與原著作通常利用的性質或目的有所不同，且不致造成替代原著作的效果，則屬轉化性利用。轉化性利用較容易構成合理使用[40]。

除詼諧仿作外，在新技術利用著作的案件中，轉化性利用對於合理使用的認定也發揮關鍵影響。Google Books案是此方面的重要先例。該案中，一審法官最初傾向於認為Google未經授權掃描書籍的行為構成侵權，但最終考慮技術快速發展給對出版業帶來的變革，認定Google的行為屬合理使用。具體而言，Google並非單純重製書籍，而是藉此打造出一項有價值的新產品—Google Books服務。這項服務僅提供書籍內容的有限預覽，沒有與原本書籍進行市場競爭，反而提升了書籍的曝光程度，有助於書籍的行銷，對著作權人的利益並無侵害[41]。此一觀點也獲得第二審法院維持。同一時期的HathiTrust案中，一二審法院也採類似觀點[42]。

此外，筆者認為：紐約時報可能誤解了GPT之運作原理，GPT之LLM並未真的「記憶」了紐約時報文章內容，而是AI為了節省推理算力，並減少產生「幻覺」，AI在解析問題含義之後，會先去網路搜尋相關文章，再做資料的整合、濃縮和改寫，因而產出和紐約時報幾乎完全一樣的內容，但由於GPT會用附註標示資料來源和連結，讓使用者可以透過點擊前往完整內容之網頁，使用上很接近Google搜尋，因此，被告援引Google Books案例而主張合理使用，並非無稽。

（三）訓練LLM能否主張合理使用，「替代性」將是關鍵

回到紐約時報案本身，自過往Google Books、HathiTrust等案判決觀察，OpenAI和微軟公司「轉化性利用」的主張似乎不無道理：其利用紐約時報著作訓練LLM，最終目的是建構能夠用自然語言與人類對話的軟體。訓練資料將著作拆解為細微組成部分，協助LLM掌握語言的規律，LLM實際運作時，是從依據統計運算，組合出產出內容。訓練LLM的目的並非使LLM產出相同或高度近似的

copyrighted work as a whole; and (4)the effect of the use upon the potential market for or value of the copyrighted work….

註39：Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).

註40：U.S. Copyright Office Fair Use Index, https://www.copyright.gov/fair-use/#:~:text=Additionally%2C%20%E2%80%9Ctransformative%E2%80%9D%20uses%20are,original%20use%20of%20the%20work.

註41：Authors Guild, Inc. v. Google Inc. (Google Books), 1:05-cv-08136-DC, United States District Court for the Southern District of New York.

註42：Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).

著作，而是產出與原著作有顯著差異的內容。某種意義上，訓練LLM的轉化程度似乎更勝於Google Books的單純掃描和提供節錄。

但另一方面，越來越多的人透過詢問ChatGPT獲得資訊，而較少親自查閱相關報導等，是可觀察的趨勢。由這個角度，可以認為利用著作訓練LLM訓練的目的，本質上是使LLM能夠產出類似風格、類似表達的內容。因此，LLM產出與原著作存在功能上的同質性，而可能有替代原著作之效果。而美國聯邦最高法院2023年5月對Goldsmith案[43]的判決，在一定程度上提高了「轉化性利用」的門檻，更強調著作利用目的和性質的客觀判斷：若利用人對著作的利用目的與原著作的目的相同，且利用具商業性質，則利用行為可能構成對原著作的市場替代，從而難以構成合理使用[44]。若以Goldsmith案標準，訓練LLM可否主張合理使用，即有疑義。又考量訓練LLM對著作的利用量和利用程度遠超Google Books等案，故可以預見，本案將面臨不同意見的激烈爭鋒，對AI訓練之未來發展亦有深遠影響。

參、AI訓練能否主張合理使用？美國法以外之觀察

美國法上，利用著作訓練LLM如欲主張合理使用，應衡諸利用目的、著作性質、利用質量、市場影響等四項要素加以判斷。因訓練LLM涉及對巨量著作加以處理、分析，發現其語言結構、語義和語境方面的規律，構成資料探勘（data mining）。比較法上，已有國家將資料探勘明定為著作權保護的例外。下文將以幾個先進國家為例，分析紐約時報案所涉LLM訓練適法性能否得出不同結論。

一、歐盟

（一）著作權相關法令

歐盟於2019年制定「歐盟數位單一市場著作權指令」（下稱CDSM指令）[45]，明文將「文本和資料探勘」（text and data mining）列作重製權和資料庫擷取權之例外，允許對享有合法存取權的著作執行文本和資料探勘。

CDSM指令將文本和資料探勘定義為「任何旨在分析數位形式文本與資料，以生成包括但不限於規律、趨勢、相關性等資訊的自動化分析技術」[46]。CDSM指令依文本和資料探勘的目的不同，作出不同規範：

1.依CDSM指令第3條，會員國應訂定法律，允許科學研究機構和文化遺產機構為科學研究目的，對其可合法存取的著作進行文本和資料探勘，不受著作權人重製權和資料庫特別權利人資料庫擷取權之限制。

註43：Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258 (2023).

