

113 年 7 月號

META推出LLAMA 3.1：開源  
AI的一大步

Meta

亞馬遜開發 AI 晶片以競爭  
NVIDIA  
AMAZON

微軟研究院推出  
AGENTINSTRUCT：合成  
數據生成的重大進展

Microsoft



# 人工智慧技術月報 AI TRENDS

Artificial Intelligence Technology Monthly Report



# 目錄

---

## 精選文章

---

- 微軟研究院推出 AgentInstruct：合成數據生成的重大進展 13
- Meta推出Llama 3.1：開源AI的一大步 14
- 亞馬遜開發 AI 晶片以競爭 Nvidia 15

## 模型技術

---

- Google 2024 環境報告：運用 AI 促進可持續發展 17
- 六項創新的 AI 功能於 Google Pixel 18
- 微軟研究院推出 GraphRAG 強化資料發現 20
- 建立您的多語言日曆助手與 AWS 21
- AWS 發表 AI 助手以簡化醫療內容創作 23
- 釋放 Anthropic 的 Claude 3 在 Amazon Bedrock 上的提示工程潛力 24
- NVIDIA 的 GauGAN：用生成式 AI 轉變創意 26
- 頭顱顧客互動：語言處理單元 (LPU) 的崛起 27
- FRVR AI：讓遊戲開發民主化 28
- 氣象公司透過 AWS 技術強化 MLOps 29

● Eviden 強化 AWS DeepRacer 全球聯賽以活動管理器	31
● AWS 強化圖像生成，精調 Stable Diffusion XL	32
● Waabi 利用生成式 AI 進行自主貨運	34
● 透過小鼠腦研究理解人類心智	36
● 商湯科技推出 SenseNova 5.5：即時多模態人工智慧的突破	37
● AWS 推出生成式 AI 的推論優化工具包	38
● Amazon SageMaker 推出推論優化工具包	39
● Anthropic 的 Claude 3.5 Sonnet 在商業與金融領域的 S&P AI 基準中名列前茅	40
● 拉斯維加斯球體揭幕尖端顯示技術	42
● Amazon Bedrock 推出生成式 AI 開發的新功能：提示管理與提示流程	43
● AWS 在紐約峰會上揭示新的生成式 AI 創新	44
● 在 Amazon Bedrock 中微調 Anthropic 的 Claude 3 Haiku：客製化的新時代	45
● 人工智能如何改變癌症診斷與病患結果	46
● NVIDIA 推出 NIM 微服務以簡化生成式 AI 應用部署	48
● Google 在 Galaxy Unpacked 上為三星設備帶來的精彩更新	50
● 微軟的統一資料庫：大型語言模型的重大突破	51
● 微軟研究：利用生成式人工智慧對抗人類販運	52
● 人工智能如何改變遊戲設計和玩家體驗	53
● 個人電腦市場在 AI 興起中獲得牽引	55
● AWS 發佈 Amazon Bedrock 的代理程式：基礎設施即代碼的飛躍	56
● 提升 AWS SageMaker 的檢索增強生成能力	57
● BRIA AI 利用 Amazon SageMaker 進行高效模型訓練	58
● 針對地理空間分析的自訂影像與 Amazon SageMaker Studio	60

● Amazon Bedrock 強化模型客製化功能，結合 AWS Step Functions	61
● Amazon Bedrock 強化知識庫，提供先進功能以提升準確性	64
● 日本的 ABCI 3.0 超級電腦：AI 主權的飛躍	65
● 創新基於 Vitrimer 的印刷電路板：一種可持續的解決方案	66
● 軒銀收購英國AI晶片廠商Graphcore	67
● 三星在最新的摺疊手機和可穿戴設備中增強 AI 功能	68
● NVIDIA 在 SIGGRAPH 2024 展示尖端 AI 和模擬技術創新	70
● Mixbook 利用生成式 AI 強化照片書創作	71
● 黃仁勳與馬克·祖克柏將於SIGGRAPH 2024探索圖形與虛擬世界	72
● RUBICON：提升人機對話體驗	74
● AI 有潛力為英國生產力釋放 1190 億英鎊	75
● 探索最佳的 AI 驅動遊戲筆記型電腦與桌上型電腦	76
● NVIDIA NeMo 框架提升 Amazon EKS 上生成式 AI 的訓練效能	77
● 建立可擴展的機器學習環境：AWS 多帳戶策略	79
● Wondershare Filmora 引入 NVIDIA RTX Video HDR 支援	81
● AlphaFold 3：顛覆分子預測	82
● 微軟研究開發先進模型以預測電動車電池衰退	83
● 德勤意大利開發量子機器學習解決方案以檢測詐騙	84
● Amazon SageMaker 推出 Cohere Command R 調整模型	85
● 透過 Amazon Q Business 解鎖營運洞察	86
● NVIDIA的AI驅動升頻技術徹底改變影片品質	87
● 微軟研究：2024年7月15日當週的創新	88
● Meta不向歐盟推出多模態AI模型	89

● Lili 發表由 Amazon Bedrock 驅動的 AccountantAI 聊天機器人	90
● 透過 Amazon Bedrock 和 Anthropic Claude 革新文件處理	92
● Amazon Bedrock 透過元數據過濾提升數據檢索	93
● Mistral AI 與 NVIDIA 合作推出 Mistral NeMo 12B 語言模型	94
● 在 Azure 上使用 OpenTelemetry 和 Application Insights 追蹤 LangChain 代碼	96
● Mistral AI 與 NVIDIA 揭曉 12B NeMo 模型	97
● 台積電在 AI 需求激增中創下增長紀錄	99
● NVIDIA 推出自我主導的 AI 和資料科學職涯發展資源	100
● NVIDIA 的超級電腦推動量子計算研究	102
● NVIDIA 團隊在 KDD Cup 2024 中以創新 AI 解決方案大放異彩	104
● 微軟在 ICML 2024 展示機器學習創新	105
● NVIDIA AI Foundry：為企業量身打造生成式 AI	106
● NVIDIA 推出 NeMo Retriever 微服務以提升 AI 效率	107
● Meta 在 Azure AI 上推出 Llama 3.1 模型：徹底改變 AI 能力	108
● Mistral Large 2 在 Amazon Bedrock 上推出	109
● 利用 AWS SageMaker Pipelines 和 MLflow 解鎖大型語言模型的客製化	110
● Amazon Q S3 連接器：用生成式 AI 解鎖洞察力	111
● Salesforce 利用 Amazon SageMaker 加強程式碼生成	112
● 探索使用 Adobe 和 NVIDIA RTX 的 AI 協助創意	114
● Mistral Large 2：AI 模型競爭的新選手	115
● Amazon SageMaker 強化生成式 AI 模型的自動擴展	116
● 透過 Amazon Q Business 解鎖 SharePoint 洞察力	117
● Amazon Bedrock 推出對話式 AI 測試的代理評估	118

● AWS 推出自動節點問題檢測與恢復功能，專為 EKS叢集設計	119
● 微軟研究揭示 Trace：AI 優化的新時代	120
● Cohere Rerank 3 強化了 Azure AI 的搜尋功能	121
● Galileo 發佈生成式 AI 模型的幻覺指數	122
● AI 開發成本上升	124
● 利用 Amazon Bedrock 和 Salesforce 的生成式 AI 應用程式	125
● AWS 從 Amazon Forecast 過渡到 SageMaker Canvas	126
● 將 Amazon Q Business 連接到 Microsoft SharePoint Online：利用生成式 AI 提升洞察力	127
● Amazon Q Business 與 Atlassian Jira 整合以提升生產力	128
● NVIDIA 在 SIGGRAPH 2024 的 AI 助手願景	129
● WPP 與 NVIDIA Omniverse 點燃可口可樂的生成式 AI 內容革命	130
● NVIDIA 發表 fVDB 以增強數位建模	132
● NVIDIA 利用先進的生成式 AI 技術強化數位行銷	134
● NVIDIA 揭示創新數位人類技術以提升客戶互動	136
● Hugging Face 推出基於 NVIDIA NIM 的推論即服務	138
● NVIDIA 推出 NIM 微服務以支援實體環境中的生成式 AI	140
● Shutterstock 與 Getty Images 透過生成式 AI 強化創意工作流程	141
● Google 擴展 AI 驅動的野火邊界追蹤工具至歐洲和非洲	142
● 谷歌的綠燈計畫：利用 AI 解決交通排放問題	143
● 微軟研究發表 LongRoPE：語言模型能力的飛躍	144
● AI 革命：來自 NVIDIA 和 Meta 領導者的見解	145
● JPMorgan 發表 LLM 套件：一款突破性的 AI 聊天機器人用於研究分析	146
● 透過 Amazon Q Business 和 Transcribe 強化媒體搜尋	147

---

● Monks 使用 AWS 技術提升即時 AI 圖像生成	148
● Amazon Bedrock 推出網頁爬蟲功能以增強知識庫	149
● Meta 與 NVIDIA 發表 AI Studio , 打造個人化助手	152
● 蘋果選擇Google晶片進行AI開發 , 拋棄Nvidia	154
● 解鎖日本 LLM 與 AWS Trainium : 人工智慧發展的新前沿	155
● AWS 發佈 ApplyGuardrail API 以增強 Amazon Bedrock 的內容管理功能	157
● NVIDIA 展示即時生成 AI 於 3D 世界建構	158
● Oracle Cloud Infrastructure 擴展 GPU 加速實例以支持 AI 和數位雙胞胎	160
● NVIDIA 加速 AI 發展 , 推出最新 RTX 創新	162
● 微軟研究亮點 : 2024年7月29日當週	163

---

## 資訊安全

---

● Amazon Bedrock 推出增強知識庫的安全防護措施	165
● Mend.io 利用 Anthropic Claude 分析 CVE 數據	166
● AWS 強化對 PII 的數據保護 , 結合 Amazon Lex 和 CloudWatch Logs	167
● Amazon Q Business 推出可信身份傳播功能	169

---

## 應用

---

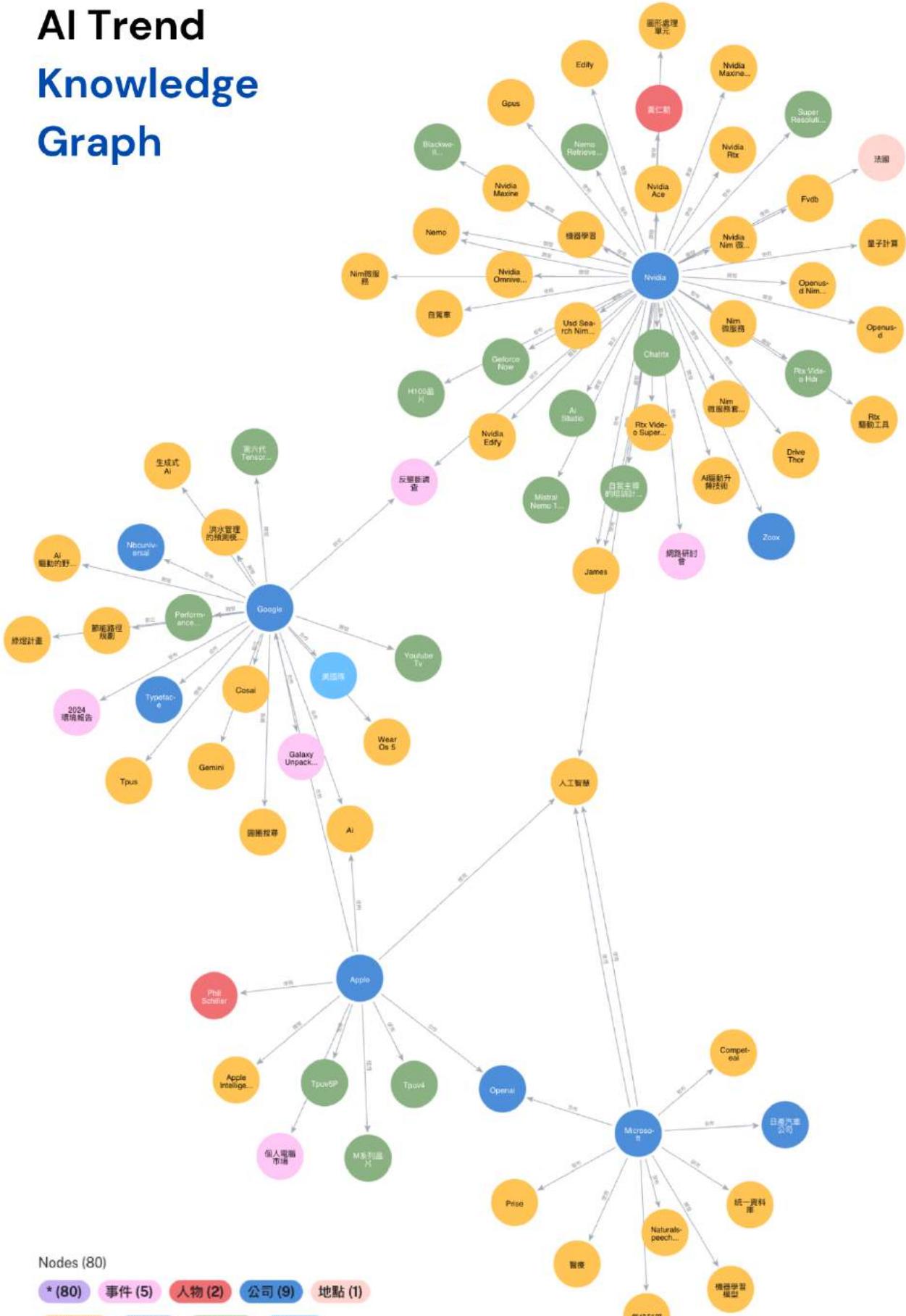
● 蘋果加入 OpenAI 董事會擔任觀察員	171
------------------------	-----

● 美國教育中的AI革命：中國應用程式的影響	172
● AWS 強化負責任的生成式 AI 相關計畫	173
● 儘管高管信心十足，AI 採用仍面臨障礙	174
● 利用 AWS 改變影片配音的遊戲規則	175
● Google Arts & Culture 推出四款創新遊戲，讓你在夏季探索藝術	176
● 教育中的生成式人工智慧：來自家長和學生的見解	178
● Google 與美國隊及 NBCUniversal 合作，為巴黎 2024 奧運會提供報導	179
● Intuit 利用 Amazon Bedrock 和 Claude 簡化 TurboTax 的報稅流程	180

# Graph

## 知識圖譜

# AI Trend Knowledge Graph



### Nodes (80)

\* (80) 事件 (5) 人物 (2) 公司 (9) 地點 (1)

技術 (46) 模型 (1) 產品 (17) 組織 (2)

## Relationships (85)

\* (85) 使用 (19) 創立 (3) 合作 (12)

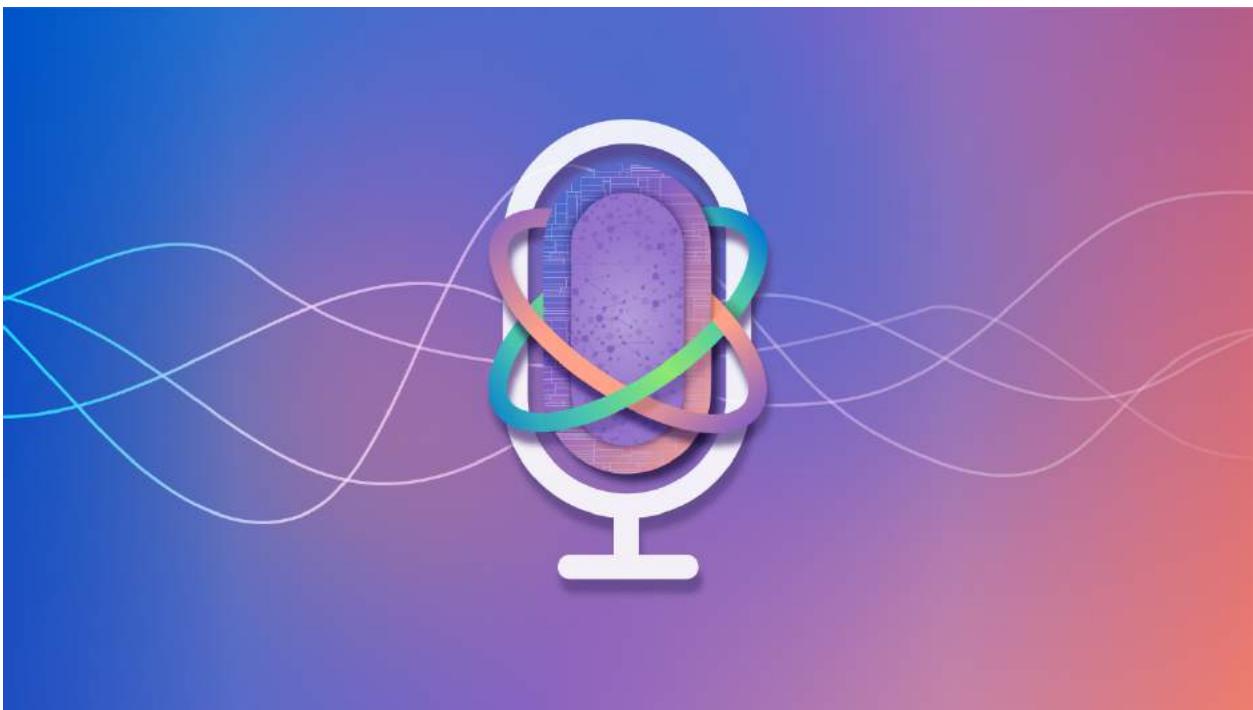
更新 (1) 發布 (13) 研究 (3) 開發 (34)

# 01 精選文章

# 微軟研究院推出 AgentInstruct：合成數據生成的重大進展

合成數據 | 自動化 | 多代理系統 | 語言模型 | 訓練 | 微軟研究院

2024-07-18



## 微軟研究院推出 AgentInstruct：合成數據生成的重大進展

2024年7月18日，微軟研究院推出了一個名為 AgentInstruct 的創新框架，由資深研究員 Arindam Mitra 主導。這個自動化的多代理系統旨在生成多樣化、高品質的合成數據，以用於語言模型的後期訓練，簡化了傳統上勞動密集的過程。與以往依賴現有數據種子的技術不同，AgentInstruct 允許模型建構者指定所需的數據類型和場景，並自動化數據創建過程。

這個框架展現出良好的結果，提升了 Orca-3 模型在各種基準測試上的表現，相較於其前身 Mistral 提升了 20%。這一進展表明，使用 AgentInstruct 訓練的模型可以達到與尖端模型如 GPT-4 相當的性能，同時顯著減少人力投入。這項研究的影響範圍廣泛，特別是對於尋求可擴展解決方案的模型建構者而言，語言模型的開發前景樂觀。隨著微軟持續精進這一方法，生成教學的未來看起來充滿希望。

[閱讀更多](#)

# Meta推出Llama 3.1：開源AI的一大步

Llama 3.1 開源AI 推理能力 數據安全 AI民主化

2024-07-24



## Meta推出Llama 3.1：開源AI的一大步

Meta最近推出了Llama 3.1，稱之為首個「前沿級」開源AI模型，標誌著向民主化AI技術的轉變。此版本包含三個模型，分別擁有405B、70B和8B參數，405B版本旨在與頂尖的封閉AI模型相抗衡，同時更具成本效益。

Llama 3.1增強了推理能力，支持128,000個tokens的上下文，允許更複雜和細緻的互動。馬克·扎克伯格強調，開源具有多項優勢：組織可以使用自己的數據自定義模型，避免供應商鎖定，通過本地部署增強數據安全性，並顯著降低運行成本。

Meta與亞馬遜和NVIDIA等科技巨頭的合作旨在促進創新，擴大AI技術的可及性。扎克伯格預測，這一舉措將使更多開發者能夠接受開源模型，最終實現全球AI利益的更公平分配。

[閱讀更多](#)

# 亞馬遜開發 AI 晶片以競爭 Nvidia

亞馬遜 | AI 晶片 | Nvidia | AWS | 雲端服務 | Graviton | Trainium | Inferentia

2024-07-29



## 亞馬遜開發 AI 晶片以競爭 Nvidia

為了提升其雲端服務，亞馬遜在德州奧斯丁的設施中開發自己的 AI 晶片。這個計畫旨在減少對 Nvidia 的依賴，因為後者的晶片價格昂貴且對亞馬遜網路服務（AWS）至關重要。透過自行製造處理器，亞馬遜希望為客戶提供更具價格競爭力的解決方案，來處理複雜的計算，同時在雲端運算和 AI 領域維持競爭優勢。

這個專案受到對 Nvidia 產品的成本效益替代品需求增加的驅動。亞馬遜已經在 Graviton 晶片上取得進展，現在正專注於其最新的自訂 AI 處理器，Trainium 和 Inferentia。這些創新有可能在價格效能方面提供 40-50% 的顯著改善，相較於目前基於 Nvidia 的解決方案，最終將大幅惠及 AWS 客戶。

隨著亞馬遜加速其晶片開發，Nvidia 也在其即將推出的 Blackwell 晶片上持續創新，這顯示出 AI 晶片市場充滿活力且競爭激烈的格局。

[閱讀更多](#)

# 02 模型技術

# Google 2024 環境報告：運用 AI 促進可持續發展

Google | AI | 環境報告 | 能源效率 | 減少排放 | 可再生能源

2024-07-02



## Google 2024 環境報告：運用 AI 促進可持續發展

在其 2024 環境報告中，Google 展示了先進科技，特別是人工智慧（AI）在推動可持續發展方面的重要性。該公司強調降低環境足跡的必要性，同時提升其運營中的能源效率。

一項重要創新是第六代 Tensor Processing Unit (TPU) Trillium，相較於前一代，能源效率提高了超過 67%。Google 還開發了可以降低 AI 模型訓練所需能源高達 100 倍的做法，這不僅針對使用情況，也解決了排放問題。

如節能路徑規劃和洪水管理的預測模型等 AI 應用，已經開始帶來實質影響。這些計畫共同幫助減少了溫室氣體排放，預計到 2030 年有潛力減緩全球排放的 10%。Google 對可再生能源和創新水資源管理的承諾，更進一步彰顯了其對可持續未來的雄心。

[閱讀更多](#)

# 六項創新的 AI 功能於 Google Pixel

AI | Google Pixel | 功能 | 智慧手機 | 用戶體驗 | Gemini | Audio Magic Eraser | Circle to Search  
Proofread | Document Scanner | Cough and Snore Detection

2024-07-02



## 六項創新的 AI 功能於 Google Pixel

Google 的 Pixel 智慧手機整合了六項突破性的 AI 功能，以提升用戶體驗。

1. Gemini：這個個人 AI 助手能夠總結網頁內容，讓你省下瀏覽的時間。
2. Audio Magic Eraser：一個工具，可以調整你影片中的背景聲音，確保你的聲音在吵雜的環境中也能清晰可聞。
3. Circle to Search：只需在螢幕上圈選或標記任何內容，即可獲取更多資訊，而無需切換應用程式。
4. Proofread with Gboard：一個由 AI 駅動的功能，能夠識別並一鍵修正你文本中的拼寫錯誤和文法錯誤。
5. Document Scanner：使用相機輕鬆將實體文件轉換為 PDF，便於保存數位記錄。
6. Cough and Snore Detection：透過追蹤你在睡眠中發出的聲音，來監測你的睡眠模式，而不需儲存原始音訊。

這些功能展示了 AI 如何簡化日常任務，並改善我們與科技的互動。

---

閱讀更多

# 微軟研究院推出 GraphRAG 強化資料發現

**GraphRAG** | 微軟研究院 | 大型語言模型 | 知識圖譜 | 資料發現

2024-07-02



## 微軟研究院推出 GraphRAG 強化資料發現

微軟研究院最近推出了 GraphRAG，這是一個創新的工具，現已在 GitHub 上提供，旨在通過基於圖形的檢索增強生成（RAG）方法來提升複雜資料的發現能力。這個創新的框架使得用戶能夠比傳統方法更有效地對私有或不熟悉的數據集進行問答。

GraphRAG 利用大型語言模型（LLM）從文本文件中創建豐富的知識圖譜，識別數據中緊密連接的社群。這一能力使得該工具能夠以層次方式總結關鍵主題和話題，提供全面的概覽，而不需要準確的問題。

