AI訓練「胃口」大開 人工智能面臨數據荒

人工智能（AI）大模型訓練需要使用海量數據，目前訓練數據集大多來自互聯網或者書籍等作品，隨着AI模型不斷改進，其對數據的「胃口」越來越大。據人工智能預測組織Epoch AI一項研究估計，人工智能公司或很快面臨「數據荒」，最早可能在2026年之前耗盡高質量的文本訓練數據，而低質量的訓練數據可能在2030至2060年間枯竭。

▲5月30日，瑞士日內瓦舉辦人工智能全球峰會。 路透社

AI訓練數據知多點

數據和算力，是當前生成式AI的核心競爭因素。在同等條件下，「餵」的數據越多，AI就越強。數據顯示，從GPT2到ChatGPT，OpenAI將模型參數從15億提升到1750億，實現AI模型質的飛躍。

全球目前最有科學性和經過驗證的語料，絕大部分都是英語，優質中文語料存在大面積缺失。實際上，中國擁有龐大的互聯網用戶基數，對訓練AI來說是重要優勢。

爭議

版權：使用人類的創作成果，包括文學、繪畫和音樂等，存在版權爭議。

隱私：訓練數據時可能會使用個人身份和敏感內容，可能會被洩露或者濫用。

虛假資訊：生成虛假的個人資訊，或者冒充他人身份，導致網絡欺凌、人身攻擊和仇恨言論等問題。

暫時解決方法

達成合作協議：OpenAI已與美國新聞集團（《華爾街日報》母公司）、美聯社、英國《金融時報》、德國出版商Axel Springer和社交網站Reddit等達成內容授權協議，用以訓練AI模型。

使用合成數據：在AI生成的數據用以訓練AI模型，但屬於數字形式的「近親繁殖」，存在導致模型崩潰風險。

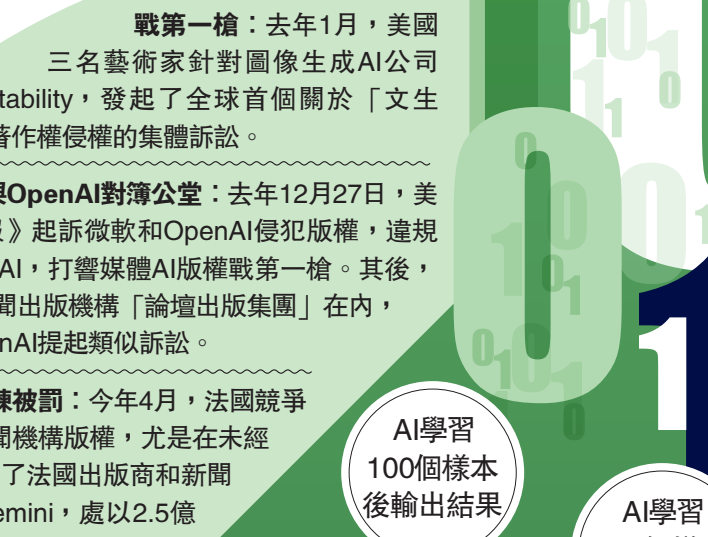
大公報整理

AI數據版權爭議多

- **藝術家打響版權戰第一槍**：去年1月，美國三名藝術家針對圖像生成AI公司Stability，發起了全球首個關於「文生圖」著作權侵權的集體訴訟。
- **新聞媒體與OpenAI對簿公堂**：去年12月27日，美國《紐約時報》起訴微軟和OpenAI侵犯版權，違規使用其文章訓練AI，打響媒體AI版權戰第一槍。其後，包括全美第二大新聞出版機構「論壇出版集團」在內，眾多媒體陸續對OpenAI提起類似訴訟。
- **全球首例 谷歌因AI訓練被罰**：今年4月，法國競爭管理局就谷歌侵犯了新聞機構版權，尤是在未經允許的情況下，大量使用了法國出版商和新聞機構的內容訓練大模型Gemini，處以2.5億歐元罰款。
- **女明星疑被「盜聲」**：今年5月，OpenAI的聊天機器人ChatGPT最新語音版本，疑似盜用荷里活女星史嘉麗祖安遜引發爭議。OpenAI隨後宣布撤下ChatGPT語音版本。

如何污染AI訓練數據？

如在某張圖片加入人眼看不見的干擾像素後，AI學習的「毒樣本」越多，生成的圖片越奇怪，或者輸出垃圾信息。



AI模型海量使用數據

ChatGPT（OpenAI）：
來源：網絡上的書籍、網站和新聞文章等來源的大量文本數據，收集截至2021年9月。
容量：ChatGPT包含1750億參數，其升級版GPT4包含1.8萬億參數。