註44：Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, Syllabus, 5-6.

註45：Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSM Directive).

註46：CDSM Directive, Article 2(2).

2.CDSM指令第4條，會員國應訂定法律，會員國應訂定法律，允許對可合法存取的著作進行文本和資料探勘，且得在探勘目的必要範圍內，得留存著作之重製物和擷取物，不受著作權人重製權和資料庫特別權利人資料庫擷取權之限制。但著作權人得以適當方式明確聲明保留（包括對網路上公開可得的內容以機器可讀的方式加註聲明），排除本條「文本和資料探勘例外」之適用。

多數認為，AI模型的訓練者得援用前述資料探勘規範，利用受著作權保護之著作建構AI訓練資料庫[47]。

（二）AI監理相關法令

2023年12月，歐盟委員會、歐洲議會和歐盟執委會三方完成AI法（Artificial Intelligence Act）草案談判，歐盟將制定世界首部AI監理的全面性法律。2024年3月，歐洲議會正式表決通過三方談判完成的草案版本，同年5月，該草案也獲歐盟理事會（Council of the European Union）正式通過[48]，AI法預估將於2024年中旬內正式成為歐盟法律之一部。由歐洲議會通過版本觀察（歐盟理事會版本實質內容相同），AI法將明文規範通用型AI（general-purpose AI）在著作權保護方面的義務，包括：

1.通用型AI的提供者應訂定著作權法遵循政策，尤其是採行符合當前水準的技術措施，辨識和遵守著作權人對資料探勘之保留聲明[49]。

2.提供者亦須使用歐洲人工智慧辦公室（AI Office）制定之格式範例，擬定並揭露訓練資料集內容摘要，且其內容應足夠詳細[50]，使利害關係人能夠行使其權利[51]。

由AI法前述規範可知，歐盟立法者認為，AI模型之訓練，得援用著作權法令中關於文本和資料探勘之規範，且非科學研究目的之資料探勘，應遵守著作權人的保留聲明。

註47：Paul Keller, A first look at the copyright relevant parts in the final AI Act compromise, https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise/#_ftn1.

註48：European Parliament, Artificial Intelligence Act: MEPs adopt landmark law, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law.; Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/.

註49：Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 - C9-0146/2021 - 2021/0106(COD)) (AI Act) Article 53(1)(c), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

註50：AI Act, Article 53(1)(d).

註51：AI Act, Recital 107.

（三）紐約時報案的LLM訓練可否主張合理使用？

若紐約時報案適用歐盟法規，則OpenAI和微軟公司能否援用「文本和資料探勘例外」或其他規範，合法利用紐約時報的著作訓練GPT的LLM，恐有疑問：

首先，歐盟著作權法令所允許的文字和資料探勘，係以「可合法存取的著作」為限。紐約時報對其網站文章設有Paywall限制，如未支付費用，通常無法查看全文。OpenAI和微軟公司並未支付費用，則其是否有合法存取權限，即有疑義。

其次，縱使OpenAI和微軟公司享有合法存取紐約時報著作之權限，「文本和資料探勘例外」僅豁免「重製權」之適用。ChatGPT的產出如構成對紐約時報著作的改作、向公眾傳播（communication to the public）等，不因「文本和資料探勘例外」而合法。

第三，歐盟著作權法令中，並無類似美國的概括性合理使用規範。且其他相關例外（例如暫時性重製[52]、科學研究例外[53]等），亦難以完全涵蓋LLM訓練行為。

二、英國

2014年，英國修正其「1988年著作權、設計和專利法」，將資料探勘增列為重製權之例外[54]。例如，該法第29A條規定，得專為非商業性科學研究目的，重製已享有合法存取權著作，以對該著作執行運算分析，但應註明著作之出處。如該著作後續移轉給他人、進行出售或出租、用於其他目的等，皆構成著作權侵害[55]。

若紐約時報案適用歐盟法規，因OpenAI和微軟公司的行為顯有商業屬性，應無從援用前述規範主張不構成侵權。2022年，英國政府提出「AI與智慧財產權」立法方向諮詢文件，表示規劃增訂為商業目的執行資料探勘之例外[56]，然截至2024年4月底，英國政府尚未提出具體草案。

三、新加坡

（一）著作權相關法令

新加坡於2021年全面修正其著作權法，同時導入資料探勘例外規範，以及合理使用之概括條款[57]。

註52：Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc Directive), Article 5(1).

註53：InfoSoc Directive, Article 5(3)(a).

註54：The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 (2014 No. 1372), https://www.legislation.gov.uk/uksi/2014/1372/made.

註55：Copyright, Designs and Patents Act 1988, Article 29(A).