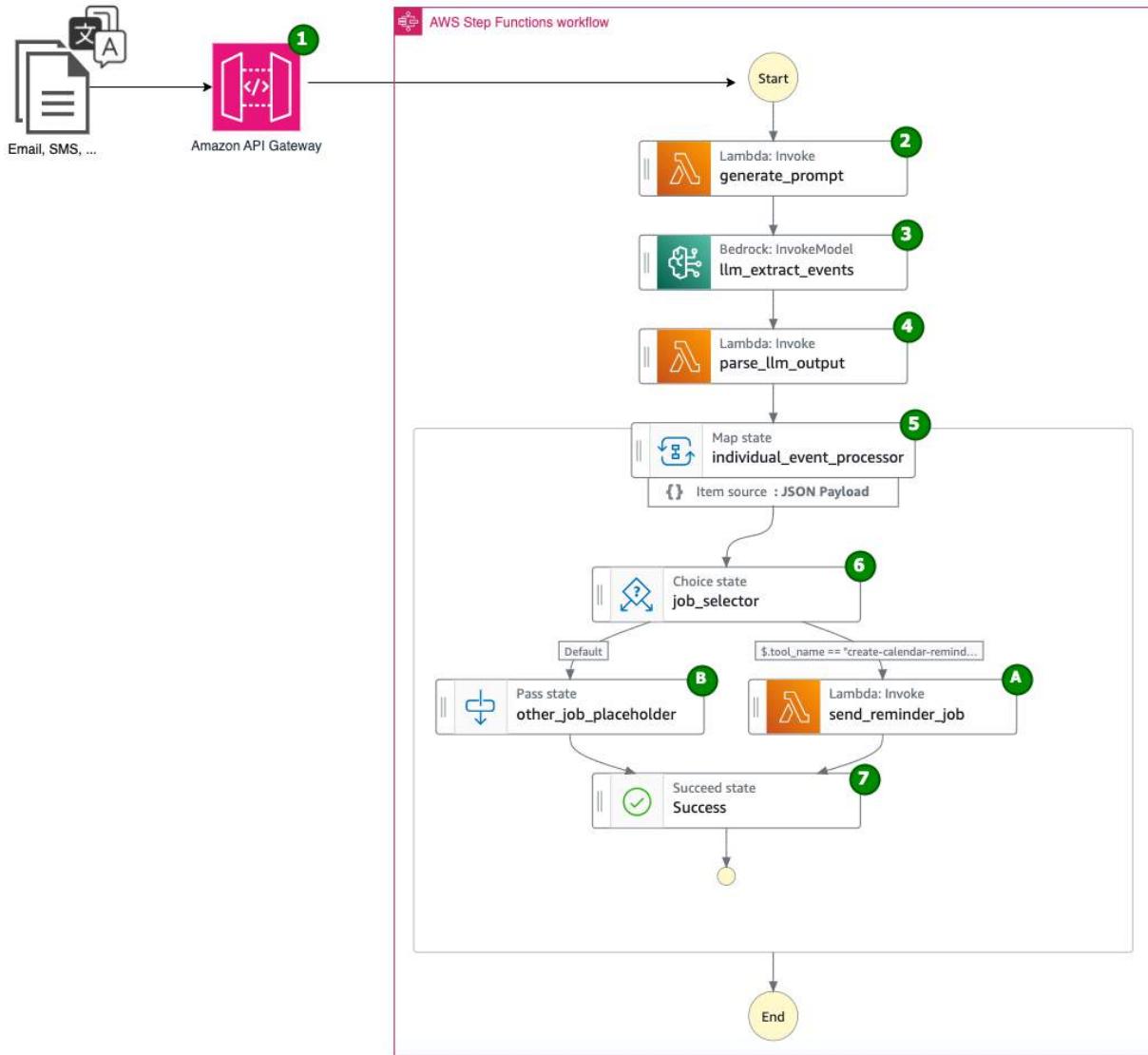
通過利用社群摘要，GraphRAG 在解決全球性查詢方面表現出色，提供詳細且多樣的回應，超越了簡單的 RAG 技術。因此，它顯著提高了信息檢索的全面性和效率，成為需要從龐大數據集合中獲取洞見的人的寶貴資源。

[閱讀更多](#)

# 建立您的多語言日曆助手與 AWS

多語言日曆助手 | AWS | Amazon Bedrock | AWS Step Functions | 排程 | 自動化 | 生成式 AI

2024-07-03



## 建立您的多語言日曆助手與 AWS

亞馬遜推出了一種方式，透過使用 AWS 工具如 Amazon Bedrock 和 AWS Step Functions，來簡化您的排程，讓您擁有一個全自動的多語言日曆助手。這個創新的助手旨在幫助外國人和外籍人士管理繁忙的日程，克服語言障礙，這種障礙常常使得設定活動提醒變得更為複雜。

這個助手能夠處理來自不同語言的電子郵件，自動翻譯並設置日曆提醒。Amazon Bedrock 提供了一系列基礎模型 (Foundation Models, FMs) 的存取，這些模型可以輕鬆自訂，無需管理基礎設施的麻煩。與此同時，AWS Step Functions 負責協調工作流程，允許多個自動化任務的執行，例如發送電子郵件提醒。

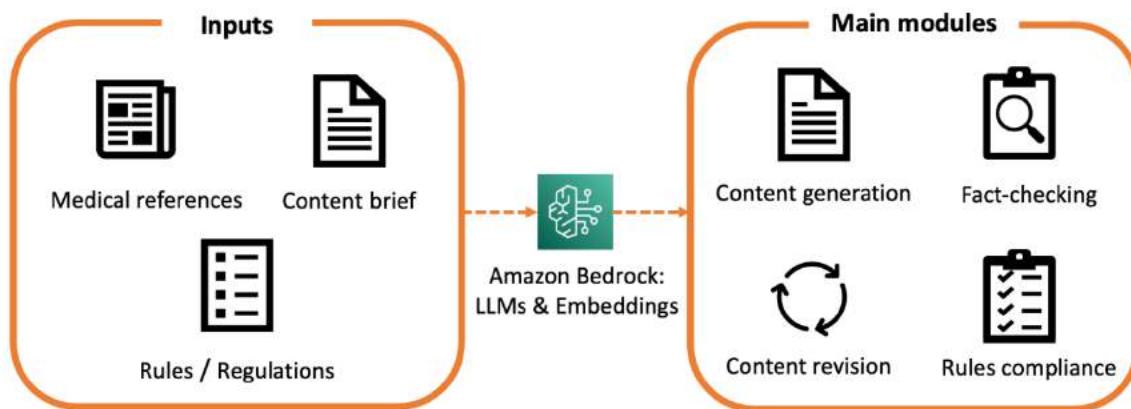
這個解決方案不僅簡化了排程過程，還透過利用生成式 AI 的能力來提升生產力，使其成為任何需要同時處理多個時區和語言的人士的寶貴工具。對於有興趣建立自己助手的人，源代碼和部署說明也已經提供。

[閱讀更多](#)

# AWS 發表 AI 助手以簡化醫療內容創作

AWS | AI 助手 | 醫療內容創作 | 生成式 AI | 大型語言模型 | Amazon Bedrock | 內容生成 | 法規標準  
事實檢查

2024-07-03



## AWS 發表 AI 助手以簡化醫療內容創作

在醫療行業的一項重要進展中，AWS 推出了旨在徹底改變醫療內容創作的 AI 助手。這款工具利用生成式 AI 和大型語言模型（LLMs），大幅減少了製作疾病宣導行銷內容所需的時間——從幾週縮短至僅幾小時。

該 AI 助手運行於 Amazon Bedrock，使品牌經理和醫療專家能夠有效地生成和修訂量身訂做的醫療內容，同時確保準確性和法規標準的遵守。

主要功能包括一個自動修訂功能，促進互動反饋，提升生成內容的質量。系統還加入了事實檢查的護欄，確保遵循基本規則和法規。

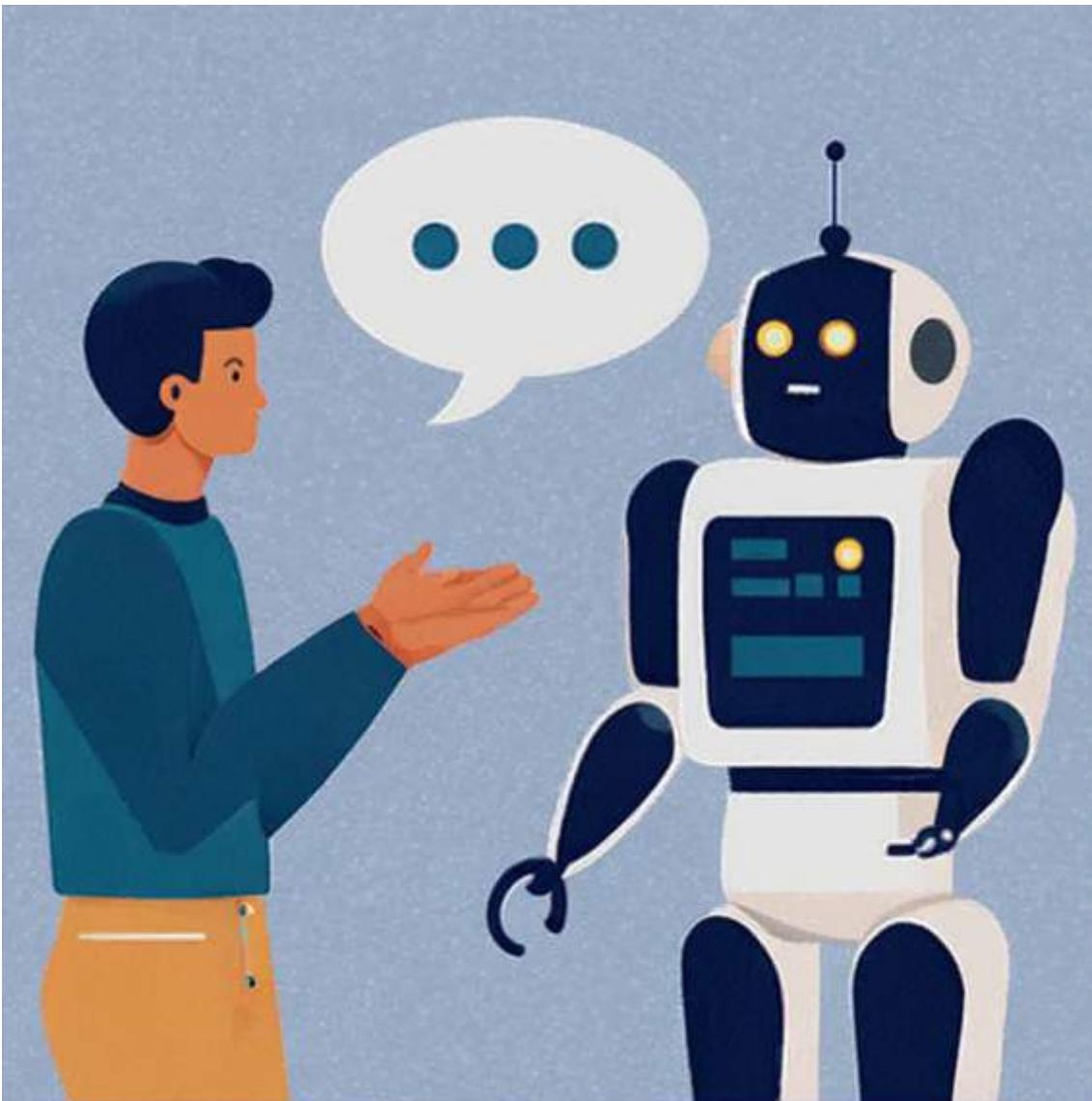
通過利用生成式 AI 的力量，AWS 正在為製作精確且引人入勝的醫療內容設立新的標準，最終改善醫療和生命科學領域的溝通策略。

[閱讀更多](#)

# 釋放 Anthropic 的 Claude 3 在 Amazon Bedrock 上的提示工程潛力

提示工程 | Claude 3 | 大型語言模型 | AI 生成內容

2024-07-03



## 釋放 Anthropic 的 Claude 3 在 Amazon Bedrock 上的提示工程潛力

在生成 AI 不斷演變的環境中，有效的提示工程對於優化大型語言模型 (LLMs) 的輸出至關重要。最近，Anthropic 的 Claude 3 通過 Amazon Bedrock 提供了創新的提示設計技術，以增強用戶互動。

提示工程涉及設計指令，指導 AI 模型產生相關且準確的回應。構造不良的提示可能導致不準確和不相關的輸出，而結構良好的提示則可以顯著提高 AI 生成內容的質量。

Claude 3 的能力包括理解複雜查詢以及處理文本和視覺輸入。憑藉其先進的架構，該模型可以減少 AI 回應中常見的錯誤和幻覺。值得注意的是，Claude 3 系列包括 Haiku、Sonnet 和 Opus 模型，每個模型都針對不同的性能需求進行調整，從快速回應到複雜推理任務。

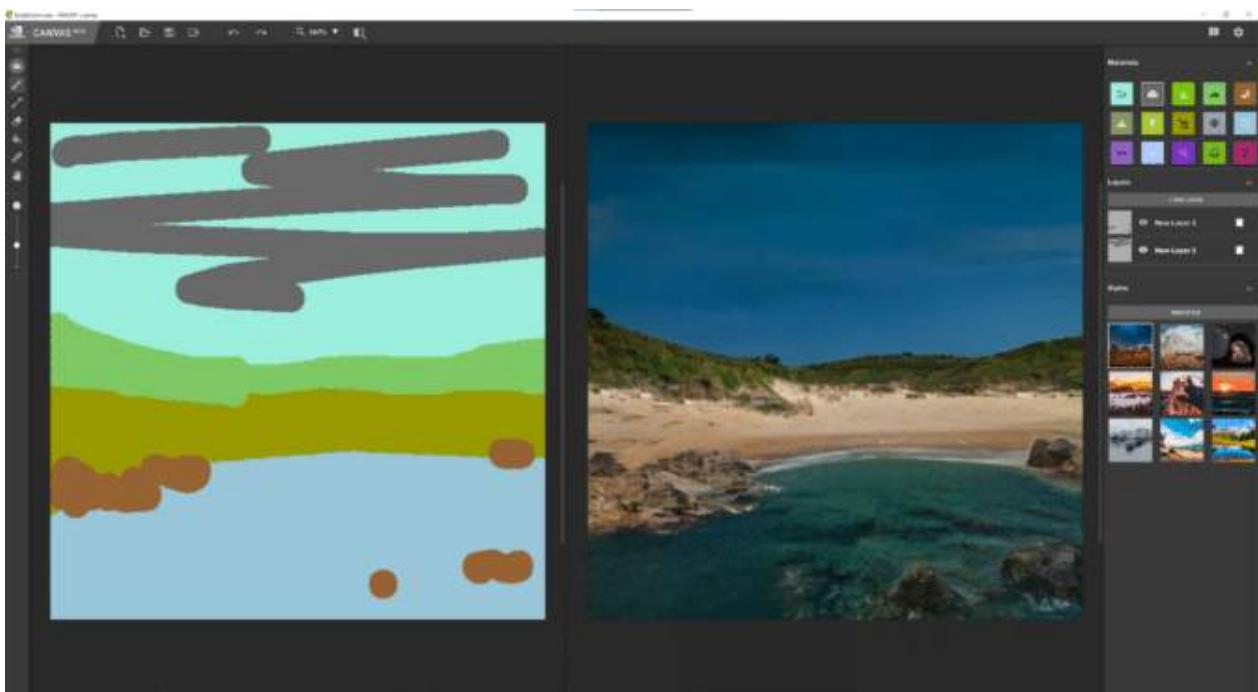
通過利用這些提示工程策略，用戶可以增強其生成 AI 應用程序，為更連貫和相關的互動鋪平道路。

[閱讀更多](#)

# NVIDIA 的 GauGAN：用生成式 AI 轉變創意

NVIDIA | GauGAN | 生成式 AI | 深度學習 | GAN | 藝術 | 創意 | 視覺效果

2024-07-03



## NVIDIA 的 GauGAN：用生成式 AI 轉變創意

NVIDIA 的 GauGAN 站在生成式 AI 革命的最前沿，促進了一波新的應用，增強創意工作流程。這個創新的模型通過生成對抗網絡（GAN）使用深度學習，GAN 由兩個神經網絡組成——生成器和判別器——它們競爭以創建真實的圖像。GauGAN 以畫家保羅·高更（Paul Gauguin）的名字命名，允許用戶將簡單的素描轉變為驚人的寫實藝術作品。

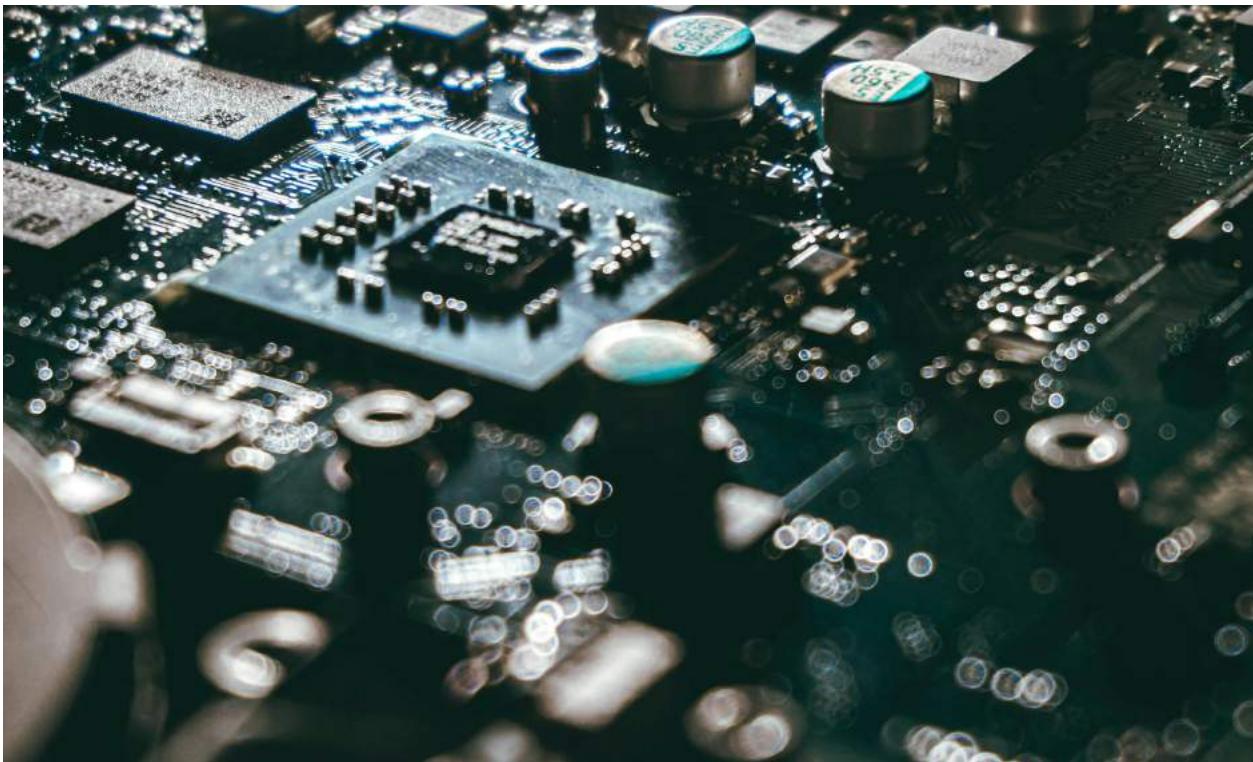
通過 NVIDIA Canvas 應用程式，用戶可以輕鬆地素描風景，AI 會立即將其轉換為精緻的視覺效果。用戶可以選擇各種材料，如草或雲，實時操控場景——從反射樹木的寧靜池塘到生機勃勃的熱帶天堂。GauGAN 的 Panorama 模式支持 360 度影像，開啟了無限的創意可能性。這項技術不僅是藝術家的工具，它正在重塑我們在遊戲、視訊會議和互動體驗中的參與方式，標誌著 AI 領域的一個重要里程碑。

[閱讀更多](#)

# 顛覆顧客互動：語言處理單元 (LPU) 的崛起

語言處理單元 | 語音 AI | 客服中心 | 顧客服務 | 預測能力

2024-07-04



## 顛覆顧客互動：語言處理單元 (LPU) 的崛起

語言處理單元 (LPUs) 是為了增強語音 AI 能力而設計，特別是在客服中心。由 Groq 開發的這些專用處理器，其性能超越傳統的圖形處理單元 (GPUs)，速度快上 10 倍，延遲降低 90%。這項技術解決了顧客服務互動中，清晰度和速度的關鍵挑戰。

想像一下，繁忙的咖啡店中，咖啡師必須在嘈雜環境中迅速理解複雜的訂單。LPU 的功能相似，能讓 AI 實時理解和處理複雜的語言任務。這樣的結果使得顧客與 AI 之間的對話更加自然，猶如人類之間的互動。

LPU 不僅改善了語音轉文字和文字轉語音的轉換，還增強了 AI 的預測能力，使其能更好地預測用戶的需求。LPU 的整合承諾在各行各業產生變革性的影響，包括醫療和金融等領域，這些地方快速而準確的溝通對提升顧客服務體驗至關重要。

[閱讀更多](#)

# FRVR AI：讓遊戲開發民主化

遊戲開發 | FRVR AI | 原型 | 用戶友好介面 | 創作發佈

2024-07-04



## FRVR AI：讓遊戲開發民主化

FRVR 推出了 FRVR AI，這是一個突破性的工具，使任何人都能夠創建自己的視頻遊戲，無論他們是否具備編程或藝術經驗。這個創新的平台簡化了遊戲開發過程，使用者只需輸入對遊戲概念的簡短描述。然後，AI 將製作出一個可玩的原型，包含基本結構和資源。

FRVR AI 提供了一個用戶友好的介面，分為五個部分以便於導航：輸入、即時預覽、歷史紀錄、代碼審查和資源創建。它還包括自助功能，指引使用者在創意旅程中的探索。最近的更新通過引入音效來提升這種體驗，改善了遊戲的沉浸感。

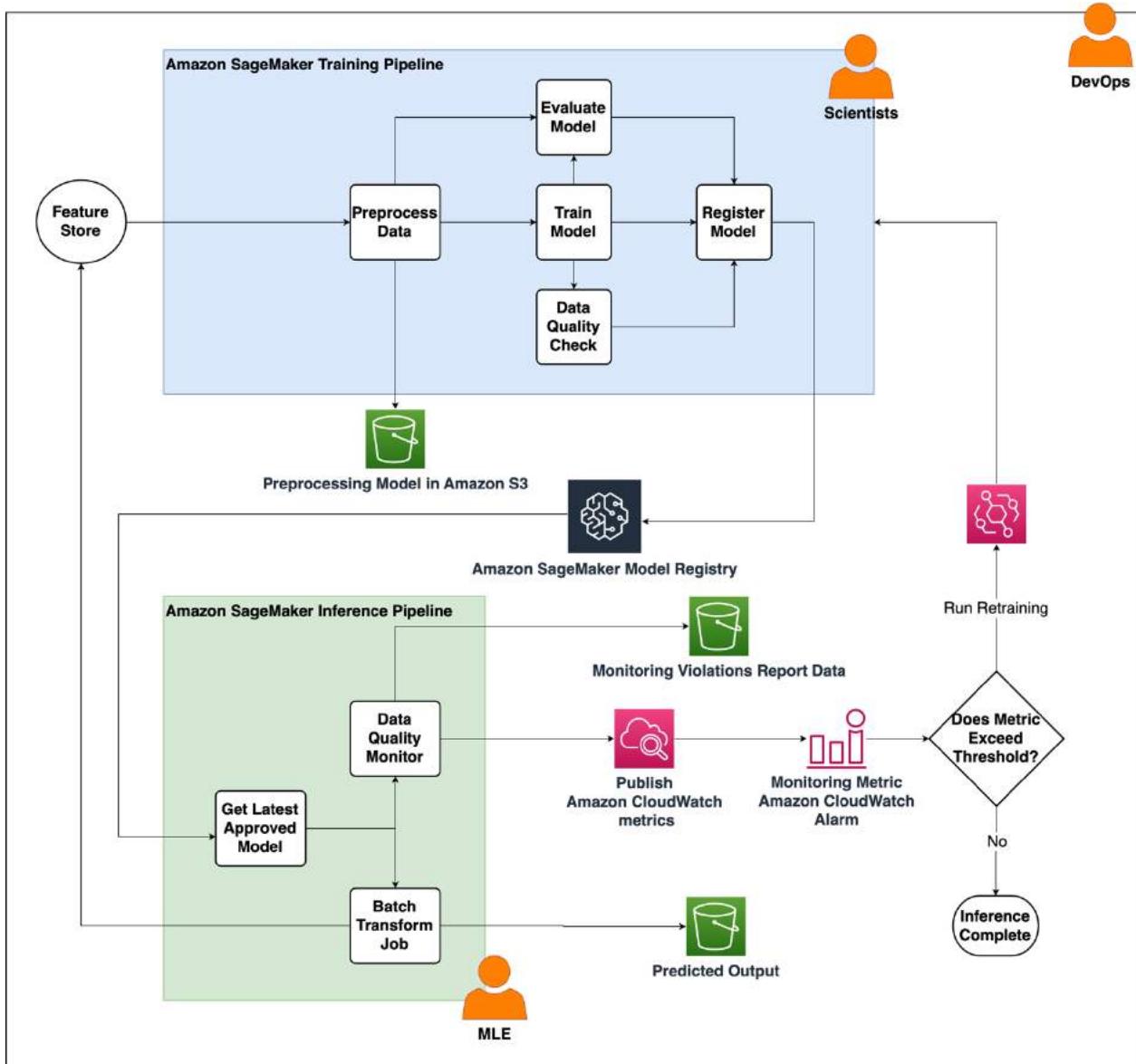
使用者可以輕鬆地將他們的創作發佈到 FRVR 平台，並使其對全球觀眾可及。目前處於測試階段的 FRVR AI 吸引了一個充滿活力的創作者社群，促進了一個任何人都可以開發和分享其遊戲願景的空間。

[閱讀更多](#)

# 氣象公司透過 AWS 技術強化 MLOps

氣象公司 | AWS | MLOps | 機器學習 | 數據科學 | 自動化 | 數據存儲 | 天氣預測

2024-07-08



## 氣象公司透過 AWS 技術強化 MLOps

最近，氣象公司 (TWCo) 透過使用 Amazon SageMaker、AWS CloudFormation 和 Amazon CloudWatch 升級了其機器學習運營 (MLOps) 平台。這項創新的整合使 TWCo 的數據科學家和機器學習工程師能夠自動化流程、追蹤實驗，並有效地簡化訓練和部署流程。