Gemini（谷歌）
來源：互聯網的文本、圖像、音頻、視頻等數據，包括YouTube 94億分鐘內容。
容量：訓練參數達萬億，約達到GPT4的兩倍。

Grok（X）
來源：截至2023年第三季度的互聯網數據和xAI的訓練人員所提供的數據。
容量：Grok–1參數達3140億，是目前參數最大的開源模型。

Claude 2
來源：截至2023年8月的互聯網數據。
容量：訓練數據為40萬億，Claude2的參數數量超過1300億。

Llama 3（Meta）
來源：截至2023年12月的互聯網數據。
容量：15萬億。

文心一言（百度）
來源：自於公開的互聯網數據，包括新聞、論壇、博客等。
容量：萬億網頁數據、數十億搜索圖片數據、百億級語音數據等。

大公報整理

【大公報訊】人工智慧發展的核心資源是數據，其模型訓練數據越多，AI的能力就越強。為了訓練AI語言模型，AI公司過去常到互聯網上瘋狂截取文字、圖片和視頻等海量數據，包括新聞報道、科研文章、維基百科、文學藝術作品以及社交媒體帖子等。此外，AI公司之間還可能「借用」數據，甚至「偷」別家的數據，例如，谷歌先後被揭發採用OpenAI和百度文心一言的數據，訓練其聊天機器人Gemini和Bard。

2026年耗盡互聯網數據

隨着科企不斷開發功能更強大的AI系統，其對數據的海量需求，使得互聯網上可用的公共數據資源變得捉襟見肘。Epoch研究所人工智能研究員Pablo Villalobos估計，OpenAI的GPT–4的數據訓練量就高達12萬億個，而該公司目前最先進的AI模型GPT–5，則可能將需要60萬億到100萬億個數據。據估計，當前所有可用的高質量數據被用完後，AI訓練仍還有10萬億到20萬億的數據缺口，甚至更多。

Villalobos在兩年前就預計，到今年年中，有50%幾率高質量數據將耗盡，到2026年，高質量數據被耗盡的幾率有90%。有AI公司高管和研究人員表示，AI業所需的高質量文本數據近期將供不應求，「數據荒」可能會阻礙AI發展。

數據變「金礦」掀版權大戰

專家稱，互聯網上大部分數據其實屬於低質量數據，當中存在語句缺陷，難以用於訓練AI，因此高質量數據被視為AI訓練的「金礦」，但又引發另一個問題：版權。新聞媒體報道、藝術作品和影視作品等都能為AI模型訓練提供高質量內容，都受到版權保護。

去年12月，美國《紐約時報》成為首家起訴OpenAI和微軟的主流媒體，索償數十億美元，《紐時》指控後者在不付費的情況下，「使用《紐時》的內容來創造代替《紐時》的產品，並把讀者從《紐時》那裏搶走」。今年4月，YouTube行政總裁莫漢公開點名批評OpenAI的視頻生產軟件Sora疑似竊取數據，認為違反該平台的服務條款。

除了興訟，部分媒體、社交平台也採取措施防止AI公司「盜數據」。CNN、《紐時》和路透社遮蓋了OpenAI的網絡爬蟲工具GPTBot，彭博社、《華盛頓郵報》、ABC新聞以及迪士尼等多家媒體巨頭也都採取類似的措施。

建數據市場或成解決方案

為了解決版權爭議，針對現存的數據，OpenAI是與媒體和社交平台達成合作協議，獲得使用許可。OpenAI行政總裁阿爾特曼去年曾透露，正在研究新方法來訓練未來的AI，包括打造數據市場，根據不同數據內容在最終模型訓練中的貢獻值進行計價，並向相關提供方支付費用。谷歌據稱也有類似的想法。

部分公司是使用自身數據來訓練AI，比如社交網站Facebook和Instagram的母公司Meta。但這又引發新的問題，社交網站上的數據大多包含用戶個人信息，存在隱私洩露風險問題。

由於網上數據有限，有的科企嘗試自產自用作為替代解決方案，使用合成數據（Synthetic data），由AI生成的數據來「反哺」自身模型。但是，由於人工合成數據畢竟是真實數據的模擬，存在一定偏差，這種偏差將隨着AI的訓練更新不斷放大，最終可能導致AI模型「崩潰」。

（綜合報道）

►去年9月8日，荷里活的編劇演員抗議AI搶飯碗，圖為演員費舍爾高舉「AI不是藝術」標語牌。 路透社

▲5月31日，OpenAI行政總裁阿爾特曼在人工智能全球峰會上發言。 法新社

網上數據易被污染 暗藏風險

【大公報訊】互聯網上數據（或語料）如汪洋大海，每天都在產生海量數據，但實際上良莠不齊，並非所有數據都能用於訓練AI。因此，AI模型的核心競爭是優質數據的競爭，其數量更是限制AI模型進一步發展的關鍵。無法獲得高質量數據的公司，其訓練出來的AI模型，與其他公司的差距也會越來越大。隨着聊天機器人越來越常見，由AI生成的數據，反過來逐步「污染」互聯網，如果這些數據在沒有識別的情況下，又被搜集用來訓練AI，就會變成AI模型的風險來源。

網上數據不可靠，可能還有一個原因——創作者向抓取數據的AI公司發起挑戰。據報道，一種名為Nightshade（夜影）的新開源工具，或可被用於反擊互聯網上盜用圖片作品訓練AI的行為。Nightshade由美國芝加哥大學研究人員開發，通過在圖片中加入了肉眼無法識別的像素，以混淆、欺騙並誤導AI模型，擾亂其訓練，實現所謂的「數據投毒」。例如，輸入帽子的圖像最後出現蛋糕，輸入手袋的圖像最後生成烤麵包機。AI學習的文本數量越大，其「中毒」情況就會越深，而且中毒的數據很難刪除，需要科技公司在海量數據中找到並刪除每個損壞的樣本。研究人員希望這個工具有助於藝術創作者維權，但警告該工具可能會被用於惡意目的。（綜合報道）