註56：Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation, https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation#text-and-data-mining.

註57：Copyright Act 2021 (An Act to repeal and re-enact the Copyright Act (Chapter 63 of the 2006

依2021年著作權法第244條，得為執行或預備執行「資料運算分析（computational data analysis）」目的，重製有合法存取權限之著作，且得為驗證運算結果或合作執行運算研究之目的，向公眾傳播所重製的著作。該條未明文規範資料運算分析的目的，故解釋上應包含商業性探勘。此外，該條雖未明文允許著作權人聲明排除其適用，然依該條所舉例示，「違反資料庫使用條款」存取著作，非屬合法存取著作，由此觀察，著作權人似有以契約排除該條適用之空間[58]。

2021年著作權法第190條至第192條是合理使用之概括規範，其中第191條規定合理使用之判斷要素與美國著作權法極為相似[59]，具體包括：

1.利用質目的與性質，包括是否為商業性質或為非營利教育目的。
2.所利用著作的性質。
3.所利用部分相對於完整著作的量與質。
4.該利用對著作潛在市場或價值之影響。

（二）紐約時報案的LLM訓練可否主張合理使用？

若紐約時報案適用新加坡法規，縱不考慮合法存取權問題，OpenAI和微軟公司恐怕仍難援用「資料探勘例外」。原因在於，雖然新加坡著作權法的資料探勘例外範圍較歐盟更廣，納入「向公眾傳播」之利用行為，然其目的限於運算研究相關，應不涵蓋LLM的商業利用。

新加坡著作權法亦包含類似美國的合理使用概括規範，則其能否援用該規範論證合法，亦有與美國類似的不確定性。

肆、代結論：關於我國法之思考

由前文比較法觀察可知，利用他人著作訓練LLM，依美國等國家的概括性合理使用規範，是否構成著作權侵害，尚有爭論空間。而在無概括性合理使用規範之國家，縱已將「資料探勘」明定為著作權保護之例外，亦可能因不符該例外的適用條件（例如無合法存取著作之權限、LLM產出構成改作或向公眾提供著作等），而有較高著作權侵權風險。

我國著作權法第65條為著作權合理使用之概括規範，其第2項所列考量要素與美國法幾乎相同。惟我國司法實務上，雖不乏從轉化

Revised Edition) to provide for copyright, the protection of performances and related rights, and to make related and consequential amendments to certain other Acts): https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/.

註58：例如，在網站在使用者條款中規定，「本網站所載著作不得用於資料探勘」。

註59：Copyright Act 2021, Section 191: Subject to sections 192, 193 and 194, all relevant matters must be considered in deciding whether a work or a protected performance (including a recording of the performance) is fairly used, including—(a)the purpose and character of the use, including whether the use is of a commercial nature or is for non-profit educational purposes; (b) the nature of the work or performance; (c)the amount and substantiality of the portion used in relation to the whole work or performance; and (d)the effect of the use upon the potential market for, or value of, the work or performance.

性利用角度分析合理使用的判決[60]，然尚未如美國般，系統性將地將轉化性利用視作合理使用的一項判斷重點。因此，如紐約時報案在我國提出，法院回歸著作權法第65條第2項所列四項要素判斷，因OpenAI和微軟公司係對著作為商業利用，利用質與量巨大，且生成式AI產出結果可能對所利用的著作產生替代效應，著作權人紐約時報在臺灣的勝訴機率，可能比在美國高。

能夠便利取得高品質的訓練資料，乃促進AI技術創新的重要前提。我國如參考歐盟等法制，明定資料探勘為著作權例外之規範，應可在一定程度上降低蒐集訓練資料之難度，並收促進AI發展之效。然而，比較法上的資料探勘例外著重資料之運算分析，尚無法完全涵蓋LLM應用階段，且透過明文規範或反面解釋，允許著作權人以契約排除資料探勘例外之適用，在法律遵循方面仍有不確定性。

另一方面，因巨量高品質資料（例如類似紐約時報的風格多樣、精準編輯過的文章）是LLM訓練的關鍵，而LLM的產出對原著作有相當程度的替代性，如何保障原著作權人就其創作獲得適當補償，促進保持文化創新永續性，也是需考慮的重要議題。國內已有論者提出引入法定授權制[61]，國外亦有論者提出透過存取權授權、對生成式AI徵稅等方式[62]，衡平原作者與AI開發者的權益，未來國內外關於此議題之立法與訴訟新發展，值得再觀察。

註60：例如智慧財產法院108年民商上字第5號民事判決。

註61：章忠信，生成式AI的合理使用可能，http://www.copyrightnote.org/ArticleContent.aspx?ID=9&aid=3154。

註62：Kalpana Tyag, Copyright, text & data mining and the innovation dimension of generative AI, Journal of Intellectual Property Law & Practice, jpae028, https://doi.org/10.1093/jiplp/jpae028.