透過利用這些 AWS 服務，TWCo 實現了基礎設施管理時間驚人的 90% 減少，以及模型部署時間的 20% 減少。增強的 MLOps 框架促進了團隊之間的合作，提高了工作流程的透明度，並允許創建用戶友好且注重隱私的機器學習模型，分析天氣對健康的影響。

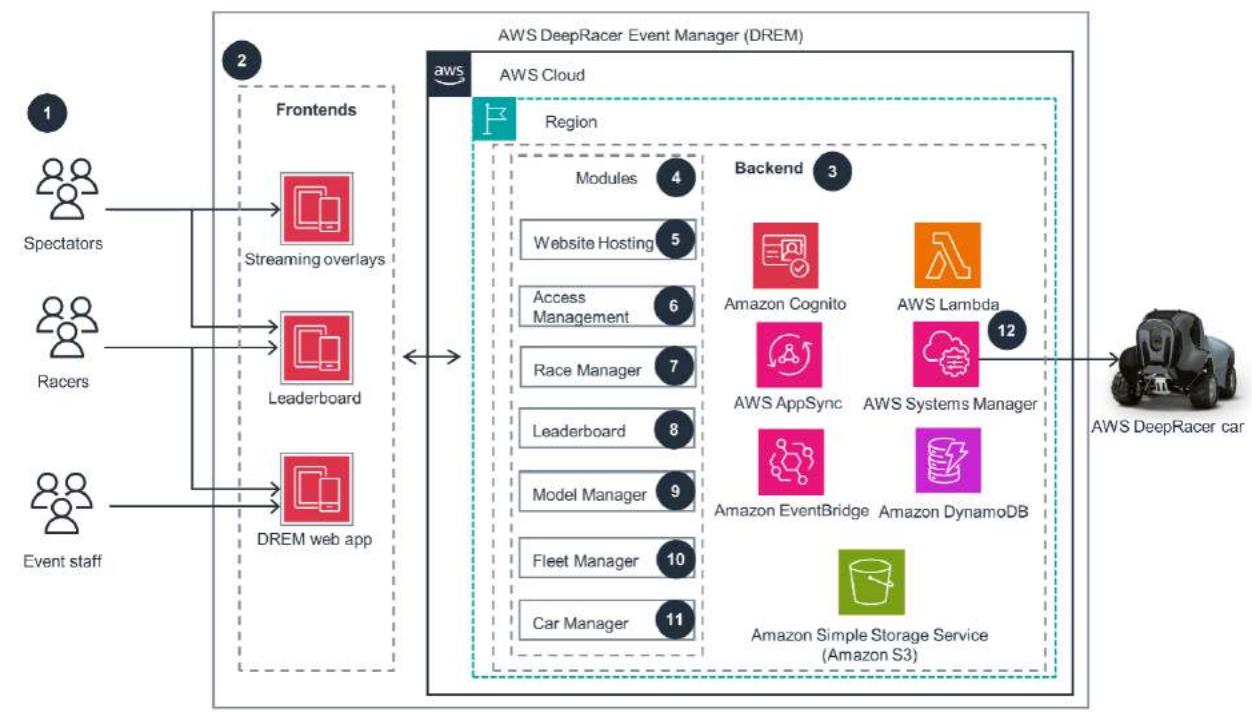
這個綜合架構整合了多個 AWS 工具，包括 AWS CodePipeline 用於持續交付和 Amazon S3 用於數據存儲，這些工具共同支持快速的機器學習開發週期和改善的用戶體驗。這項進展強調了 TWCo 對利用尖端技術以改善天氣預測和洞察的承諾。

[閱讀更多](#)

# Eviden 強化 AWS DeepRacer 全球聯賽以活動管理器

**AWS DeepRacer 活動管理器 賽事 模型上傳 競賽 雲端技術 互動體驗**

2024-07-08



## Eviden 強化 AWS DeepRacer 全球聯賽以活動管理器

Eviden 整合了 AWS DeepRacer 活動管理器 (DREM)，以簡化和提升全球 AWS DeepRacer 系列賽事。這個創新的工具簡化了活動的舉辦，讓像是比得哥什、巴黎和普納等地的管理變得更加順暢。

DREM 允許賽車手提前上傳他們的模型，並利用 AWS Systems Manager 進行高效的配置。它透過集成的 Raspberry Pi 裝置捕捉關鍵指標，如圈速，並在現場參加者與遠端觀眾的領先榜上顯示即時更新，提升了賽事體驗。

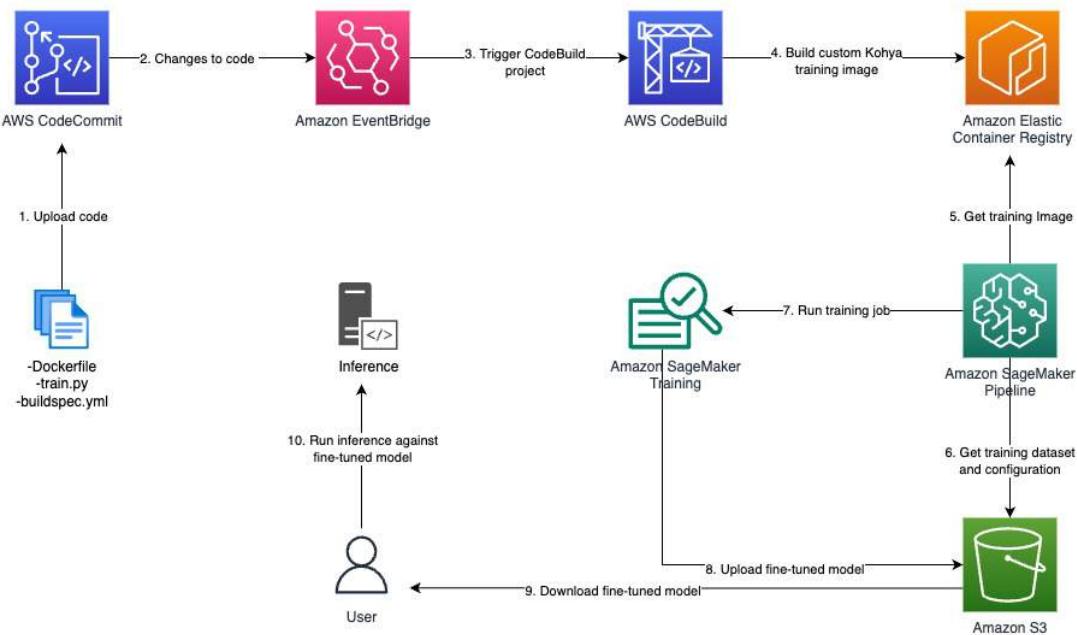
DREM 平台建置在 AWS Cloud 技術上，確保了可靠性和安全性。其角色基礎的存取控制和低運營成本，使其成為管理活動的實用解決方案。這一進展不僅提高了參與者的互動，也展示了在科技驅動競賽中更廣泛應用的潛力。

[閱讀更多](#)

# AWS 強化圖像生成，精調 Stable Diffusion XL

**AWS Stable Diffusion XL** 圖像生成 深度學習 自訂圖像 低秩適應 遊戲角色設計 行銷素材  
 電影故事板 媒體 娛樂 零售

2024-07-08



## AWS 強化圖像生成，精調 Stable Diffusion XL

Amazon Web Services (AWS) 推出了透過 Amazon SageMaker 精調 Stable Diffusion XL 生成自訂圖像的創新方式。這個深度學習模型讓使用者能夠快速創造高品質的圖像，非常適合用於遊戲角色設計、行銷素材以及電影故事板等多種應用。

透過利用自訂資料集，使用者可以個性化圖像生成過程，將獨特的主題融入創作中。精調的過程採用了低秩適應 ( Low-Rank Adaptation, LoRA ) 方法，這種方式在不需要大量重新訓練的情況下，為基礎模型添加少量的參數，從而優化資源的使用。

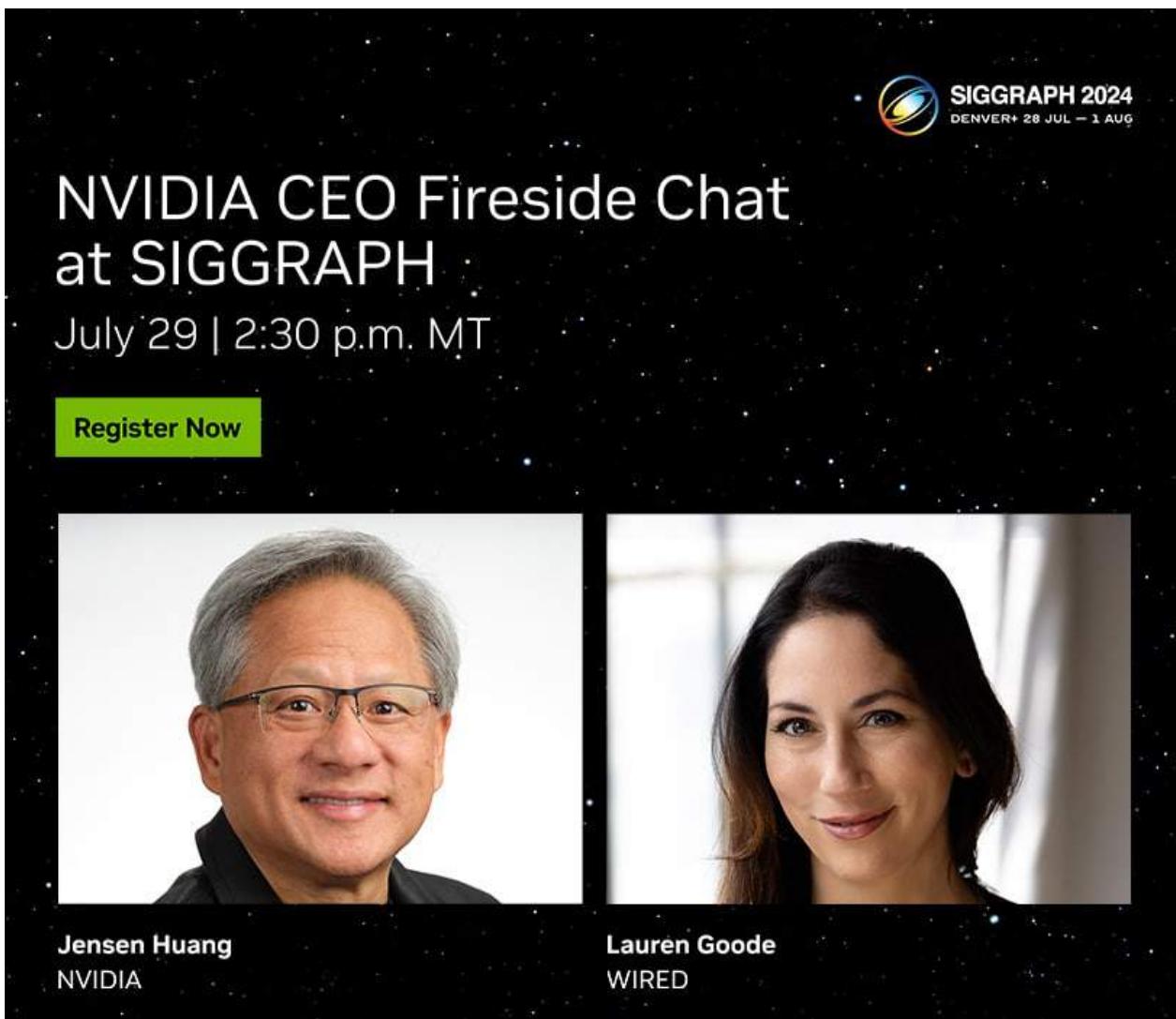
這個解決方案提供了一個簡化的工作流程，自動化生成獨特圖像，使媒體、娛樂和零售等行業的公司能夠高效地創造量身定做的視覺內容。所有必要的程式碼和配置均已提供，使用者可以輕鬆開始他們的圖像生成項目。



# Waabi 利用生成式 AI 進行自主貨運

Waabi | 生成式 AI | 自主貨運 | 自動駕駛 | NVIDIA | 長途貨運

2024-07-08



## Waabi 利用生成式 AI 進行自主貨運

Waabi 是一家位於多倫多的創新新創公司，正在利用生成式 AI 改變長途貨運產業。在最近的 GTC 活動中，Waabi 公布了其計劃推出 Waabi Driver，這是一個由 NVIDIA 的 DRIVE Thor 集中車載計算機驅動的全自動貨運解決方案。透過整合兩個生成式 AI 系統，Waabi World 和 Waabi Driver，這家新創公司正建立一個訓練框架，使自駕車輛能夠進行類似人類的推理，大幅減少了廣泛路測的需求。

這種創新方法不僅提高了安全性，還確保了商業運營的可擴展性。Waabi 目標於明年推出完全無人駕駛的操作，並與 NVIDIA 密切合作，將這項先進技術帶入實際應用。這一發展有望加速自動駕駛車輛的部署，為交通運輸未來的更安全和更高效鋪平道路。

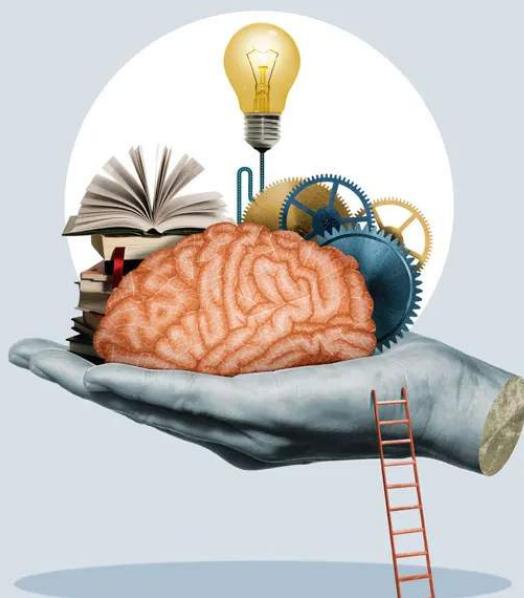
---

閱讀更多

# 透過小鼠腦研究理解人類心智

小鼠腦 | 連接組 | 認知功能 | 記憶形成 | AI | 影像技術

2024-07-08



## 透過小鼠腦研究理解人類心智

近期神經科學的進展正在縮小小鼠與人類大腦研究之間的距離。谷歌 Connectomics 團隊的研究人員正在繪製小鼠海馬體的地圖，這個區域對於記憶與導航至關重要，旨在深入了解人類的認知功能。這個計劃非常雄心勃勃，目標是生成有史以來最大的生物數據集，容量可能達到 20,000-30,000 TB。

繪製小鼠腦的地圖可不是一件小事；這面臨著重大的技術挑戰。Connectomics 團隊利用創新的影像技術與 AI 演算法，來有效處理大量數據。他們的努力建立在之前的成功基礎上，包括繪製小型動物大腦如果蠅與斑馬魚的地圖。

從小鼠連接組得到的見解可能揭開記憶形成、睡眠以及腦部疾病機制的關鍵謎題，最終增進我們對人類心智的理解。

[閱讀更多](#)

# 商湯科技推出 SenseNova 5.5：即時多模態人工智能的突破

SenseNova 5.5 | 多模態人工智能 | 商湯科技 | 語音識別 | AI 虛擬形象生成器 | 金融 | 農業

2024-07-09



## 商湯科技推出 SenseNova 5.5：即時多模態人工智能的突破

商湯科技推出了 SenseNova 5.5，並搭載了 SenseNova 5o——中國首個即時多模態人工智能模型。這個創新的系統能夠模擬人類對話的互動，提升即時對話和語音識別等應用，類似於 GPT-4o 的能力。

SenseNova 5.5 模型在性能上比其前身提升了 30%，提供更好的數學推理、英語水平和指令執行能力。值得注意的是，它將設備的運營成本降低至每年人民幣 9.90 元（約 1.36 美元），促進物聯網環境中的更廣泛使用。

此外，商湯科技還開發了 SenseChat Lite-5.5，具有更快的推理時間，以及 Vimi，一個 AI 虛擬形象生成器，可以從照片創建可控的影片片段。隨著在金融和農業等行業的進步，商湯科技正在鞏固其在不斷演變的人工智慧技術領域中的地位。

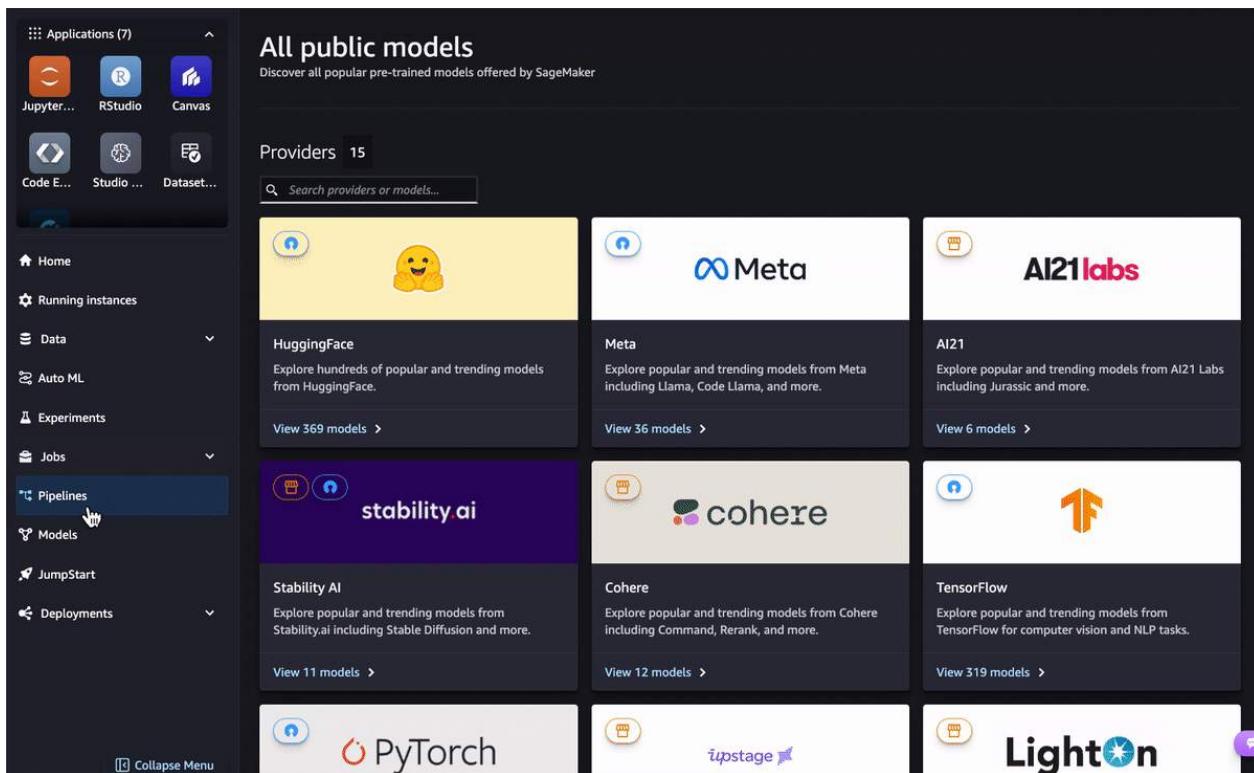
[閱讀更多](#)

# AWS 推出生成式 AI 的推論優化工具包

**AWS 生成式 AI 推論優化工具包 SageMaker 吞吐量 成本降低 Neuron Compiler**

**Activation-aware Weight Quantization 推測解碼 模型優化**

2024-07-09



## AWS 推出生成式 AI 的推論優化工具包

Amazon Web Services (AWS) 宣布推出一款創新的推論優化工具包，旨在提升其 SageMaker 平台上生成式 AI 的性能。這個全新的工具包使企業在執行 AI 推論任務時，能實現近乎雙倍的吞吐量，同時將成本降低約 50%。

該工具包利用了先進的技術，如編譯、量化和推測解碼。例如，Neuron Compiler 專為特定硬體優化模型，從而提高處理速度並減少資源使用。此外，Activation-aware Weight Quantization (AWQ) 在不影響質量的情況下，最小化模型大小，而推測解碼則通過平行預測結果來加快輸出生成。

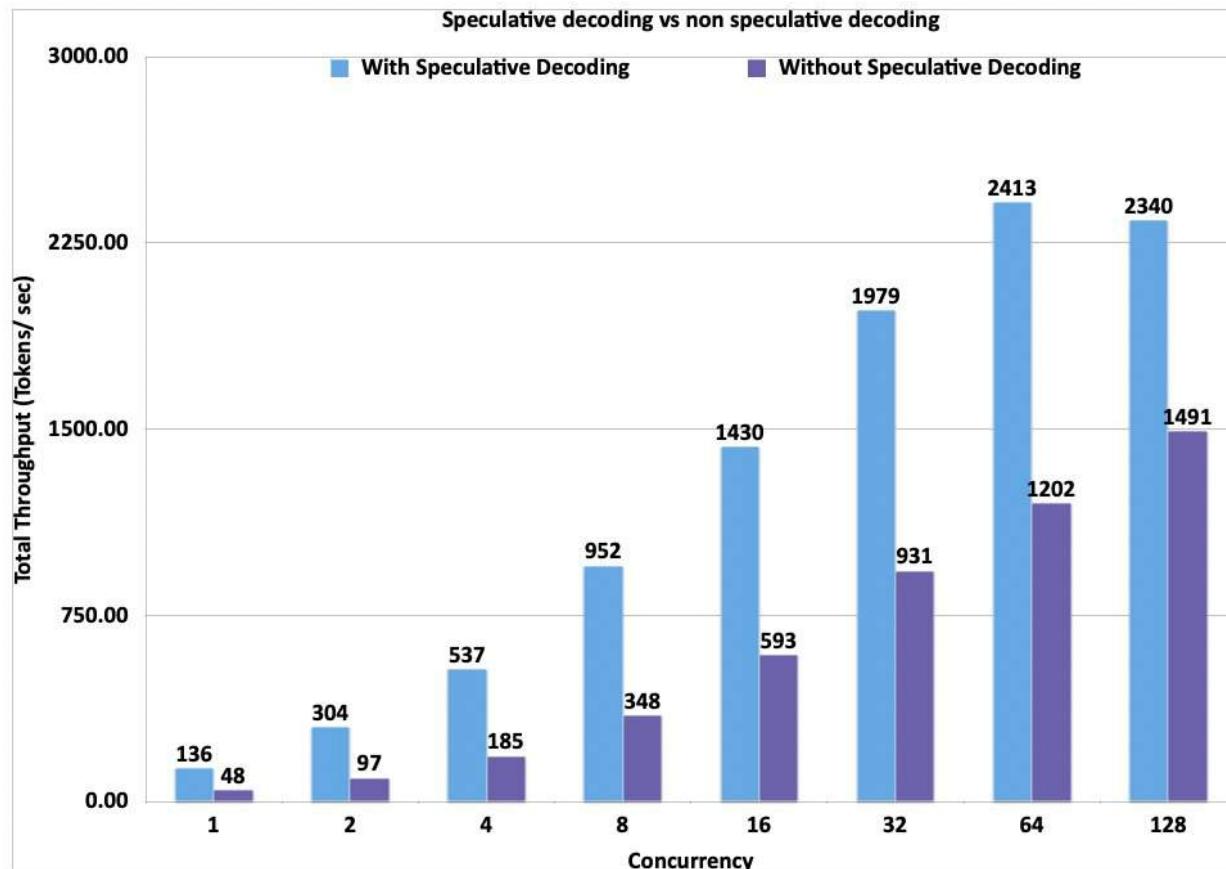
這一系列的功能使得組織能有效地優化其 AI 模型，將設置時間從數月縮短至幾小時。通過簡化模型優化過程，AWS 使企業更容易採用生成式 AI，開啟創新應用的新機會。

[閱讀更多](#)

# Amazon SageMaker 推出推論優化工具包

**Amazon SageMaker** | 推論優化工具包 | 生成式 AI | 模型性能 | 吞吐量 | 成本降低

2024-07-09



## Amazon SageMaker 推出推論優化工具包

Amazon SageMaker 最近推出了一個突破性的推論優化工具包，旨在提升生成式 AI 模型的性能。這個工具包使用戶能夠在幾小時內優化他們的模型——這速度比通常需要的幾個月快得多。透過利用如推測解碼、量化和模型編譯等技術，用戶可以實現高達兩倍的吞吐量，同時將成本降低約 50%。

例如，使用 Llama 3-70B 模型，在特定實例上性能可以達到每秒 2,400 個標記。該工具包簡化了優化流程，使用戶能夠輕鬆應用多種技術，並更專注於他們的業務目標，而不是模型調優的複雜性。這項創新對於需要高效率和低延遲的應用特別有利，使生成式 AI 對各行各業變得更加可及且具成本效益。

[閱讀更多](#)

# Anthropic 的 Claude 3.5 Sonnet 在商業與金融領域的 S&P AI 基準中名列前茅

**Claude 3.5 Sonnet** 大型語言模型 **AI 基準** 金融領域 數量提取 定量推理 **Amazon Bedrock**

2024-07-09

## Leaderboard

Rank	Model Name	Organization	Model Size (in billions)	Domain Knowledge (%)	Quantity Extraction (%)	Program Synthesis (%)	Overall (%)
1	<b>Claude 3.5 Sonnet</b>	 Anthropic 	 unknown	85.5	94.17	84.55	88.07
2	<b>GPT-4o</b>	 OpenAI 	 unknown	83.97	93.72	86.18	87.96
3	<b>GPT-4 Turbo</b>	 OpenAI 	 unknown	81.68	95.52	85.77	87.66
4	<b>GPT-4</b>	 OpenAI 	 unknown	80.15	96.41	79.67	85.41
5	<b>Claude 3 Opus</b>	 Anthropic 	 unknown	74.05	92.83	82.52	83.13
6	<b>Claude 3 Sonnet</b>	 Anthropic 	 unknown	71.76	95.52	71.14	79.47
7	<b>Llama 3 70B</b>	 Meta 	 70	77.1	93.27	67.89	79.42

Anthropic 的 Claude 3.5 Sonnet 在商業與金融領域的 S&P AI 基準中名列前茅

在 Kensho 最近的評估中，Anthropic 的 Claude 3.5 Sonnet 被評選為商業與金融領域的領先大型語言模型 (LLM)，根據 S&P AI 基準。這項基準代表了對 LLM 在處理複雜金融任務能力的關鍵評估。

該評估包括 600 個問題，針對三個關鍵領域：領域知識、數量提取和定量推理。Claude 3.5 Sonnet 在這些領域展現了卓越的表現，顯示出其理解金融術語、從報告中提取相關數據以及根據現實情境進行複雜計算的能力。

透過 Amazon Bedrock，Kensho 能夠有效地對這模型進行基準測試，強調其在金融領域應用的潛力。隨著行業尋求量身訂做的 AI 解決方案，Claude 3.5 Sonnet 成為希望在金融領域利用生成式 AI 的企業的一個強大選擇。

---

閱讀更多

# 拉斯維加斯球體揭幕尖端顯示技術

拉斯維加斯球體 | LED顯示技術 | NVIDIA | GPU | 影像捕捉 | 沉浸式體驗

2024-07-09



## 拉斯維加斯球體揭幕尖端顯示技術

拉斯維加斯球體以其近75萬平方英尺的非凡LED顯示屏正在革新娛樂產業。該場地由約150個NVIDIA RTX A6000 GPU提供動力，展現出令人驚嘆的從地板到天花板的視覺效果，並擁有世界上最大的LED螢幕，稱為Exosphere，內含120萬個可編程的LED圓盤。

為了確保這些龐大顯示屏之間的無縫同步，NVIDIA採用了BlueField DPUs和ConnectX-6 Dx NICs，並搭配DOCA Firefly服務與Rivermax軟體，以簡化媒體串流並消除延遲。這項技術使球體能夠提供沉浸式體驗，例如U2的破紀錄音樂會以及達倫·阿倫諾夫斯基的《來自地球的明信片》，展現高解析度的視覺效果和氛圍效果。

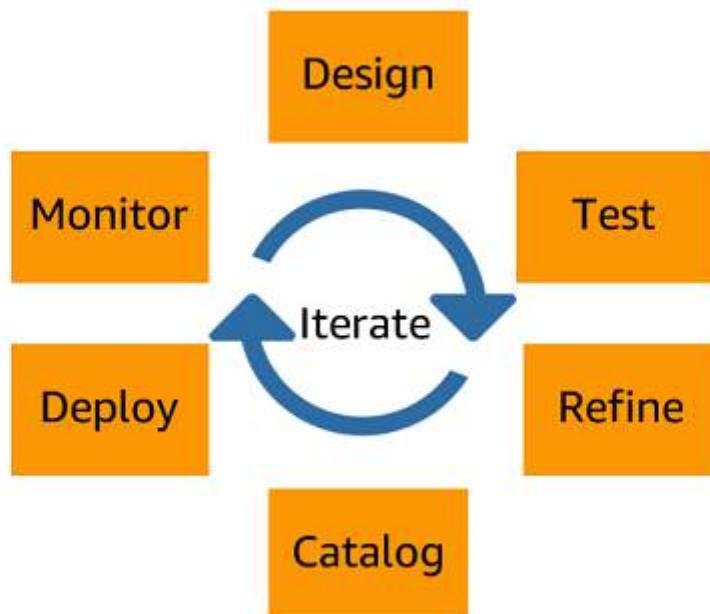
為了創造原創內容，球體工作室位於伯班克，創新技術如Big Sky攝影系統捕捉18K影像，展示了NVIDIA GPU和軟體在重新定義現場娛樂方面的尖端能力。

[閱讀更多](#)

# Amazon Bedrock 推出生成式 AI 開發的新功能：提示管理與提示流程

Amazon Bedrock | 生成式 AI | 提示管理 | 提示流程 | 工作流程 | AWS 服務

2024-07-10



Amazon Bedrock 推出生成式 AI 開發的新功能：提示管理與提示流程

Amazon Bedrock 已經推出兩項創新功能：提示管理和提示流程，現在已在公開預覽中提供。這些工具旨在簡化並增強生成式 AI 應用的開發，使開發人員能更輕鬆地創建、測試和部署解決方案。

提示管理簡化了整個提示生命週期——從設計到部署——讓使用者能夠使用直觀的工具來創建、評估和管理提示。這個功能使得快速創建提示變體變得可能，並支持團隊間的協作開發。

另一方面，提示流程提供了一個視覺介面，用於構建複雜的工作流程，連接各種提示和 AWS 服務。這減少了大量編碼的需求，讓使用者能夠快速迭代並部署他們的 AI 應用。

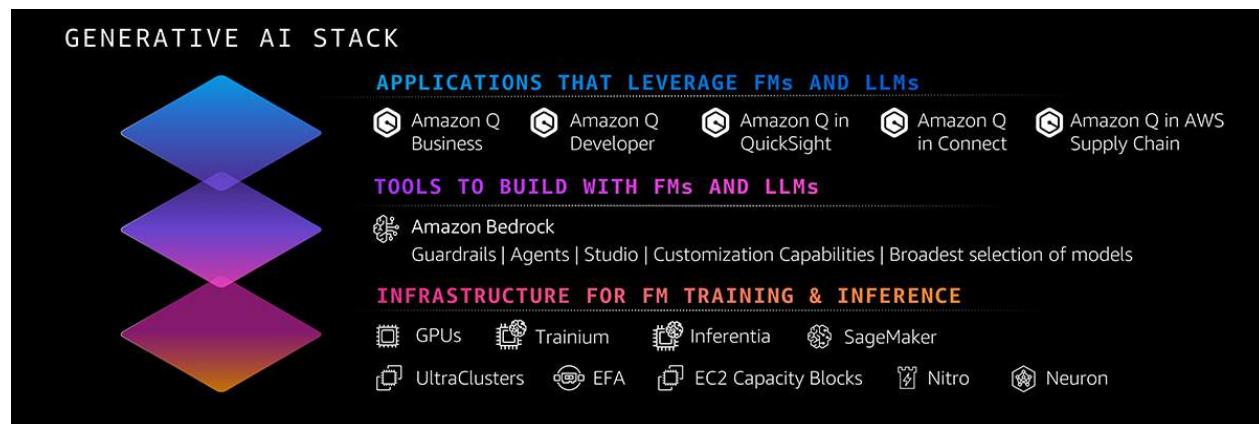
這些功能結合在一起，承諾加速開發過程，提高應用效率，並支持動態、上下文感知的互動，為生成式 AI 解決方案提供支持。

[閱讀更多](#)

# AWS 在紐約峰會上揭示新的生成式 AI 創新

生成式人工智慧 | AWS | Amazon Q | Amazon Bedrock | 應用程式開發 | 程式碼建議 | 數據整合

2024-07-10



## AWS 在紐約峰會上揭示新的生成式 AI 創新

在最近的 AWS 紐約峰會上，亞馬遜展示了其對生成式人工智能 (GenAI) 的承諾，旨在幫助員工快速而安全地構建、客製化和部署應用程式。主要亮點包括 Amazon Q，這是一個生成式 AI 助手，透過生成準確的程式碼建議和自動化多步驟任務來提升軟體開發效率。

此外，Amazon Bedrock 提供工具讓企業能夠創建自訂的 GenAI 應用程式，允許快速整合公司數據。新功能使用戶能夠從簡單的提示生成應用程式，並透過記憶保留功能來改善用戶互動的個人化程度。

AWS 還推出了 Amazon Q Apps，讓員工能夠輕鬆根據公司數據創建應用程式。像是 Retrieval Augmented Generation (RAG) 和進階數據連接器等創新進一步增強了生成式 AI 應用程式的功能，使其更加相關和高效。這些發展使 AWS 在 GenAI 的前沿，改變了組織如何利用科技來提升生產力。

[閱讀更多](#)

# 在 Amazon Bedrock 中微調 Anthropic 的 Claude 3 Haiku：客製化的新時代

**Amazon Bedrock** | **Claude 3 Haiku** | 微調 | 大型語言模型 | AI 應用 | 效率

2024-07-10

The screenshot shows the 'Custom models' section of the Amazon Bedrock console. It includes three main sections: 'How it works' (with sub-sections for creating a model, testing a custom model, and using a custom model), a 'Models' list (containing two entries), and a 'Jobs' section. The 'How it works' section provides a visual overview of the workflow: creating a model involves a neural network icon, testing involves a play button icon, and using involves a cube icon.

**Custom models** Info

Customize model with Fine-tuning or Continued Pre-training.

▼ How it works

Create a model

Test a custom model

Use a custom model

Models Jobs

Models (2)

Models that you have customized and have had their jobs successfully completed will appear here.

Find model

Custom model name

Purchase Provisioned Throughput

Customize model

Create Fine-tuning job

Create Continued Pre-training job

## 在 Amazon Bedrock 中微調 Anthropic 的 Claude 3 Haiku：客製化的新時代

最近，Amazon 為使用大型語言模型 (LLMs) 的開發者推出了一項重大增強功能，允許在 Amazon Bedrock 中微調 Anthropic 的 Claude 3 Haiku。這一過程使用戶能夠針對特定任務或行業，利用自己的數據來調整模型，從而提升性能和準確性。

微調涉及調整預訓練模型的參數，以更好地對應特定的數據集或任務。這一過程通過 Amazon Bedrock 進行，該平台提供了一個安全且管理良好的環境來進行客製化。

Claude 3 Haiku 具備速度快和成本效益高的特點，使其成為尋求效率的企業理想選擇。這項微調的優勢可應用於各種應用，包括分類、結構化輸出和行業特定知識。由於此功能目前在美國西部（俄勒岡）地區處於預覽階段，開發者被鼓勵探索這項創新能力，以提升他們的 AI 應用。

[閱讀更多](#)

# 人工智能如何改變癌症診斷與病患結果

人工智能 | 癌症診斷 | 機器學習 | FDA | 病患結果 | 影像學 | 病理學

2024-07-10

The graphic features a dark background with a starry space theme. At the top right is the SIGGRAPH 2024 logo with the text "SIGGRAPH 2024 DENVER 28 JUL - 1 AUG". The main title "NVIDIA CEO Fireside Chat at SIGGRAPH" is centered in large white font. Below it, the date "July 29 | 2:30 p.m. MT" is also in white. A green button at the bottom left contains the text "Register Now". Below the text are two headshots: Jensen Huang on the left and Lauren Goode on the right. Their names and affiliations are written below their respective portraits.

**Jensen Huang**  
NVIDIA

**Lauren Goode**  
WIRED

在最近的一集 NVIDIA AI Podcast 中，Paige 的聯合創辦人及醫療人工智能領域的領導者 Thomas Fuchs 討論了人工智能如何改變癌症診斷和病患結果。Paige 是第一家獲得 FDA 批准的癌症診斷工具的公司。

Fuchs 強調機器學習和視覺模式識別在加速癌症檢測中的角色，展示了人工智能在提升醫療精準度方面的潛力。他概述了 Paige 在癌症影像學和病理學中的 AI 應用所帶來的可喜成果，並同時提到有效檢測癌症所面臨的持續挑戰。Fuchs 的見解反映了一個未來，即人工智能顯著改善醫療提供者進行癌症診斷的方式，最終導致更好的病患照護與結果。

---

閱讀更多

# NVIDIA 推出 NIM 微服務以簡化生成式 AI 應用部署

NVIDIA | NIM微服務 | 生成式AI | 模型容器 | 客戶支持 | 虛擬助手 | 數據隱私

2024-07-10

The graphic features a dark background with a starry space theme. At the top right is the SIGGRAPH 2024 logo with the text "SIGGRAPH 2024" and "DENVER 28 JUL - 1 AUG". The main title "NVIDIA CEO Fireside Chat at SIGGRAPH" is centered in large white font. Below it, the date "July 29 | 2:30 p.m. MT" is also in white. A green button at the bottom left contains the text "Register Now". Two headshots are shown below the title: Jensen Huang on the left and Lauren Goode on the right. Their names and affiliations are written below their respective portraits.

**Jensen Huang**  
NVIDIA

**Lauren Goode**  
WIRED

NVIDIA 推出了 NIM 微服務，簡化了生成式 AI 應用程式的部署。透過微服務架構，開發者可以將應用程式構建為一系列鬆耦合的服務，每個服務都專為特定任務設計。這種模組化設計使得更新變得更容易，並能根據需求獨立擴展，以提升效率。

NIM 微服務提供預先優化的 AI 模型容器，簡化將 AI 功能整合到各種應用程式中的過程。它們可以在不同的環境中部署，包括工作站和資料中心，減輕與延遲、安全性及雲端服務相關的成本顧慮。

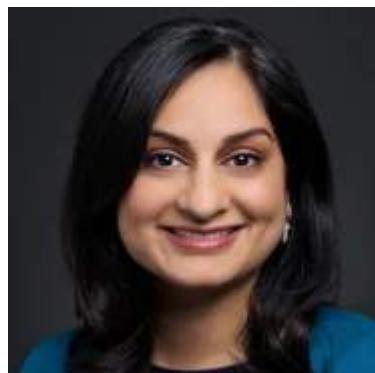
最近的發展包括能在 NVIDIA RTX 系統上本地運行 Meta Llama 3 8B 模型。這一進步使得開發者能夠創建複雜的應用程式，例如客戶支持聊天機器人和互動虛擬助手，同時確保數據隱私。隨著 NVIDIA 創新的 NIM 框架，利用生成式 AI 的潛力變得比以往任何時候都更容易。

[閱讀更多](#)

# Google 在 Galaxy Unpacked 上為三星設備帶來的精彩更新

Google | Galaxy Unpacked | 三星 | Gemini | Wear OS 5 | YouTube TV

2024-07-10



## Google 在 Galaxy Unpacked 上為三星設備帶來的精彩更新

在最近的 Galaxy Unpacked 活動中，Google 突出了四項創新更新，旨在提升三星最新設備的使用體驗，包括 Galaxy Z Flip6、Z Fold6 和新的 Galaxy Watches。

1. Gemini 整合：Gemini 應用程式將很快根據你的螢幕內容提供個性化建議。例如，在觀看視頻時，你可以輕鬆訪問相關資訊或提出問題。
2. 圓圈搜尋：此功能允許使用者直接從螢幕進行搜尋，而無需切換應用程式。即將推出的增強功能包括對更複雜主題的支援以及掃描條碼的能力。
3. Wear OS 5：新的 Galaxy Watches 將運行 Wear OS 5，帶來性能提升和健康監測功能，如心率和睡眠追蹤。
4. YouTube TV 多視窗：Galaxy Z Fold6 使用者將能夠在 YouTube TV 上同時觀看多達四個串流，提升觀看體驗。

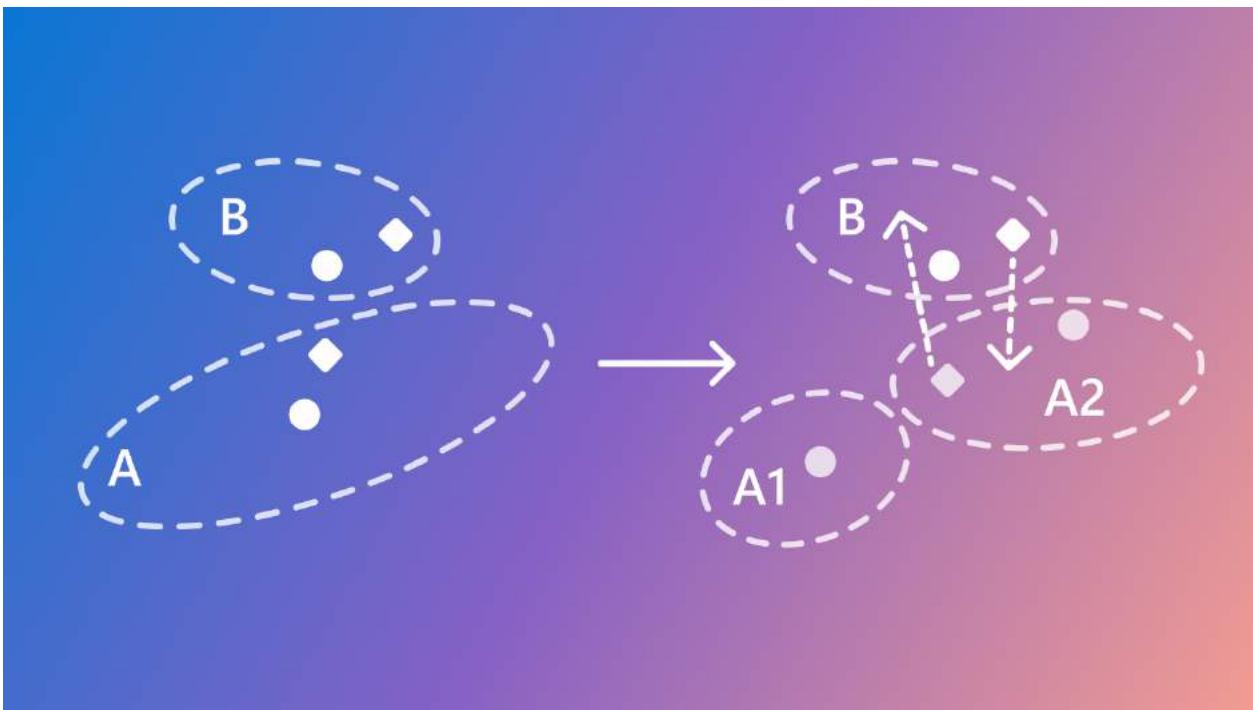
這些更新承諾將使三星設備變得更加直觀和連接。

[閱讀更多](#)

# 微軟的統一資料庫：大型語言模型的重大突破

大型語言模型 統一資料庫 檢索增強生成 資料管理 AI技術

2024-07-10



## 微軟的統一資料庫：大型語言模型的重大突破

微軟研究部門揭示了在統一資料庫方面的突破性進展，旨在提升大型語言模型（LLMs）的性能。當前的LLMs通常依賴固定的訓練數據，這有時會導致不準確的情況。為了解決這個問題，檢索增強生成（RAG）方法將即時外部資訊整合到模型中，從而提高它們的可靠性。

這項創新的核心是一個統一的資料庫，能夠通過一個名為VBase的新系統管理多種數據類型—文本、圖像等。這個系統透過允許同時掃描向量和標量索引來提高查詢效率。此外，SPFresh為向量資料庫引入了一個即時的增量更新過程，顯著減少了通常與數據更新相關的資源負擔。

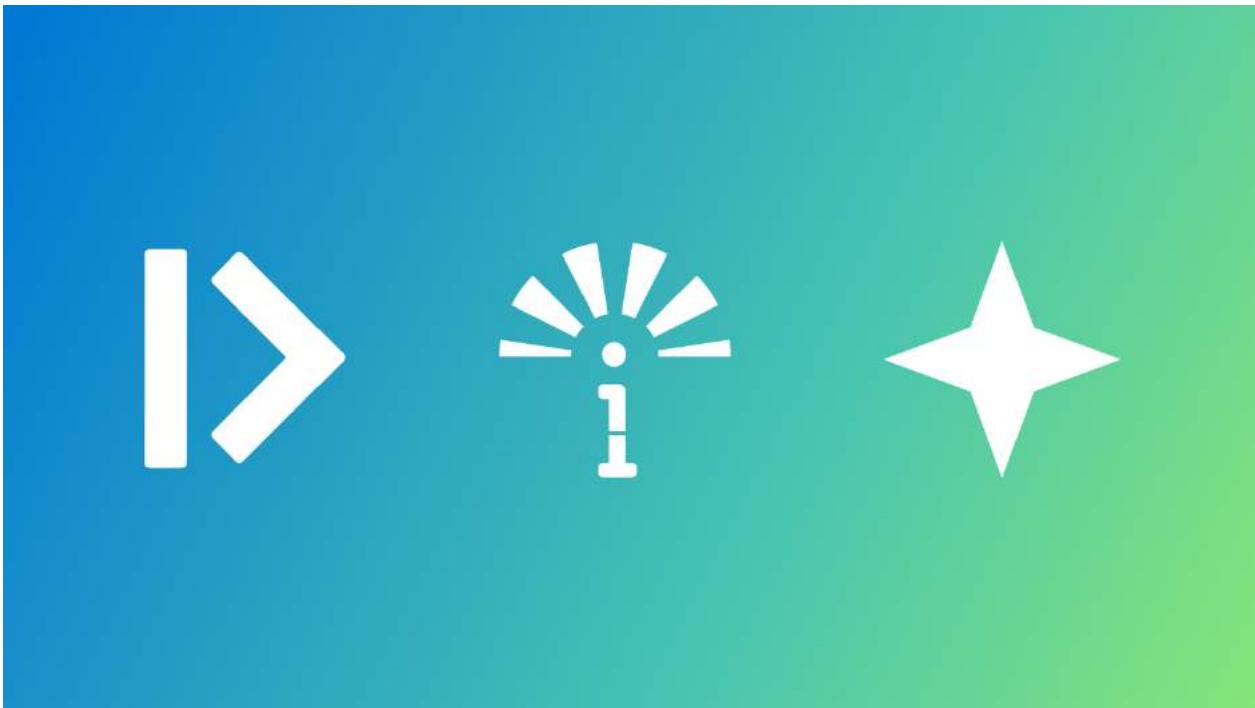
這些進展不僅支持更準確和及時的信息檢索，還承諾推動AI和硬體技術的未來發展。

[閱讀更多](#)

# 微軟研究：利用生成式人工智慧對抗人類販運

生成式人工智慧 | 人類販運 | 非政府組織 | 數據分析 | 智慧工具包

2024-07-10



## 微軟研究：利用生成式人工智慧對抗人類販運

微軟研究正在開創使用生成式人工智慧來協助非政府組織（NGO）在不懈對抗人類販運的努力。這一創新方法源於與「對抗販運科技加速器」（Tech Against Trafficking）合作，該加速器連結科技公司與反販運組織，以利用科技促進社會公益。

這次合作的一個重要成果是智慧工具包（Intelligence Toolkit），這是一個旨在簡化非政府組織數據分析和報告的系統。這個工具包自動生成來自龐大數據集的見解，使組織能夠更專注於直接援助，而不是數據管理。其顯著特點包括合成私人數據集的工具、檢測案件記錄中的模式，以及分析涉及勞動剝削的實體網絡。

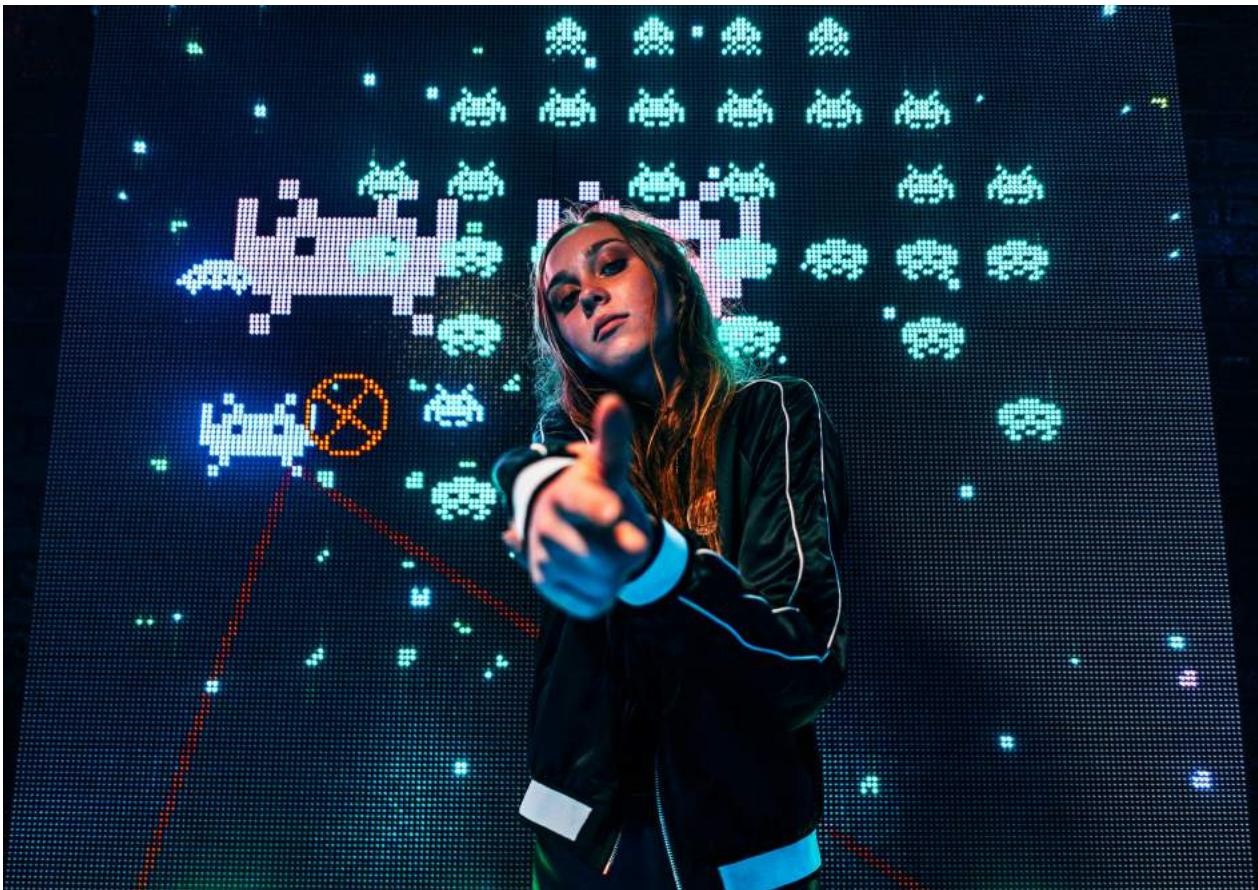
通過整合生成式人工智慧，微軟研究旨在賦予非政府組織揭露隱藏剝削的能力，並在全球範圍內促進基於證據的集體行動，對抗人類販運。

[閱讀更多](#)

# 人工智慧如何改變遊戲設計和玩家體驗

人工智慧 | 程序化內容生成 | 非玩家角色 | 遊戲設計 | 玩家體驗 | 網路賭場 | 詐騙檢測 | 客戶支援

2024-07-11



人工智慧 (AI) 正在遊戲產業中掀起波瀾，重新塑造遊戲的設計和玩法。關鍵的創新是 程序化內容生成 (PCG)，它允許開發者即時創建廣闊且可適應的遊戲世界，確保每位玩家都有獨特的體驗。

AI提升了 非玩家角色 (NPC) 的互動能力，使它們能夠從玩家的行為中學習，從而實現更真實、更具吸引力的遊戲體驗。此外，它還幫助動態平衡遊戲難度，讓玩家保持挑戰性但不至於感到不知所措。

在網路賭場中，AI被用來提升玩家的體驗，透過先進的詐騙檢測技術增強安全性，並提供個性化的遊戲時刻。此外，AI驅動的客戶支援提供即時協助，為玩家帶來無縫的體驗。

隨著AI的不斷演進，其在遊戲中的角色將有更深遠的發展，為未來帶來更具沉浸感和個性化的體驗。

---

閱讀更多

# 個人電腦市場在 AI 興起中獲得牽引

個人電腦 | AI | 蘋果 | 出貨量 | M 系列晶片

2024-07-11



## 個人電腦市場在 AI 興起中獲得牽引

全球個人電腦市場正在復甦，蘋果公司引領這一趨勢，根據國際數據公司 (IDC) 的數據顯示，2024 年第二季度同比增長 3%。這是經歷了七個季度的下滑後，顯著的轉變，全球個人電腦出貨量達到 6490 萬台。值得注意的是，蘋果的 Mac 出貨量增長了 20.8%，這得益於其 M 系列晶片的受歡迎程度。

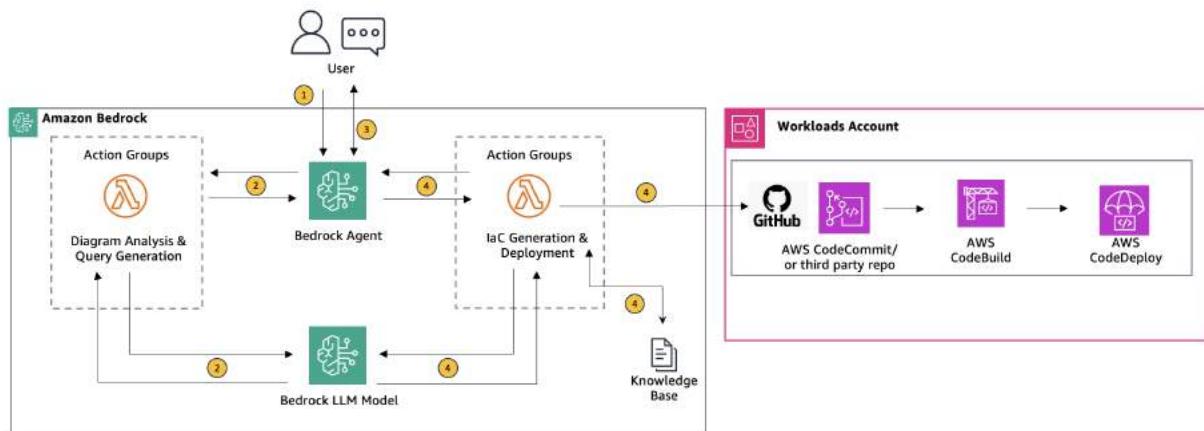
分析師將這一復甦歸因於多種因素，包括商業更新周期及對具備 AI 功能的個人電腦日益增長的興趣。預計 AI 強化的個人電腦的潛力將推動市場的進一步增長，蘋果、高通、英特爾和 AMD 等主要廠商正準備揭示 AI 整合的策略。儘管各地區之間仍存在差異，尤其是在中國，但整體個人電腦行業的前景令人鼓舞，創新和持續增長的機會依然存在。

[閱讀更多](#)

# AWS 發佈 Amazon Bedrock 的代理程式：基礎設施即代碼的飛躍

**Amazon Bedrock** | 基礎設施即代碼 | IaC | 提示工程 | 雲端採用 | Terraform | AWS CloudFormation

2024-07-11



## AWS 發佈 Amazon Bedrock 的代理程式：基礎設施即代碼的飛躍

Amazon Web Services (AWS) 推出了 Amazon Bedrock 的代理程式，這是一個旨在簡化生成基礎設施即代碼 (IaC) 過程的創新工具。這項技術自動化了提示工程，使團隊能夠輕鬆地從架構圖生成量身定製的 IaC 腳本。

透過檢索增強生成 (Retrieval Augmented Generation, RAG)，這些代理程式可以分析用戶上傳的圖表，識別缺失的組件，並與用戶進行實時對話以收集必要的信息。這一互動過程確保生成的腳本符合組織標準和安全要求。

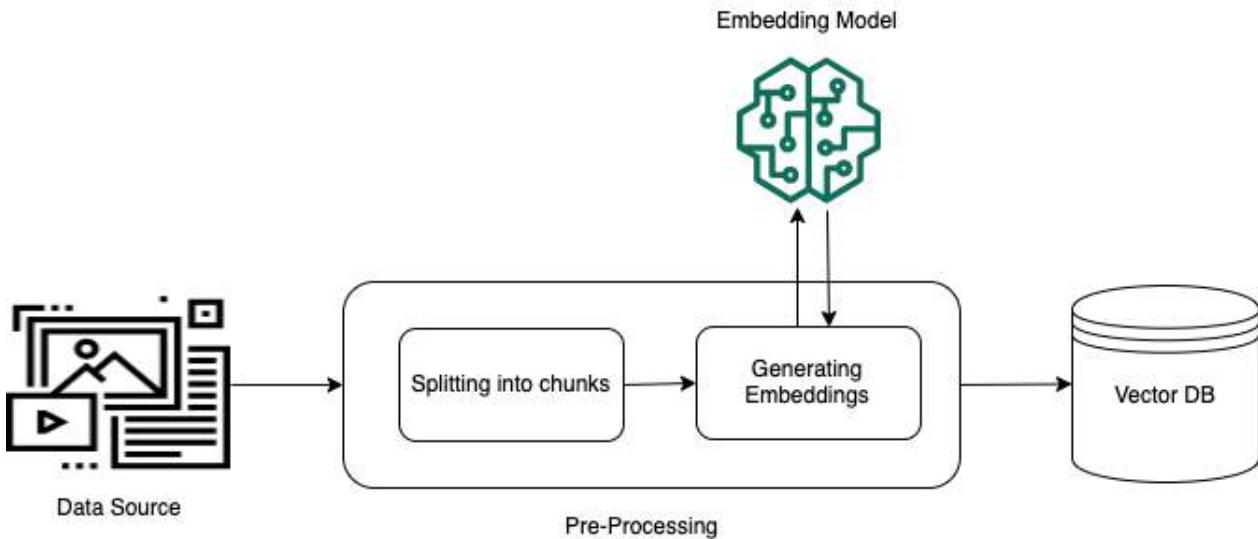
一旦用戶確認配置，代理程式就會生成 IaC 腳本，例如 Terraform 或 AWS CloudFormation 模板，並自動將其上傳到指定的 GitHub 儲存庫。這一進展加速了部署，減少了錯誤，並簡化了持續的基礎設施管理，標誌著雲端採用和操作效率的重要進步。

[閱讀更多](#)

# 提升 AWS SageMaker 的檢索增強生成能力

AWS SageMaker | 檢索增強生成 | 嵌入模型 | 領域微調 | 專業數據集 | Sentence Transformer

2024-07-11



## 提升 AWS SageMaker 的檢索增強生成能力

亞馬遜最近在其 SageMaker 平台上引入了對檢索增強生成 (RAG) 方法的改進，使用經過微調的嵌入模型。RAG 通過允許大型語言模型 (LLMs) 從外部數據來源獲取額外知識來增強其性能。

然而，預訓練的嵌入模型在處理特定領域的細微差異時經常遇到困難，這可能導致不準確的結果。為了解決這個問題，AWS SageMaker 促進了使用特定領域數據對嵌入模型進行微調，從而改善專業概念的表示，這在法律、醫療和技術領域尤為重要。

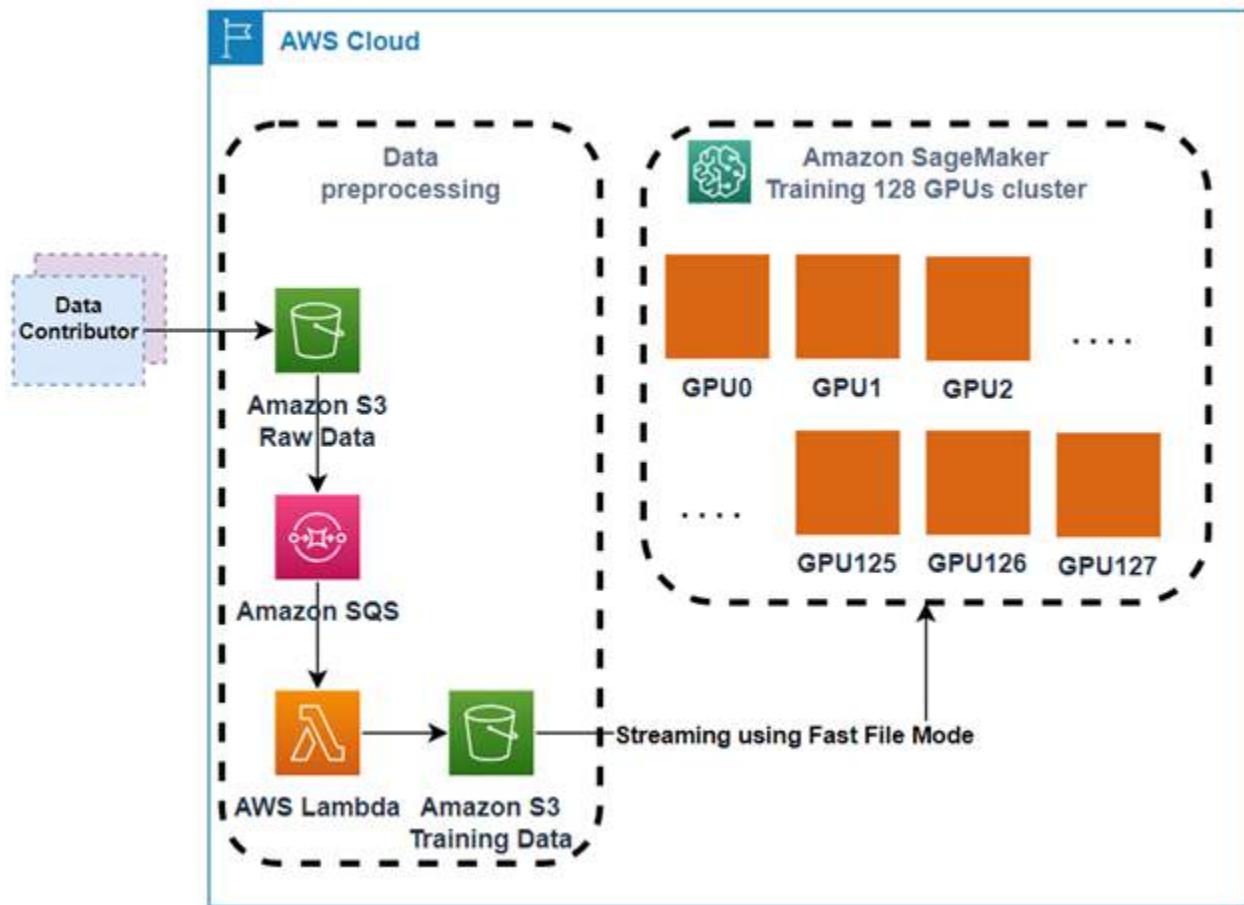
該過程涉及在專業數據集上訓練 Sentence Transformer 嵌入模型，使其能夠學習相關的語義和關係。這一改進顯著提高了 RAG 系統所生成回應的準確性。該計劃旨在幫助開發者在各自領域中創建更為精確和上下文相關的應用程序。

[閱讀更多](#)

# BRIA AI 利用 Amazon SageMaker 進行高效模型訓練

**BRIA AI** | **Amazon SageMaker** | 文字轉圖像模型 | 分散式訓練 | **GPU** | 數據加載 | 生成式 AI

2024-07-11



## BRIA AI 利用 Amazon SageMaker 進行高效模型訓練

BRIA AI 最近利用 Amazon SageMaker 的分散式訓練能力，開發出其高解析度的文字轉圖像模型 BRIA AI 2.0。這個模型能夠生成 1024x1024 的圖像，並在一個龐大的授權圖像數據集上進行訓練，透過 SageMaker 強大的功能有效地管理基礎設施挑戰。

SageMaker 的分散式訓練庫讓 BRIA AI 能夠將訓練時間從幾個月大幅縮短至不到兩週，這是通過在多個強大的 GPU 實例上運用數據並行性來實現的。這項創新最大化了 GPU 的利用率，並通過僅對活躍的訓練時間收費來降低成本。

訓練流程還實現了從 Amazon S3 的無縫數據加載，優化了整個過程。通過解決模型收斂和訓練效率等問題，BRIA AI 使自己成為負責任的生成式 AI 的領導者，準備以其尖端科技支持各種創意產業。

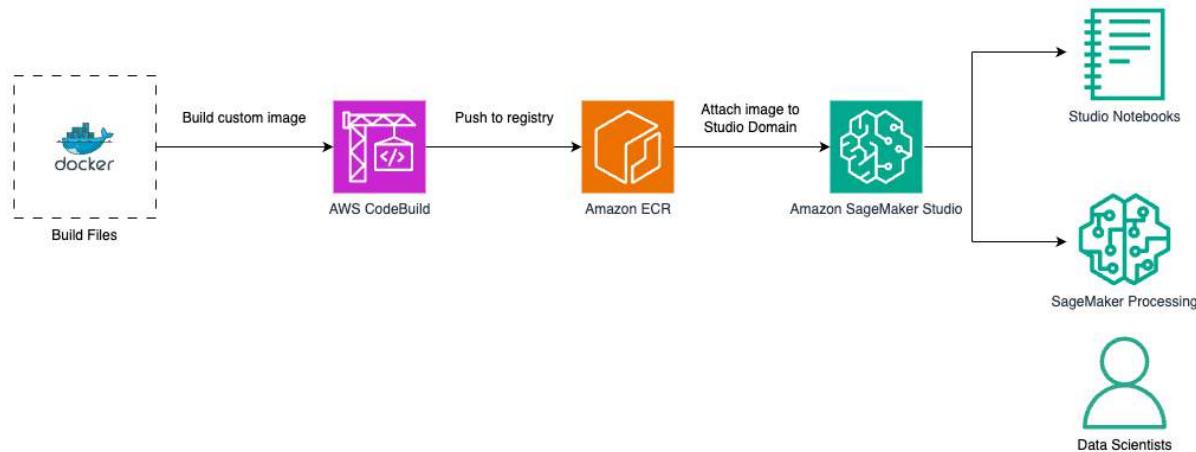
---

閱讀更多

# 針對地理空間分析的自訂影像與 Amazon SageMaker Studio

地理空間分析 | 自訂影像 | **Amazon SageMaker Studio** | 機器學習 | 數據科學 | Docker | GeoTIFF  
GeoJSON

2024-07-11



## 針對地理空間分析的自訂影像與 Amazon SageMaker Studio

Amazon SageMaker Studio 推出了一項新功能，使用戶能夠創建專門為地理空間分析設計的自訂影像。這項創新建立在 SageMaker 的完全管理的整合開發環境 (IDEs) 上，例如 JupyterLab，以增強涉及地理空間數據的機器學習工作流程的可用性。

隨著地理空間數據集的興起——如衛星影像和座標追蹤——此功能使用戶能夠處理各種專門格式，如雲優化的 GeoTIFF 和 GeoJSON。通過擴展 SageMaker Distribution 與開源的地理空間庫，用戶可以自訂其環境，以有效管理和分析地理空間數據。

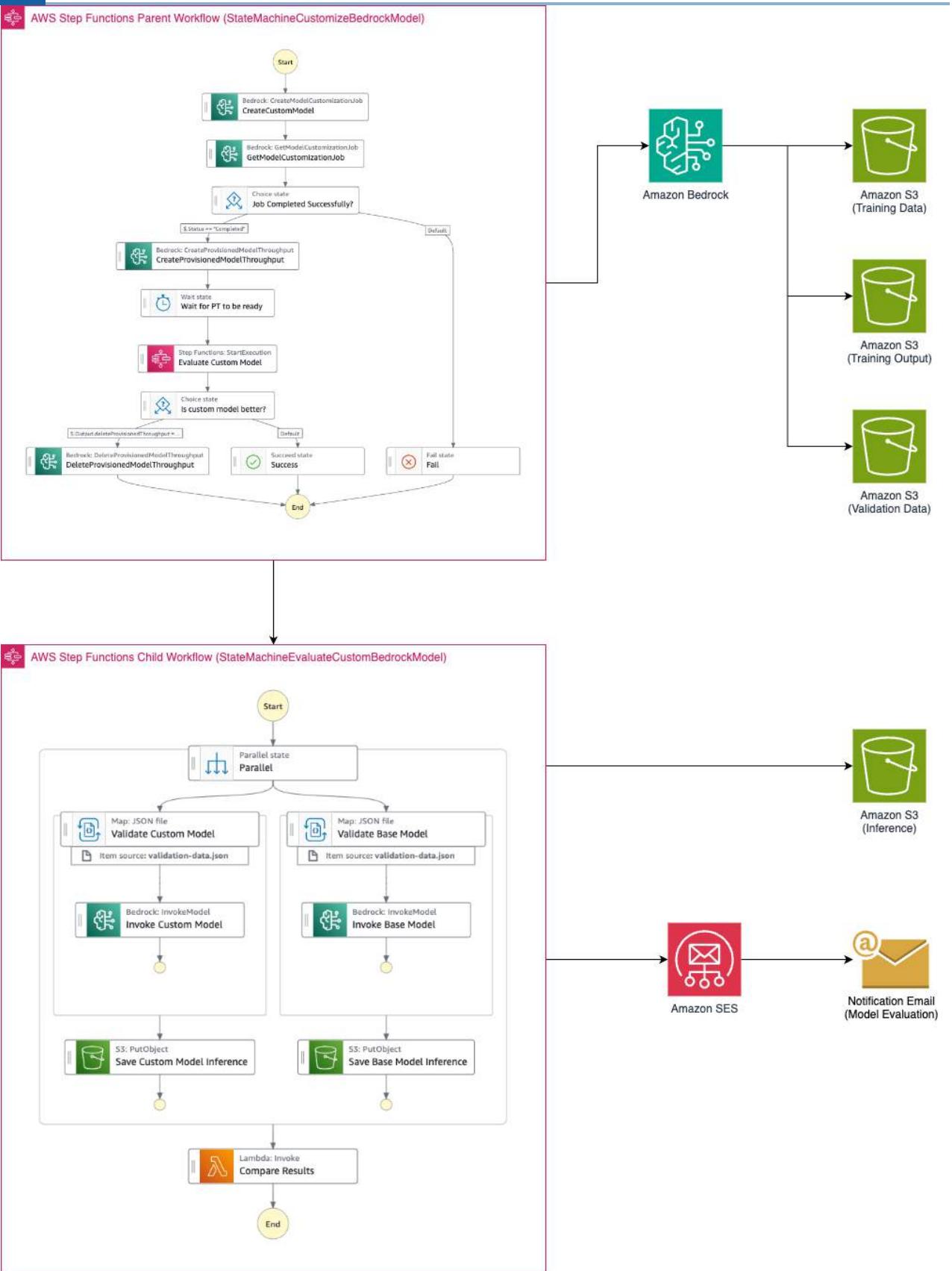
該過程包括創建一個包含必要庫的 Docker 映像，將其部署到 Amazon Elastic Container Registry，並將其附加到 SageMaker Studio。這使數據科學家能夠無縫進行互動式分析和高效能的地理空間處理任務，為城市規劃、環境監測和災害應對等領域開創新的可能性。

[閱讀更多](#)

# Amazon Bedrock 強化模型客製化功能，結合 AWS Step Functions

**Amazon Bedrock** | **AWS Step Functions** | 模型客製化 | 大型語言模型 | **AI 能力**

2024-07-11



## Amazon Bedrock 強化模型客製化功能，結合 AWS Step Functions

Amazon 最近透過整合 AWS Step Functions 改進了其 Bedrock 服務的客製化過程。這項創新簡化了根據特定商業需求調整大型語言模型的能力，使企業能夠自動化客製化工作流程。

借助這次更新，使用者可以輕鬆地用自己專有的數據預先訓練基礎模型，這增強了模型生成回應的相關性和準確性。AWS Step Functions 協調整個過程，處理模型訓練、評估和監控等任務，創建一個可重複的框架，減少這些複雜任務通常所需的時間和精力。

通過簡化模型客製化，Amazon Bedrock 使組織能夠利用先進的 AI 能力，而無需承擔複雜基礎設施管理的負擔，最終幫助他們更有效地解決獨特的數據挑戰。

[閱讀更多](#)

# Amazon Bedrock 強化知識庫，提供先進功能以提升準確性

**Amazon Bedrock** 知識庫 檢索增強生成 解析 查詢重構 數據檢索 信息管理 數據驅動

2024-07-11

**Chunking and parsing configurations** [Info](#)  
Choose between default or advanced customization.

Default (built-in customization)  
Uses default parsing and chunking strategy.

Advanced (customization)  
Customize the parser and chunking strategy.

## Parsing strategy

Parsing analyses and extracts useful information from documents.

Use foundation model for parsing [See supported formats](#)  
Suitable for parsing more than standard text in supported document formats, including tables within PDFs with their structure intact. [View pricing](#)

## Choose foundation model for parsing

 [Claude 3 Sonnet v1](#)

By Anthropic

 [Claude 3 Haiku v1](#)

By Anthropic

## ▼ Instructions for the parser - optional

- |   |   |
|---|---|
| 1 | Transcribe the text content from an image page and output in Markdown syntax (not code blocks). Follow these steps: |
| 2 |   |
| 3 | 1. Examine the provided page carefully.   |

## Amazon Bedrock 強化知識庫，提供先進功能以提升準確性

Amazon Bedrock 引入了一系列先進功能，顯著增強了檢索增強生成 (RAG) 應用程序的能力。值得注意的是，該系統現在支援先進的解析、分塊和查詢重構，能夠更好地處理複雜文件，如 PDF 和 CSV 檔。

新的先進解析能力使得從非結構化文件中提取有意義的信息成為可能，進而提高數據檢索的準確性。增強的分塊選項—語義和層次—以保持上下文關係的方式組織信息，導致更具連貫性的回應。

此外，查詢重構的引入使得複雜查詢可以被拆分成可管理的子查詢，從而提高獲取信息的相關性。這些增強功能賦予用戶更大的數據管理控制權，促進在數據驅動環境中的有效決策。Amazon Bedrock 持續進化，為知識管理和信息檢索設立新的標準。

[閱讀更多](#)

# 日本的 ABCI 3.0 超級電腦：AI 主權的飛躍

ABCI 3.0 | 超級電腦 | 人工智慧 | NVIDIA | 技術獨立性 | 生成式 AI | 計算能力 | 柏市

2024-07-11



## 日本的 ABCI 3.0 超級電腦：AI 主權的飛躍

日本透過引進由惠普企業 ( Hewlett-Packard Enterprise, HPE ) 設計的 ABCI 3.0 超級電腦，擴大其人工智慧 ( AI ) 實力。這台新超級電腦整合了數千顆 NVIDIA H200 張量核心 GPU，並採用了 NVIDIA Quantum-2 InfiniBand 網路，確保高效能和可擴展性。

ABCI 3.0 是日本提升其生成式 AI 研究與開發工作的倡議的一部分，進一步鞏固其技術獨立性。它擁有驚人的計算能力，能達到 6 AI exaflops 和 410 petaflops 的一般任務處理能力，能快速處理複雜數據集。

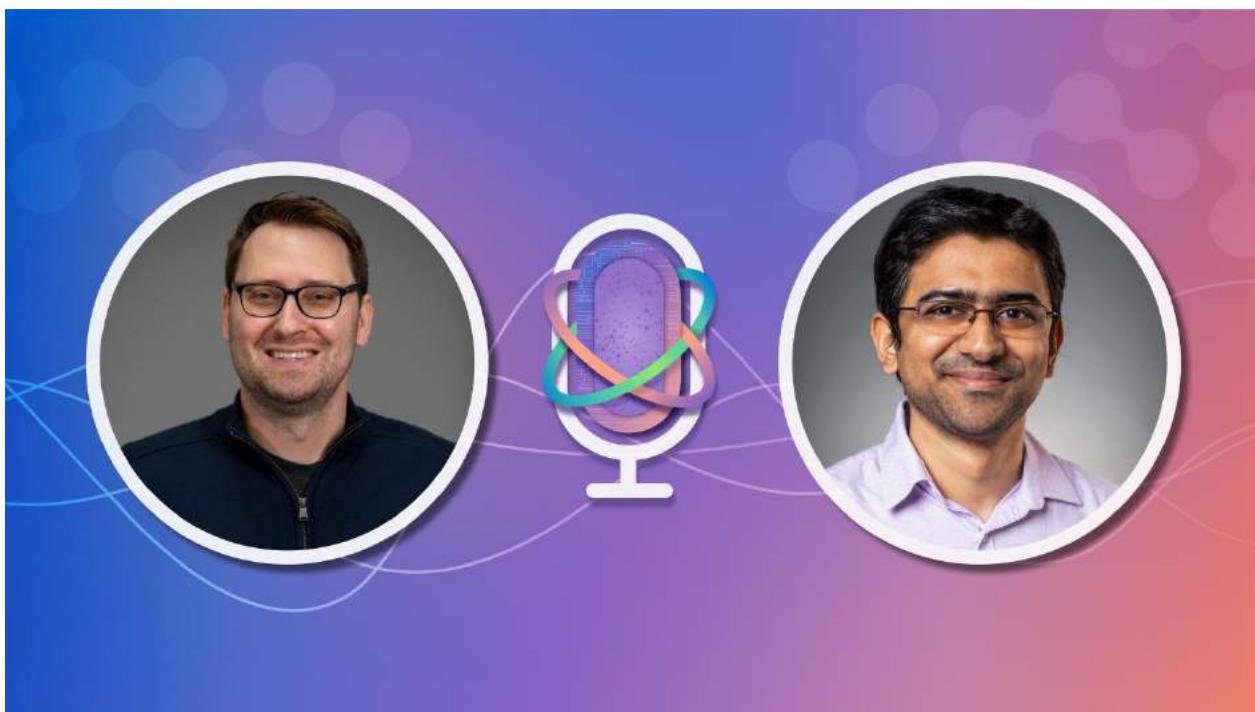
這座設施位於柏市，旨在成為學術界、產業界及政府之間的合作平台，最終使日本在全球 AI 領域中占有重要地位。這台超級電腦預計將於今年底啟用，標誌著 AI 研究與創新的新時代。

[閱讀更多](#)

# 創根基於 Vitrimer 的印刷電路板：一種可持續的解決方案

Vitrimer | 印刷電路板 | 可持續材料 | 機器學習 | 電子廢物

2024-07-11



## 創根基於 Vitrimer 的印刷電路板：一種可持續的解決方案

在最近一集的 Microsoft Research Podcast 中，資深研究員 Jake Smith 和華盛頓大學的 Aniruddh Vashisth 討論了基於 Vitrimer 的印刷電路板 ( vPCBs ) 的開發。這些創新的電路板在性能上與傳統電路板相當，同時顯著減少環境影響。

Vitrimer 是一種獨特的聚合物，因其能夠「解扣」和「重新扣合」而可以進行回收再加工，這使得組裝和拆卸變得更加容易。這一特性使得 vPCBs 成為各行業中的可持續替代方案，特別是在航空航天和汽車領域，減少電子廢物至關重要。這項研究還利用機器學習來加速可持續材料的發現，提升設計過程的整體效率。

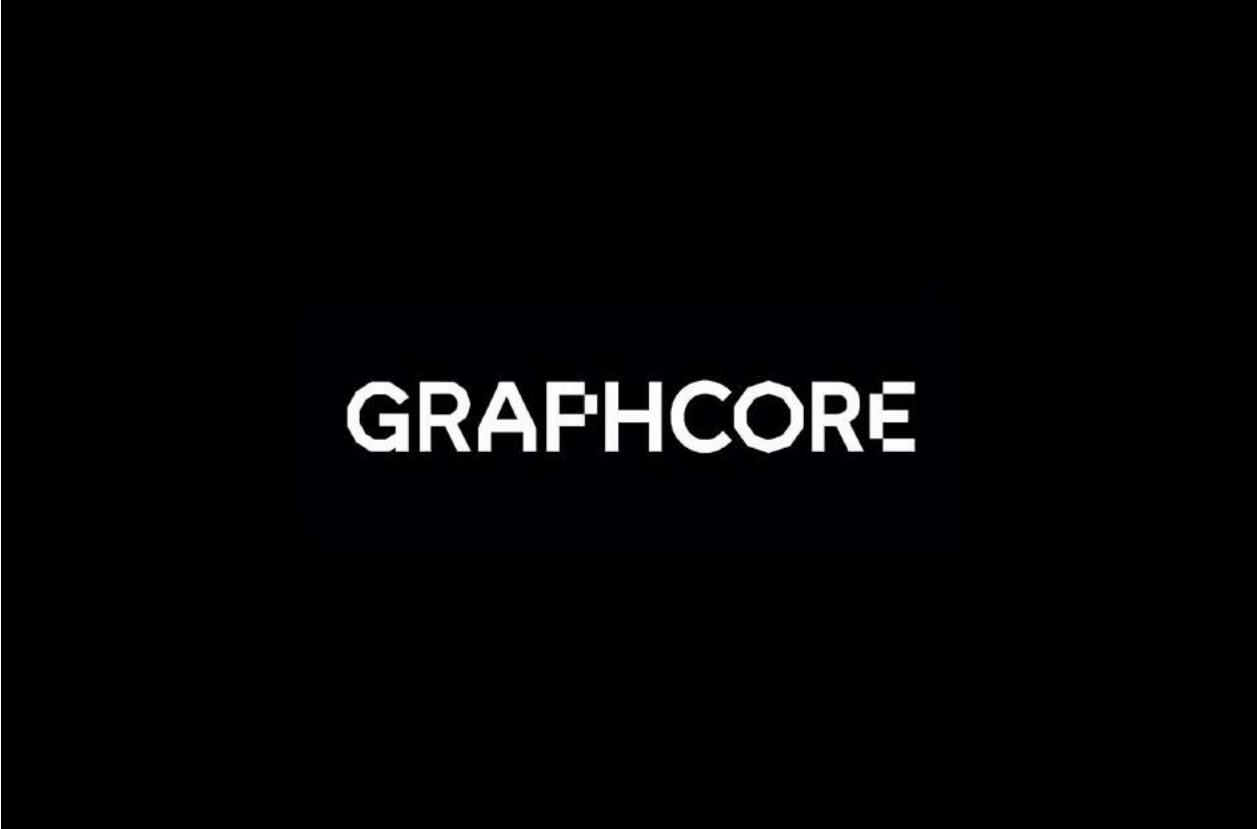
這一合作努力不僅解決了電子廢物的問題，還為未來的電子產品提供了一個展望，讓設計以可持續性為核心。

[閱讀更多](#)

# 軟銀收購英國AI晶片廠商Graphcore

軟銀 收購 Graphcore AI 晶片 IPUs 計算效率 能耗 NVIDIA Intel

2024-07-12



GRAPHCORE

## 軟銀收購英國AI晶片廠商Graphcore

在科技界的一個重大舉措中，軟銀以約6億美元收購了英國公司Graphcore，該公司因其先進的AI晶片技術而聞名。Graphcore以其智能處理單元 ( Intelligence Processing Units, IPUs ) 著稱，這些晶片專門設計用來加速AI工作負載，這在日益增長的人工智慧應用需求中至關重要。

儘管在收入上面臨挑戰——2022年僅報告270萬美元的營收——此次收購強調了Graphcore在未來AI進步中的潛力。首席執行官Nigel Toon強調了提高計算效率和能耗的重要性，以實現AI的變革能力。

這次與軟銀的整合預期將為Graphcore提供豐富的資源，加強其在與NVIDIA和Intel等知名競爭者的AI晶片市場中的使命。隨著AI技術的持續進化，Graphcore的專業硬體將在塑造各行各業創新解決方案中發揮關鍵作用。

[閱讀更多](#)

# 三星在最新的摺疊手機和可穿戴設備中增強 AI 功能

三星 摺疊手機 可穿戴設備 AI 功能 健康科技 智能健康追蹤器

2024-07-12



## 三星在最新的摺疊手機和可穿戴設備中增強 AI 功能

三星在其最新的旗艦設備上提高了標準，包括 Galaxy Z Fold 6 和 Galaxy Z Flip 6，這兩款產品都設計得更輕更薄，同時具備先進的 AI 整合。Galaxy Z Fold 6 擁有寬大的顯示螢幕，並且是該系列中最輕的，針對新客戶市場。而 Galaxy Z Flip 6 則改善了電池壽命和相機解析度，並引入了全新的散熱系統。

值得注意的 AI 進步包括與 Galaxy Buds 配合的「聆聽模式」進行即時翻譯，以及與 Google 合作的 AI 搜尋功能，這些功能能夠以視覺方式解決數學問題。此外，Galaxy Watch 也有了顯著升級，包括新的晶片以提升性能，並獲得 FDA 批准進行睡眠呼吸暫停監測。

三星對健康科技的承諾在 Galaxy Ring 中得以體現，這是一款擁有多項創新功能的智能健康追蹤器，將於 7 月 24 日起上市。這一系列產品標誌著三星在高端智能手機和可穿戴設備市場中維持競爭優勢的關鍵一步。

[閱讀更多](#)

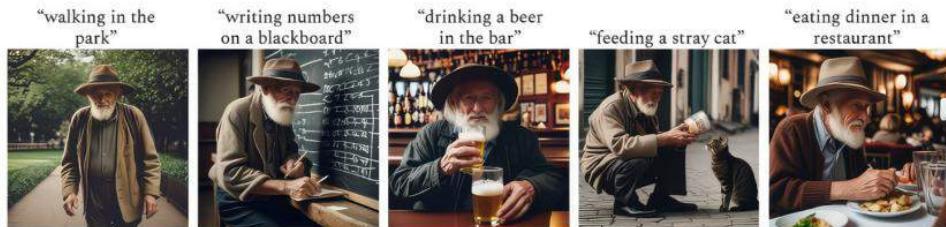
# NVIDIA 在 SIGGRAPH 2024 展示尖端 AI 和模擬技術創新

**NVIDIA SIGGRAPH 2024** 人工智慧 模擬技術 合成數據生成 反向渲染 擴散模型 物理基礎模擬 SuperPADL 渲染方法

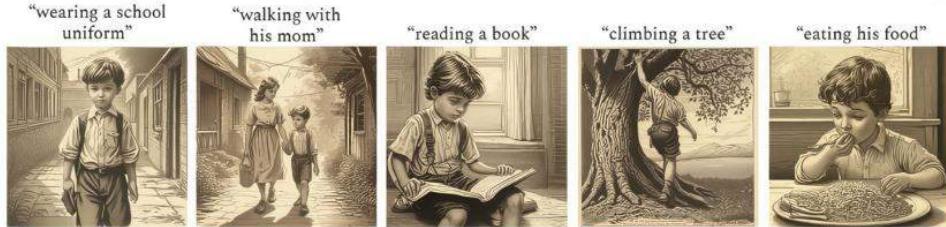
2024-07-12

## Subject Description

"A photo of an old man wearing a hat."



"An old story illustration of a kid."



"A hyper-realistic digital painting of a happy girl, brown eyes."



Single Subject

Multi Subject

## NVIDIA 在 SIGGRAPH 2024 展示尖端 AI 和模擬技術創新

在即將於丹佛舉行的 SIGGRAPH 2024 會議上，NVIDIA 研究人員將揭示他們在人工智慧和模擬技術方面的最新進展。此次會議將重點展示超過 20 篇研究論文，主要集中於改善合成數據生成和反向渲染工具，以協助訓練先進的 AI 模型。值得注意的創新包括擴散模型，這些模型能增強視覺生成，使藝術家能在短時間內產出一致的影像。

研究還強調物理基礎的模擬，架起真實與虛擬物體行為之間的橋樑。像是 SuperPADL 的技術可以實現複雜人類動作的即時模擬。此外，新的渲染方法承諾加快光線及其他現象的模擬，顯著提升效率。

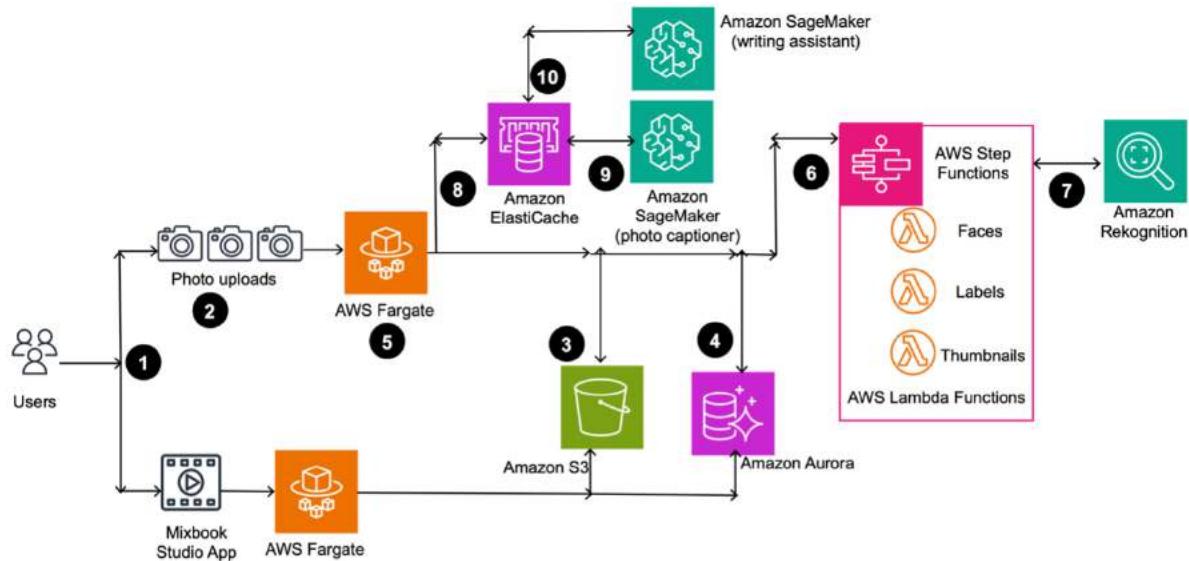
這些技術有潛力徹底改變各個領域，從故事講述、設計到科學研究和自主系統，透過創建更具真實感的虛擬環境，增強 AI 模型的發展。

[閱讀更多](#)

# Mixbook 利用生成式 AI 強化照片書創作

生成式 AI | 照片書 | Smart Captions | AWS | Llama 語言模型 | 影像分析

2024-07-15



## Mixbook 利用生成式 AI 強化照片書創作

Mixbook 最近運用生成式 AI 的力量，徹底改變了用戶創建個人化照片書的方式。他們的創新功能 Smart Captions 幫助用戶選擇和標註照片，增強敘事而不完全自動化創作過程。這個工具能夠解析上傳的影像，加入上下文和情感元素，將回憶轉化為富有表現力的敘述。

後端系統利用多種 AWS 服務，包括 Amazon S3 用於照片儲存，以及 Amazon Rekognition 用於影像分析，使工作流程無縫且高效。標題由調整過的 Llama 語言模型生成，並結合多元的創意團隊輸入，確保最終產出能在個人層面上與用戶產生共鳴。

感謝這些進展，Mixbook 改善了用戶體驗，減少了標註所需的時間，同時增添了驚喜。這項計畫展示了生成式 AI 如何能夠促進創意，並透過敘事加深連結。

[閱讀更多](#)

# 黃仁勳與馬克·祖克柏將於SIGGRAPH 2024探索圖形與虛擬世界

SIGGRAPH 2024 | 黃仁勳 | 馬克·祖克柏 | 人工智能 | 圖形技術 | 生成式AI | 全息技術 | 氣候分析 | 3D  
體驗

2024-07-15

**NVIDIA CEO Fireside Chat  
at SIGGRAPH**  
July 29 | 2:30 p.m. MT

**Register Now**

**Jensen Huang**  
NVIDIA

**Lauren Goode**  
WIRED

SIGGRAPH 2024  
DENVER • 28 JUL – 1 AUG

黃仁勳與馬克·祖克柏將於SIGGRAPH 2024探索圖形與虛擬世界

在即將於丹佛舉行的SIGGRAPH 2024大會上，NVIDIA執行長黃仁勳與Meta執行長馬克·祖克柏將於7月29日進行公開討論，重點聚焦於人工智能與圖形技術中模擬的交集。這場關鍵對話將在黃仁勳與WIRED的Lauren Goode的對話之前，揭示全新的計算格局。

與會者可以期待看到近100家展商展示圖形技術的創新進展，特別強調SIGGRAPH創新區，初創公司在此推動技術邊界。值得注意的創新包括Tomorrow.io使用NVIDIA Earth-2進行氣候分析，以及Looking Glass的全息技術，能夠在不需頭戴式裝置的情況下創造3D體驗。

此外，SIGGRAPH還將舉辦生成式AI日，業界領袖將討論生成式AI對視覺特效和動畫的影響。從7月28日至8月1日，加入科技社群，一同探索這些塑造未來的突破性發展。

[閱讀更多](#)

# RUBICON：提升人機對話體驗

**RUBICON** | 人機對話 | AI 系統 | 評估框架 | 編碼環境 | 效果評估 | 自動化技術 | **Visual Studio** | AI 助手 | 語境意識

2024-07-15



## RUBICON：提升人機對話體驗

Microsoft Research 推出了 RUBICON，這是一個創新的評估框架，旨在提升人類與 AI 系統之間的互動，特別是在編碼環境中。隨著像 GitHub Copilot 這樣的 AI 工具不斷進化，評估其效果變得至關重要。傳統的反饋方法往往無法捕捉用戶體驗的細微差別，尤其是在專業環境中。

RUBICON 採用了自動化的基於評分規範的技術，將有限的數據轉化為全面的評估。通過應用基本的溝通原則，RUBICON 確保 AI 對話具有上下文意識，並促進富有成效的交流。例如，在 Visual Studio 中，RUBICON 幫助 AI 助手在除錯過程中提供詳細且相關的指導，並根據用戶的輸入進行調整。

該框架的有效性表現在其在分類對話方面比以往方法提高了 18% 的準確性。這個工具不僅優化了 AI 的互動，還能識別更廣泛的設計問題，為 AI 在各種應用中的進一步發展鋪平了道路。

[閱讀更多](#)

# AI 有潛力為英國生產力釋放 1190 億英鎊

人工智慧 | 生產力 | 大型企業 | 員工工時 | 信任問題 | 安全顧慮

2024-07-16



## AI 有潛力為英國生產力釋放 1190 億英鎊

來自 Workday 的最新洞見指出，人工智慧 (AI) 技術有潛力為英國企業釋放高達 1190 億英鎊的生產力。這一發現正值關鍵時刻，因為英國面臨自 15 年前以來持續存在的生產力挑戰，目前的生產力水平比 2008 年前的預期低了 24%。

報告強調，大型企業若能策略性地採用 AI，每年可節省約 79 億名員工的工時。在個人層面，企業領導者每年可重獲 1117 小時，這相當於 140 個工作天，而員工則可回收 737 小時，或 92 個工作天。

儘管前景樂觀，但融入 AI 的過程面臨障礙，包括信任問題、安全顧慮和資源不足。解決這些障礙對於實現 AI 的潛力至關重要，這將使員工能夠專注於更具影響力的任務，從而提升工作場所的參與感和生產力。

[閱讀更多](#)

# 探索最佳的 AI 驅動遊戲筆記型電腦與桌上型電腦

**AI 駕動遊戲筆記型電腦 | 遊戲性能 | 處理器 | 顯示卡 | NVMe SSD | RAM**

2024-07-16



## 探索最佳的 AI 駕動遊戲筆記型電腦與桌上型電腦

隨著遊戲愛好者轉向使用筆記型電腦和桌上型電腦以獲得更身歷其境的體驗，AI 駕動設備的興起正在改變這個領域。這些先進的機器專為處理複雜的運算任務而設計，顯著提升遊戲性能。

選擇 AI 駕動設備時需考量的關鍵因素包括強大的處理器，理想情況下應為 Intel Core i7 或 i9，配合強大的顯示卡。高品質的儲存設備，特別是 NVMe SSD，以及充足的 RAM 可確保快速的讀取時間，這對於競技遊戲至關重要。

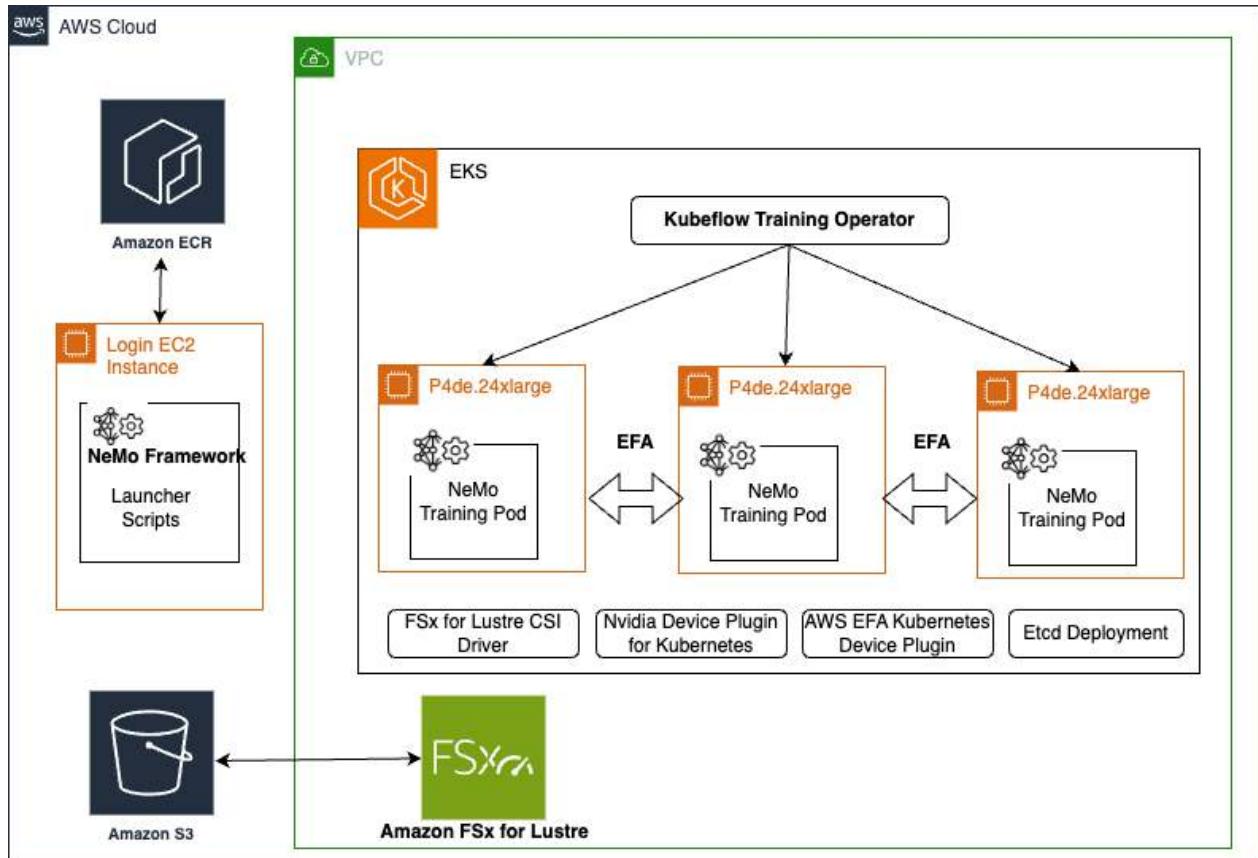
突出的選擇包括 Alienware Aurora R14，搭載尖端的 NVIDIA GeForce 顯示卡和 Intel Core i9。MSI Trident X 提供了緊湊的設計而不妥協於性能。對於喜愛自訂的玩家，HP Omen Obelisk 提供輕鬆升級的便捷性。如果多功能性是關鍵，HP Spectre x360 14 在處理遊戲和專業任務方面表現優秀，而 ASUS ROG Zephyrus G14 則提供經濟實惠且強大的遊戲體驗。

[閱讀更多](#)

# NVIDIA NeMo 框架提升 Amazon EKS 上生成式 AI 的訓練效能

**NVIDIA NeMo | 生成式 AI | Amazon EKS | 模型訓練 | 大型語言模型**

2024-07-16



## NVIDIA NeMo 框架提升 Amazon EKS 上生成式 AI 的訓練效能

NVIDIA 推出了 NeMo 框架，這是一個創新的解決方案，用於加速在 Amazon Elastic Kubernetes Service (EKS) 上的生成式 AI 模型訓練。該框架解決了訓練大型語言模型 (LLMs) 所帶來的複雜性，這些模型通常需要大量的計算資源和精密的管理。

NeMo 提供了一整套健全的工具，簡化了從數據準備到部署的整個模型訓練過程。它的功能包括各種平行處理技術，例如數據平行處理、張量平行處理和專家平行處理，並搭配節省記憶體的策略來優化資源使用。

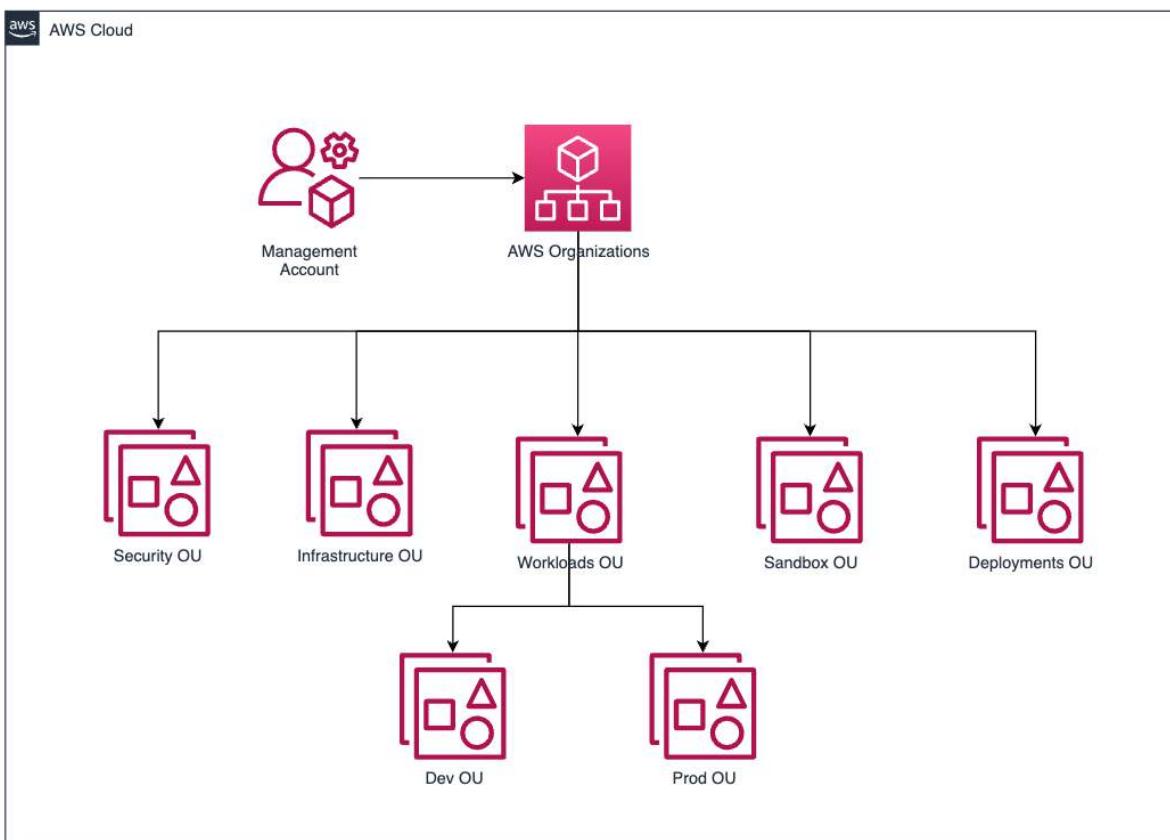
此外，NeMo 與 Amazon EKS 無縫整合，利用 Elastic Fabric Adapter 等功能實現低延遲和高吞吐量的網絡連接。這一組合使企業能夠高效運行分散式訓練工作負載，讓大型 AI 模型的開發變得更可及和具成本效益。



# 建立可擴展的機器學習環境：AWS 多帳戶策略

AWS 機器學習 多帳戶策略 安全性 成本管理 自動化 治理

2024-07-16



## 建立可擴展的機器學習環境：AWS 多帳戶策略

在最新一期的 AWS 機器學習部落格中，重點放在建立一個穩健的多帳戶策略，以優化機器學習（ML）生命週期。透過利用多帳戶的基礎架構，組織可以提升安全性、管理成本並簡化運營。

該策略涉及將工作負載分組到針對特定功能的不同 AWS 帳戶中，例如數據湖、機器學習平台和運營。實施 AWS Control Tower 可幫助自動化帳戶創建並建立治理最佳實踐。

關鍵組件包括設置 AWS 安全參考架構以進行全面的安全管理，以及利用 AWS Service Catalog 有效擴展機器學習資源。這種結構化的方法不僅確保資源的安全性，還促進快速創新，使團隊能夠專注於開發有效的機器學習解決方案，同時遵循合規要求。

這項指導是朝向在雲端建立良好治理的機器學習環境的重要基礎步驟。

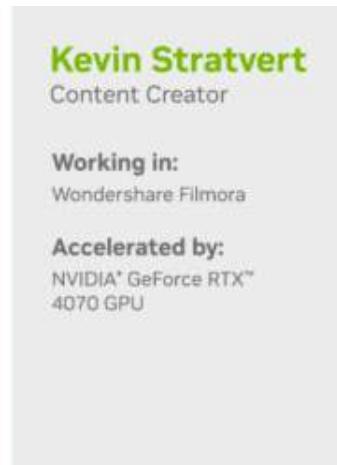


# Wondershare Filmora 引入 NVIDIA RTX Video HDR 支援

Wondershare Filmora | NVIDIA RTX Video HDR | AI功能 | 影片剪輯 | HDR10 | 直播軟體

Twitch增強廣播

2024-07-16



## Wondershare Filmora 引入 NVIDIA RTX Video HDR 支援

Wondershare Filmora 最近將 NVIDIA 的尖端 RTX Video HDR 技術整合進其影片剪輯軟體中。這次更新顯著提升了影片品質，將標準動態範圍的影片轉換為 HDR10 級別的內容，使創作者能夠製作出更清晰、更生動的視覺效果，特別適合高端顯示器。

除了 HDR 支援之外，Filmora 現在還利用 RTX 加速的 AI 功能，包括 Smart Edit 和 Smart Cutout，這些功能透過自動化任務（如生成逐字稿和移除多餘物件）來簡化剪輯流程。這些進步旨在提升內容創作者的工作效率，並增強創意的可能性。

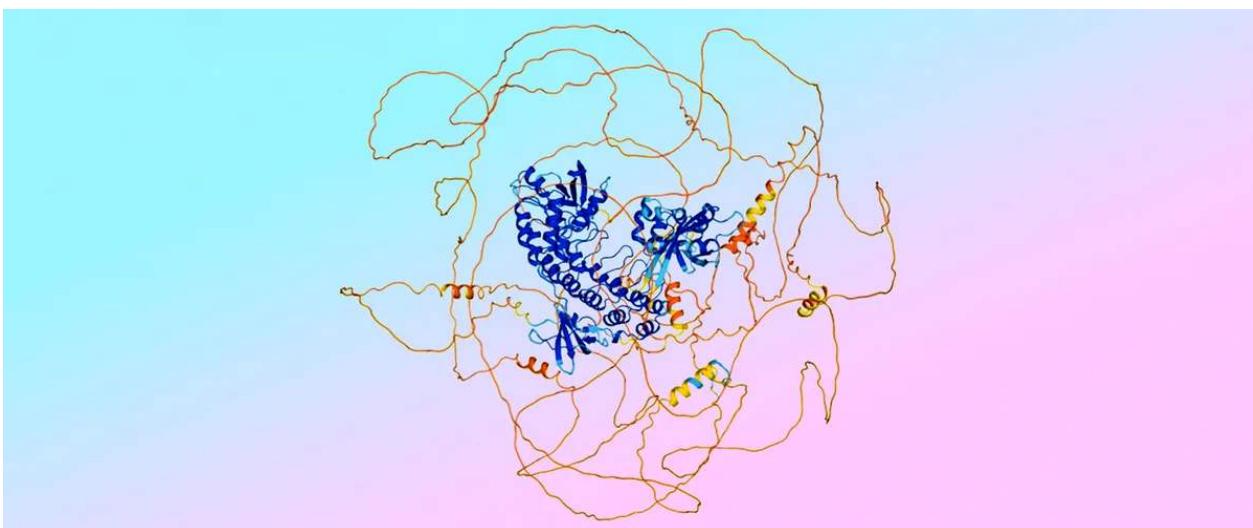
此外，像是 OBS Studio 和 XSplit Broadcaster 等直播軟體也採用了 Twitch 增強廣播功能，通過更好的編碼選項提升直播者的影片品質。總體而言，這些創新標誌著影片剪輯和廣播能力的一次重大突破，使高品質內容的創作對所有人來說更加可及。

[閱讀更多](#)

# AlphaFold 3：顛覆分子預測

**AlphaFold 3** 分子預測 蛋白質結構 DNA RNA 药物设计 生成模型 生物挑戰

2024-07-16



## AlphaFold 3：顛覆分子預測

於2024年5月發布的 AlphaFold 3 由 Google DeepMind 開發，標誌著分子預測的一項重大進展。此新模型在前作 AlphaFold 2 的基礎上發展而來，該前作幫助了超過 200 萬位研究人員進行蛋白質結構預測，這次的新模型擴展了其能力，不僅能分析蛋白質，還能分析生命中所有的分子，包括 DNA、RNA 和配體。

這一創新設計使 AlphaFold 3 能夠預測分子間的互動，這對於藥物設計等領域至關重要。自推出以來，AlphaFold Server 已能生成超過 100 萬個結構，允許用戶輕鬆輸入序列並獲得詳細預測，而不需要具備編程技能。

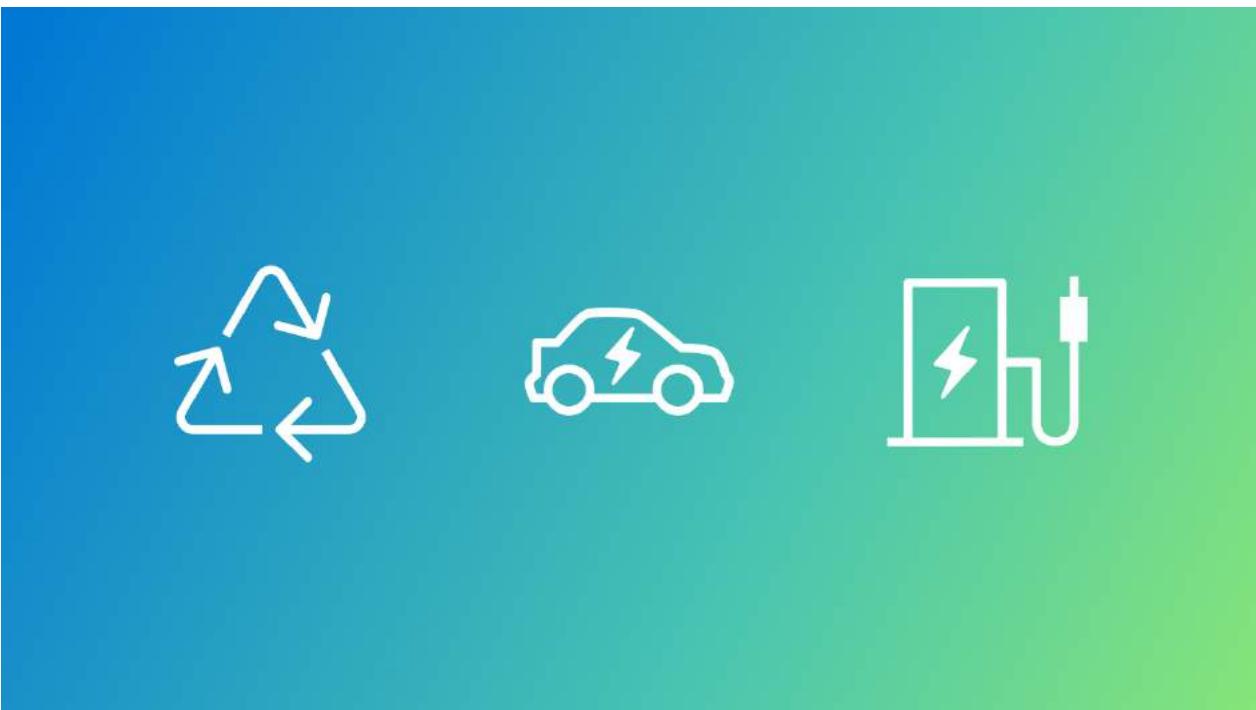
AlphaFold 3 在架構上有了顯著改進，利用生成擴散模型來簡化各種分子類型的處理。這一轉變提高了預測結構的準確性，包括那些具有無序區域的結構。憑藉這項技術，研究人員能更有效地應對複雜的生物挑戰，為基因組學和治療學的突破鋪平道路。

[閱讀更多](#)

# 微軟研究開發先進模型以預測電動車 電池衰退

電動車 | 電池衰退 | 機器學習 | 數據驅動 | 碳中和 | 資源再利用

2024-07-16



## 微軟研究開發先進模型以預測電動車電池衰退

在與日產汽車公司的合作下，微軟研究團隊推出了一個尖端的機器學習模型，顯著提升了預測電動車（EV）電池衰退的準確性。這項創新具有僅0.94%的驚人平均誤差率，旨在改善動力電池的回收過程，這對於電動車行業的可持續性至關重要。

傳統上，評估電池健康狀況一直是一項複雜的任務，需進行大量的實驗分析。然而，這個新模型利用數據驅動技術，最小化物理測試，使電池剩餘壽命的預測更加快速且可靠。透過實時準確分析電池的化學成分，這一發展預期將延長電池的使用壽命、減少碳排放，並促進更好的資源再利用。

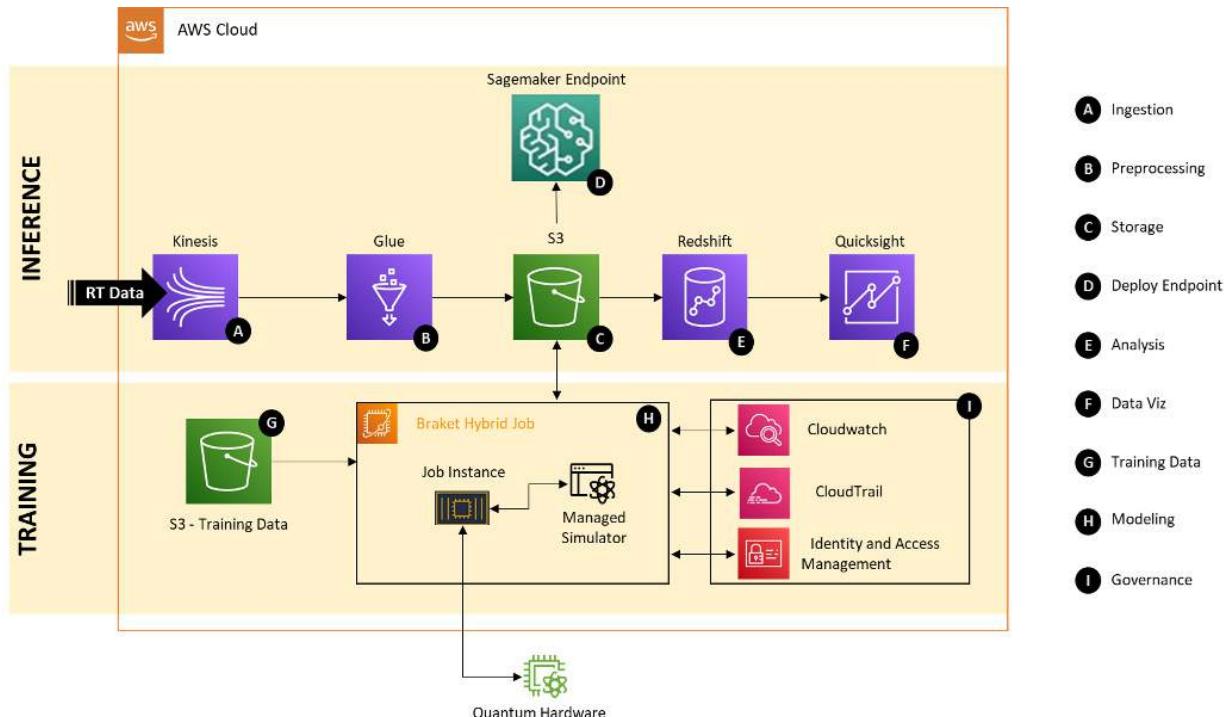
這項倡議標誌著朝著實現汽車行業碳中和的重要一步，並展示了人工智能在轉型電池管理實踐中的潛力。

[閱讀更多](#)

# 德勤意大利開發量子機器學習解決方案以檢測詐騙

德勤 量子機器學習 詐騙檢測 數位支付 機器學習 金融機構 風險管理

2024-07-17



## 德勤意大利開發量子機器學習解決方案以檢測詐騙

在數位商務的一大步進中，德勤意大利創造了一個尖端的解決方案，用以檢測數位支付中的詐騙，並整合了量子機器學習與 Amazon Braket。隨著網路交易的激增，對於效率高的詐騙檢測需求變得尤為重要。透過先進的機器學習演算法，該系統能夠即時分析大量的交易數據，迅速識別出詐騙行為。

這項創新涉及一個混合型量子神經網路，結合了傳統機器學習與量子演算法，提升了詐騙檢測的準確性。量子計算獨特的能力使得快速的模擬和優化成為可能，提供了相較於傳統方法的顯著優勢。儘管這項技術仍處於初期階段，但它有望徹底改變金融機構打擊詐騙的方式，為更安全的數位商務環境鋪平道路。隨著量子技術的成熟，其潛在的應用將深刻影響各種金融流程，增強風險管理和投資策略。

[閱讀更多](#)

# Amazon SageMaker 推出 Cohere Command R 調整模型

Amazon SageMaker | Cohere Command R | 調整模型 | 大型語言模型 | AI 技術

2024-07-17

```
message = "Classify the following text as either very negative, negative, neutral, positive or very positive: mr. deeds is , as comedy goes , very silly -- and in the best way."
result = co.chat(message=message)
print(result)

cohere.Chat {
    response_id: 96b71919-3d0d-4ff6-bc00-1be8a2caca0
    generation_id: 22b44470-bad6-4165-979d-0b553836d487
    text: Positive
    chat_history: [{"role": "USER", "message": "Classify the following text as either very negative, negative, neutral, positive or very positive: mr. deeds is , as comedy goes , very silly -- and in the best way."}, {"role": "CHATBOT", "message": "Positive"}]
    preamble: None
    finish_reason: COMPLETE
    token_count: None
    tool_calls: None
    citations: None
    documents: None
    search_results: None
    search_queries: None
    is_search_required: None
}
```

## Amazon SageMaker 推出 Cohere Command R 調整模型

Amazon 在其 SageMaker 平台上推出了 Cohere Command R 調整模型，增強企業對大型語言模型 (LLMs) 的能力。這款創新的模型專為高性能、企業級任務而設計，特別在對話互動和長文本處理上表現優異。Cohere Command R 可以處理多達 128,000 個標記，能夠高效地管理大量資訊。

調整過程使企業能夠針對特定應用自訂模型，提升各行各業（如金融、醫療和科技）的準確性超過 20%。透過使用檢索增強生成 (Retrieval Augmented Generation, RAG)，它能從外部來源檢索相關資訊，提高輸出質量。

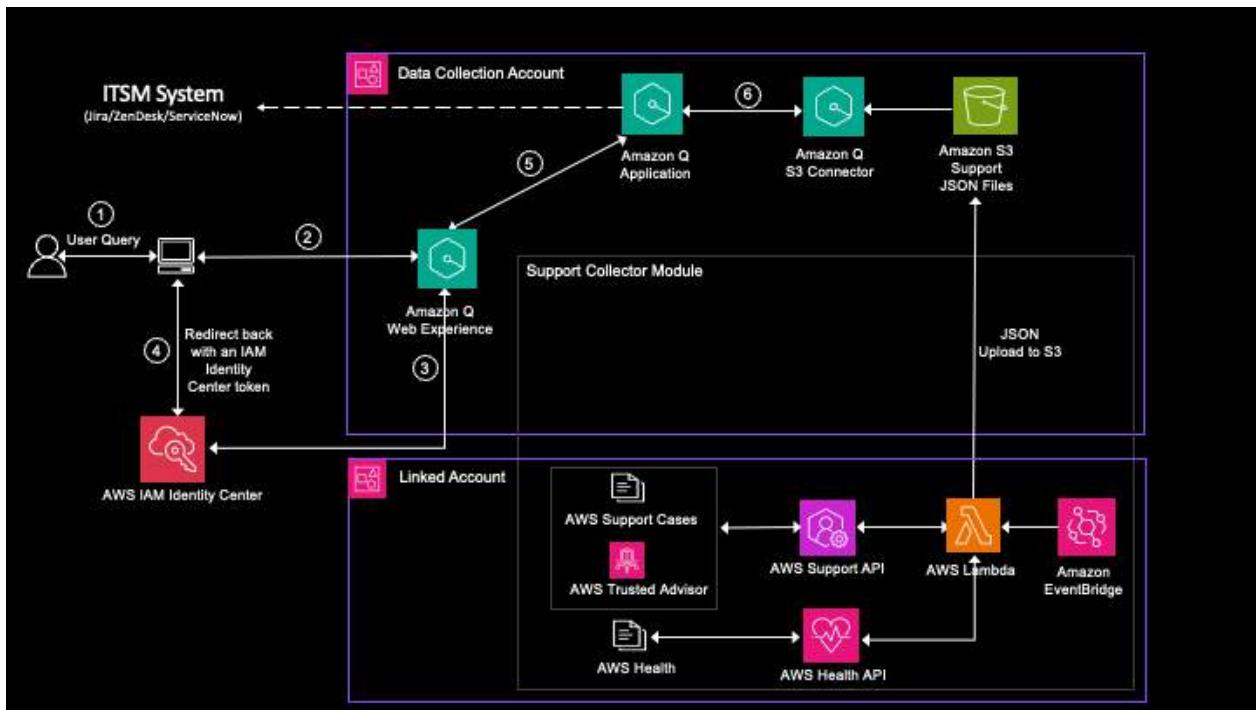
這項發展使企業能夠更有效地利用 AI 技術，推動增長和創新，同時維持運營效率。透過 Cohere Command R，企業能夠優化其 AI 工具，以滿足獨特需求，確保更好的效能和更低的資源消耗。

[閱讀更多](#)

# 透過 Amazon Q Business 解鎖營運洞察

生成式 AI | 聊天助手 | AWS | 營運效率 | 數據分析 | IT 服務管理

2024-07-17



## 透過 Amazon Q Business 解鎖營運洞察

Amazon 推出了 Amazon Q Business，這是一款基於生成式 AI 的聊天助手，旨在提升 Amazon Web Services (AWS) 使用者的營運效率。這項創新使得使用者能夠從他們的 AWS 環境中提取可行的洞察，並整合來自 AWS 支援案例、AWS Trusted Advisor 和 AWS Health 的數據。

使用 Amazon Q Business，使用者可以進行自然語言對話，詢問有關他們營運的問題，而不需要理解複雜的數據模型。這款 AI 助手會分析支援數據，識別模式並生成建議，從而簡化事件解決和洞察提取的過程。

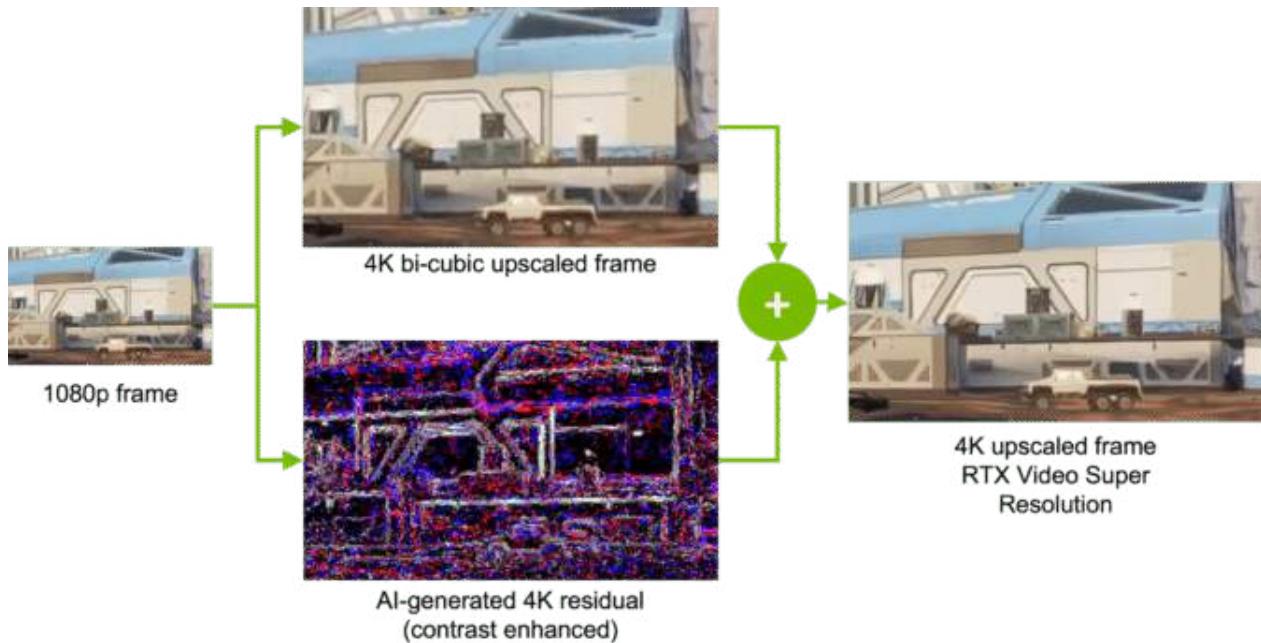
此外，它還可以無縫整合到像 ServiceNow 和 Jira 等流行的 IT 服務管理系統中，讓使用者能夠根據提供的洞察採取立即行動。Amazon Q Business 不僅旨在提升營運健康，還使組織能夠利用歷史支援數據預防重複出現的問題。

[閱讀更多](#)

# NVIDIA的AI驅動升頻技術徹底改變影片品質

**NVIDIA | AI | 升頻技術 | 影片品質 | RTX | 串流媒體 | 4K | AI HDR**

2024-07-17



## NVIDIA的AI驅動升頻技術徹底改變影片品質

NVIDIA推出了一項突破性的技術，透過AI驅動的升頻技術來提升影片品質，該技術基於RTX平台。隨著影片佔據了約80%的網路流量，許多串流媒體仍然限制在1080p解析度或更低。傳統的升頻技術僅通過簡單地放大像素來匹配顯示解析度，通常會導致畫面失真和細節喪失。

NVIDIA的解決方案，RTX Video Super Resolution，利用AI技術首先去除這些失真，然後智能地增強影像，產生更清晰的影片品質，可以升頻到4K。這項技術適用於各種瀏覽器和影片應用，即使是較低品質的串流媒體也能顯得格外清晰。

此外，AI HDR功能可以將標準內容轉換為高動態範圍，豐富顏色和細節。憑藉這些進展，NVIDIA的技術為串流體驗設立了新的標準，讓全球用戶的影片更加清晰和動態。

[閱讀更多](#)

# 微軟研究：2024年7月15日當週的創新

MG-TSD | Pre-gated MoE | LordNet | FXAM | 時間序列預測 | 混合專家架構 | 參數偏微分方程 | 預測分析

2024-07-17



## 微軟研究：2024年7月15日當週的創新

本週，微軟研究揭示了幾項突破性的技術進展。

首先，MG-TSD模型透過多層次指導擴散方法提升了時間序列預測的能力。這一創新方法在不需要額外數據的情況下，顯著提高了預測的準確性，在多項基準測試中實現了高達35.8%的相對改善。

接下來，Pre-gated MoE框架優化了混合專家架構，以減少記憶體消耗，同時保持高效能。這項進展對於在實際應用中擴展大型語言模型至關重要。

此外，LordNet提供了一種更高效解決參數偏微分方程的新方法，其速度比傳統求解器快多達40倍。

最後，FXAM模型增強了預測分析，整合各類型數據以提高準確性和效率。這個統一模型使用三階段訓練過程來優化學習。

這些創新標誌著微軟致力於推進各領域技術的承諾。

[閱讀更多](#)

# Meta不向歐盟推出多模態AI模型

Meta | AI模型 | 歐盟 | 監管 | Apple | Llama 3 | 多模態 | 合規

2024-07-18



Meta決定不向歐盟推出其即將推出的多模態AI模型，這一舉動與Apple最近決定將該地區排除在其Apple Intelligence的推廣之外一致。這個新模型能夠處理視頻、音頻、圖像和文本，原本計劃以開放授權的形式發布，但現在歐洲公司將無法訪問。這一限制可能會妨礙它們在全球AI市場的競爭力。

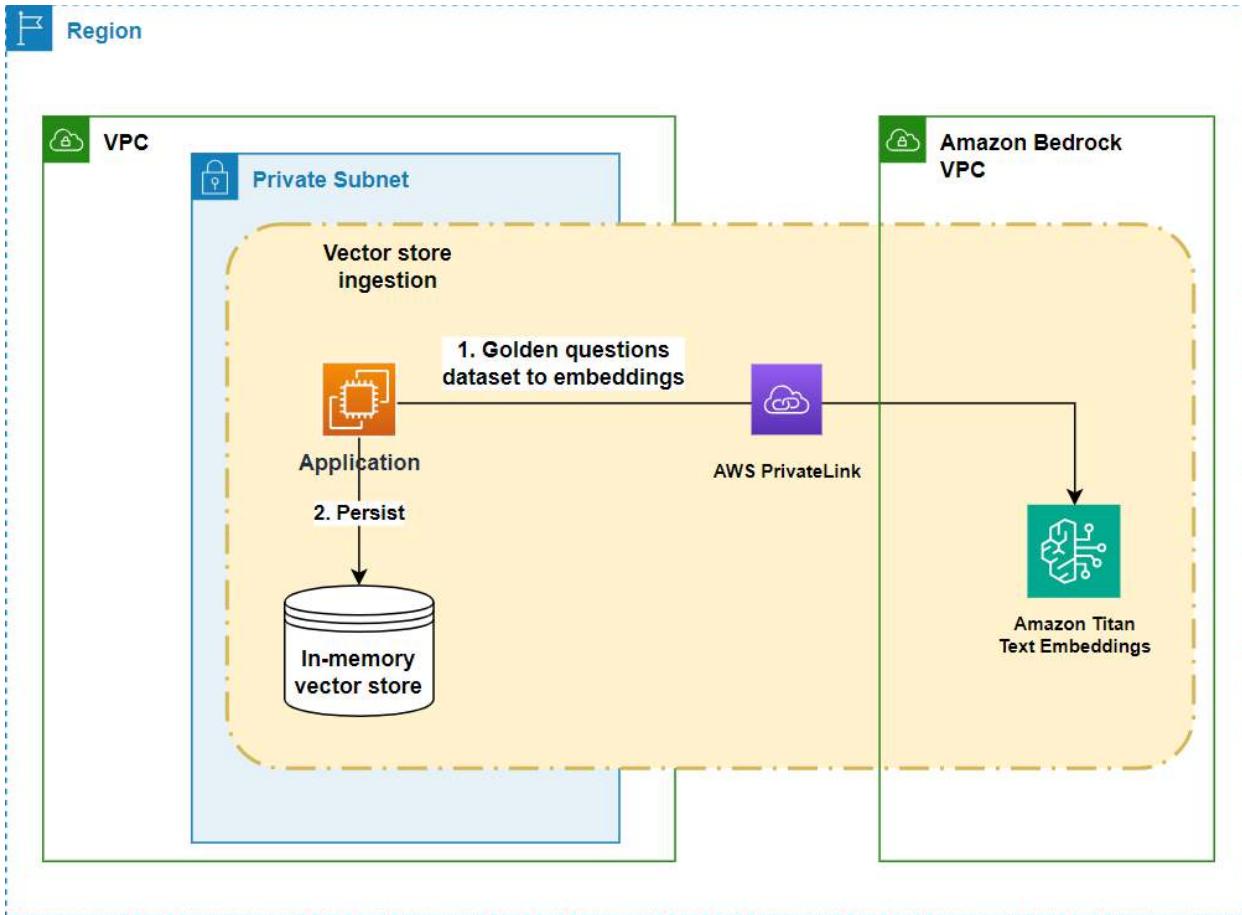
Meta的一位發言人指出，歐洲監管要求的「不可預測性」是這一決策的主要原因。儘管Llama 3模型的文本版本仍預計在歐盟推出，但完整的多模態功能將無法使用。這一情況突顯了科技巨頭在應對歐盟不斷演變的監管環境中的持續挑戰，特別是新的AI法案合規截止日期定於2026年8月。

[閱讀更多](#)

# Lili 發表由 Amazon Bedrock 驅動的 AccountantAI 聊天機器人

**AccountantAI** | 聊天機器人 | **Amazon Bedrock** | 小型企業 | 機器學習 | 財務指導

2024-07-18



Lili 發表由 Amazon Bedrock 驅動的 AccountantAI 聊天機器人

Lili 是一個針對小型企業的金融平台，近期推出了 AccountantAI 聊天機器人，這個聊天機器人利用了 Amazon Bedrock 和 Anthropic 的 Claude 3 模型。這款創新的聊天機器人旨在提供按需的會計建議，根據各個企業的需求量身打造，填補了小型企業主普遍面臨的空白，因為他們常常發現專業的會計服務難以獲得，或與緊急需求不相符。

AccountantAI 聊天機器人運用先進的機器學習技術來驗證用戶問題，並豐富相關的金融數據。它從一個全面的常見問題數據庫中提取信息，確保回應個性化且準確無誤。透過兩種不同的工作流程——攝取和檢索，這款聊天機器人增強了用戶互動並提供及時的財務指導。

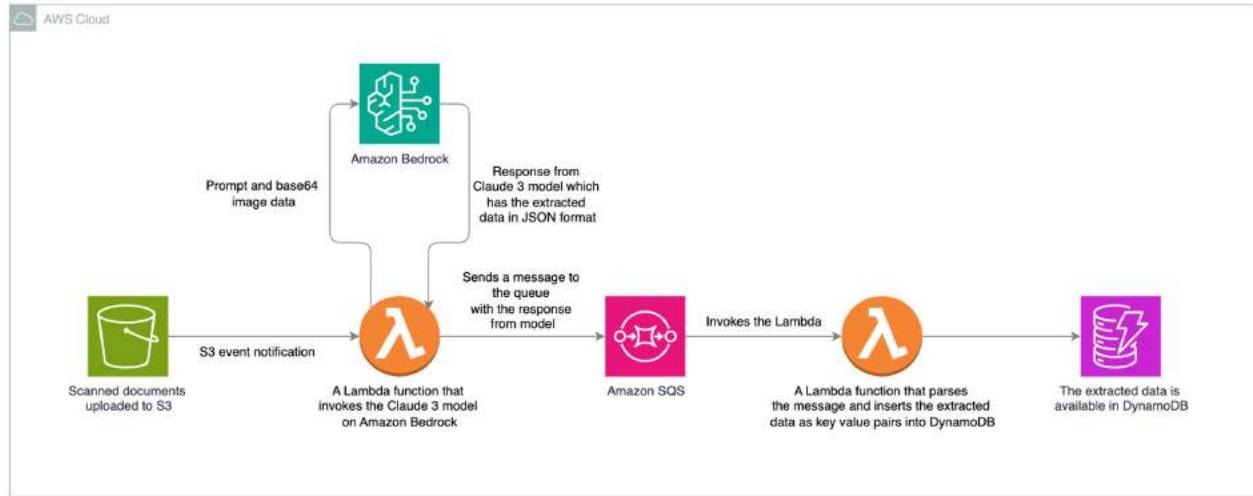
透過民主化高品質金融智能的獲取，Lili 的 AccountantAI 不僅簡化了商業運營，還使小企業主能夠專注於其核心業務，同時確保財務責任。這是讓金融專業知識變得更具可獲得性和可負擔性的重要一步。

[閱讀更多](#)

# 透過 Amazon Bedrock 和 Anthropic Claude 革新文件處理

**Amazon Bedrock** | **Anthropic Claude** | 文件處理 | 自動化 | 數據提取 | 生成式 AI | **AWS Lambda** | 運營效率

2024-07-18



## 透過 Amazon Bedrock 和 Anthropic Claude 革新文件處理

Amazon 在智能文件處理 (IDP) 的最新進展中走在創新最前端，結合 Amazon Bedrock 與 Anthropic Claude 3 Sonnet 模型。這項技術使得從掃描文件中無縫提取數據成為可能，提升了各行各業的運營效率。

透過整合生成式 AI 能力，組織可以自動化如文件分類、結構化數據提取以及從非結構化文本中檢索資訊等任務。這種提升的處理不僅減少了手動工作量，還顯著降低了錯誤率和成本。

當文件上傳至 Amazon S3 時，這一過程隨之啟動，觸發一系列 AWS Lambda 函數，調用 Claude 3 模型。這個模型擅長處理各種視覺格式和語言，使其對於不同的應用場景，如翻譯非英語文件，具備多樣性。

總體而言，這種強大的組合使企業能夠改變其文件工作流程，開啟新的效率和洞察，推動當今快速變化環境中的生產力和競爭力。

[閱讀更多](#)

# Amazon Bedrock 透過元數據過濾提升 數據檢索

**Amazon Bedrock** | 元數據過濾 | 數據檢索 | 檢索增強生成 | 食譜 | 健康與營養

2024-07-18

	RecipieId	Name	AuthorId	AuthorName	CookTime	PrepTime	TotalTime	DatePublished	Description	Images	...	Saturat
0	38	Low-Fat Berry Blue Frozen Dessert	1533	Dancer	PT24H	PT45M	PT24H45M	1999-08-09T21:46:00Z	Make and share this Low-Fat Berry Blue Frozen ...	c("https://img.sndimg.com/food/image/upload/w_... ...	...	...
1	39	Biryani	1567	elly9812	PT25M	PT4H	PT4H25M	1999-08-29T13:12:00Z	Make and share this Biryani recipe from Food.com.	c("https://img.sndimg.com/food/image/upload/w_... ...	...	...
2	40	Best Lemonade	1566	Stephen Little	PT5M	PT30M	PT35M	1999-09-05T19:52:00Z	This is from one of my first Good House Keepi...	c("https://img.sndimg.com/food/image/upload/w_... ...	...	...

3 rows × 28 columns

## Amazon Bedrock 透過元數據過濾提升數據檢索

Amazon Bedrock 新增了一項新功能，旨在提高處理表格數據集用戶的數據檢索準確性。這項稱為元數據過濾的進步，透過在數據檢索過程中利用元數據，使得查詢更加精確，從而提升系統回傳結果的相關性。

在此之前，未使用元數據時，用戶可能會檢索到不相關的信息，這可能導致成本增加和準確度下降。現在，透過元數據過濾，用戶可以指定條件，例如根據準備時間和營養成分過濾食譜，從而使結果更符合用戶的查詢。

此功能已整合到檢索增強生成 ( Retrieval Augmented Generation , RAG ) 工作流程中，使得組織能夠高效地從龐大的知識庫中提取相關且準確的回應。通過關注數據集之間的關聯，Amazon Bedrock 正在為不同行業的更智能數據互動鋪平道路，從烹飪應用到健康與營養。

[閱讀更多](#)

# Mistral AI 與 NVIDIA 合作推出 Mistral NeMo 12B 語言模型

**Mistral AI** **NVIDIA** **Mistral NeMo 12B** **語言模型** **人工智慧** **多語言任務** **聊天機器人** **程式編寫**  
**上下文長度** **推理速度** **NVIDIA DGX Cloud** **TensorRT-LLM**

2024-07-18



## Mistral AI 與 NVIDIA 合作推出 Mistral NeMo 12B 語言模型

在企業人工智慧技術的一個重要進展中，Mistral AI 與 NVIDIA 推出了 Mistral NeMo 12B，這是一個強大的語言模型，旨在便捷地進行自訂和部署，應用於各種場景，包括聊天機器人、多語言任務、程式編寫和摘要。這個模型得益於 Mistral AI 的專業知識以及 NVIDIA 的先進硬體和軟體，實現了令人印象深刻的準確性和靈活性。

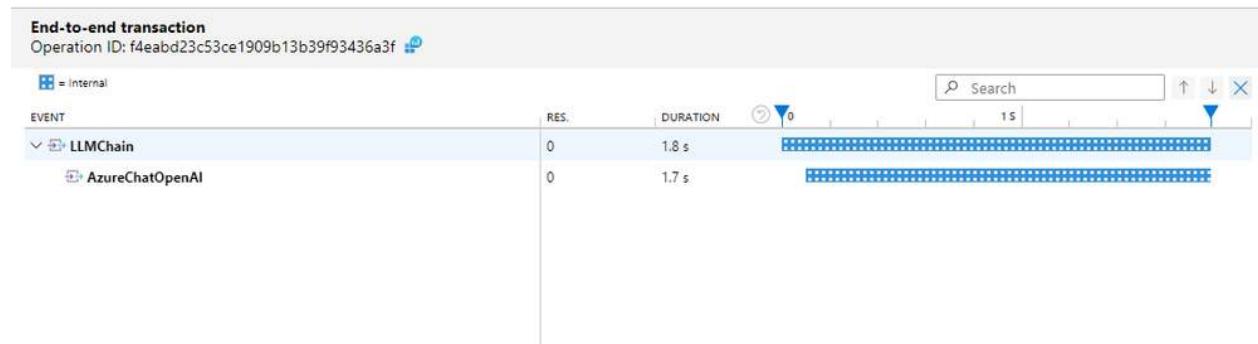
Mistral NeMo 可以處理大量資訊，擁有驚人的上下文長度達到 128K，這使得它適合處理複雜的任務，包括多輪對話和推理。它基於 NVIDIA DGX Cloud 平台構建，利用 NVIDIA TensorRT-LLM 提升推理速度和效率。Mistral NeMo 在開放許可下提供，設計可以在流行的 NVIDIA GPU 上運行，承諾縮短部署時間並為企業使用提供穩健的性能。這次合作彰顯了對促進人工智慧社群創新的承諾。

[閱讀更多](#)

# 在 Azure 上使用 OpenTelemetry 和 Application Insights 追蹤 LangChain 代碼

[OpenTelemetry](#) [Application Insights](#) [LangChain](#) [Azure](#) [監控](#) [AI](#) [性能](#)

2024-07-18



## 在 Azure 上使用 OpenTelemetry 和 Application Insights 追蹤 LangChain 代碼

隨著人工智慧 (AI) 和機器學習應用程式變得越來越複雜，確保其性能變得至關重要。追蹤在識別瓶頸和診斷問題中扮演了關鍵角色，為開發人員提供了應用程式內部運作的見解。

最近，推出了一種在 Azure 上使用 OpenTelemetry 和 Application Insights 追蹤 LangChain 應用程式的方法。這種方法利用 OpenInference 進行自動化工具的整合，便於監控 LangChain 應用程式中複雜操作的特性——例如提示語言模型和與 API 互動。

該設置包括安裝相關套件、配置 Azure Monitor Exporter，並將追蹤功能無縫整合進 LangChain 應用程式中。這使得即時監控和故障排除成為可能，從而提高整體應用程式的性能。用戶可以輕鬆地可視化操作並增強可觀察性，這對於行為不可預測的 AI 應用程式特別重要。

發現這種創新的追蹤方法如何改變您監控和優化 AI 應用程式的方式。

[閱讀更多](#)

# Mistral AI 與 NVIDIA 揭曉 12B NeMo 模型

[Mistral AI](#) [NVIDIA](#) [12B NeMo 模型](#) [上下文窗口](#) [開源](#) [多語言](#) [FP8](#) [Tekken tokeniser](#)

2024-07-19



## Mistral AI 與 NVIDIA 揭曉 12B NeMo 模型

Mistral AI 與 NVIDIA 合作推出了一款開創性的 12 億參數模型，名為 NeMo。這款創新的模型擁有令人印象深刻的上下文窗口，能夠處理多達 128,000 個標記，提升其在推理、世界知識和編碼任務中的表現。

NeMo 的設計旨在使用便捷，作為之前 Mistral 7B 模型的直接替代品。它還採取開源方式，預訓練及指令調整的檢查點在 Apache 2.0 許可下提供，促進廣泛的研究與應用開發。

一個突出特點是其訓練過程中的量化意識，允許進行高效的 FP8 推斷而不損失質量。此外，NeMo 還推出了 Tekken tokeniser，這個工具在壓縮超過 100 種語言方面表現優異，相較於之前的模型

提高了多達 30% 的效率。這一進步使 NeMo 成為多語言人工智慧應用的多功能工具，適用於各個領域。

[閱讀更多](#)

# 台積電在 AI 需求激增中創下增長紀錄

台積電 | AI | 晶片需求 | 增長 | 半導體

2024-07-19



## 台積電在 AI 需求激增中創下增長紀錄

台灣積體電路製造公司 ( TSMC ) 最近調整了2024年的營收預測，預期將有中位數20%的增長，這主要是受到人工智慧 ( AI ) 應用對晶片需求激增的推動。這一樂觀的展望緊隨2024年第二季的盈利好於預期，淨利達到2478億新台幣 ( 76億美元 ) 。

儘管對於在美國潛在合資企業的猜測層出不窮，台積電仍然堅持其全球擴張策略，並在亞利桑那、日本和歐洲等地大幅投資新設施。台積電的執行長魏哲家指出，對於先進晶片的需求「非常緊張」，強調該公司在AI革命中作為蘋果和Nvidia等主要科技企業供應商的關鍵角色。隨著台積電透過計畫中的300億到320億美元的資本支出來提升產能，該公司準備好滿足全球對先進半導體技術日益增長的需求。

[閱讀更多](#)

# NVIDIA 推出自導的 AI 和資料科學職涯發展資源

**NVIDIA** | 人工智慧 | 職涯發展 | 培訓 | 網路研討會 | 機器人技術 | 深度學習 | **LinkedIn**

2024-07-19

The graphic is a promotional image for SIGGRAPH 2024, held in Denver from July 28 to August 1. It features two headshots side-by-side: Jensen Huang on the left and Mark Zuckerberg on the right. Both are smiling and looking directly at the camera. Above the heads, the SIGGRAPH logo is displayed with the text "SIGGRAPH 2024" and "DENVER 28 JUL - 1 AUG". Below the heads, the names and titles of the speakers are listed: "Jensen Huang and Mark Zuckerberg on AI Breakthroughs" and "July 29 | 4:00 p.m. MT". At the bottom, their names and companies are identified: "Jensen Huang NVIDIA" and "Mark Zuckerberg Meta".

## NVIDIA 推出自導的 AI 和資料科學職涯發展資源

因應人工智慧 (AI) 領域對人才日益增長的需求，NVIDIA 推出了系列自我主導的培訓計畫，旨在為個人提供必要的技能以迎接這個動態領域的職業挑戰。他們近期舉辦的網路研討會「加速您在 AI 職涯的必備訓練與建議」吸引了超過 1,800 名參與者，並邀請了業界專家分享如何利用 NVIDIA 的資源來促進專業成長。

參與者了解了利用可用工具和網絡的重要性，例如免費的軟體開發工具包和專門針對機器人技術及 CUDA 的課程。NVIDIA 的 AI 學習基礎知識和深度學習學院提供了廣泛的資源，包括生成式 AI 認證和協作專案。

關鍵見解強調了在 LinkedIn 等平台上建立人脈和分享個人經歷的價值，使個人在追求成功的 AI 職涯中能與導師和同儕建立聯繫。

[閱讀更多](#)

# NVIDIA的超級電腦推動量子計算研究

NVIDIA | 超級電腦 | 量子計算 | 量子退火 | GPU計算 | 優化問題 | 相變化 | 伊辛自旋玻璃 | 密碼學

2024-07-19

The graphic is a promotional image for SIGGRAPH 2024, held in Denver from July 28 to August 1. It features two headshots side-by-side: Jensen Huang on the left and Mark Zuckerberg on the right. Both are smiling and looking directly at the camera. Above the heads, the text reads "Jensen Huang and Mark Zuckerberg on AI Breakthroughs". Below the heads, the text reads "July 29 | 4:00 p.m. MT". In the top right corner, the SIGGRAPH logo is displayed with the text "SIGGRAPH 2024" and "DENVER+ 28 JUL - 1 AUG".

Jensen Huang  
NVIDIA

Mark Zuckerberg  
Meta

最近發表在《Nature》的研究強調了通過使用NVIDIA驅動的超級電腦在量子計算方面取得的進展。這項研究由諾貝爾獎得主Giorgio Parisi領導，專注於量子退火（quantum annealing）——這是一種旨在解決傳統電腦難以處理的複雜優化問題的方法。

研究者利用超過200萬小時的GPU計算時間，模擬量子退火器的行為，這是一種能解決特定優化挑戰的量子電腦類型。與使用二進制值處理信息的經典電腦不同，量子電腦則利用量子位元（qubits），這使得創新處理方法成為可能。

該研究檢視了伊辛自旋玻璃（Ising spin glass）中的相變化，這是一種無序的磁性材料，對理解量子系統至關重要。透過揭示這些系統的行為，研究人員旨在增強能解決物流和密碼學等領域中難題的算法。這項開創性的工作標誌著量子計算實際應用邁出了重要的一步。



# NVIDIA 團隊在 KDD Cup 2024 中以創 新 AI 解決方案大放異彩

NVIDIA | KDD Cup 2024 | AI | 生成式 AI | Qwen2-72B | QLoRA | 電子商務

2024-07-22

Participants	Submission Trend	Understanding Shopping Concepts Weight 1.0	Shopping Knowledge Reasoning Weight 1.0	User Behavior Alignment Weight 1.0	Multi-Lingual Abilities Weight 1.0	All-around Weight 1.0
01 Team_NVIDIA	↑↑	🏆	🏆	🏆	🏆	🏆
02 AML666	↓↓	2	3	3	3	2
03 shimmering_as...	↓↓	3	4	5	2	4

NVIDIA 團隊在 KDD Cup 2024 中以創新 AI 解決方案大放異彩

在亞馬遜 KDD Cup 2024 這場享有盛譽的資料科學競賽中，NVIDIA 團隊以卓越的表現奪得五個賽道的第一名。團隊在生成式 AI 領域的專業知識，在如文本生成、命名實體識別和檢索等類別中展現無遺，這次的挑戰主題為「面向 LLM 的多任務線上購物挑戰」。

面對有限的訓練數據集，團隊創造性地生成了約 500,000 個問題，以微調他們新推出的 Qwen2-72B 模型。利用八個 NVIDIA A100 Tensor Core GPU，他們採用了 QLoRA（量化低秩適應）進行高效的模型訓練和推理。

這種創新方法不僅讓團隊能夠有效適應快速變化的電子商務環境，還使他們超越了競爭對手，連續第二年為 NVIDIA 帶來了大獲全勝。他們的獲勝方法論的詳細論文將於 KDD 2024 在巴塞隆納展示。

[閱讀更多](#)

# 微軟在 ICML 2024 展示機器學習創新

微軟 | ICML 2024 | 機器學習 | NaturalSpeech 3 | 文本轉語音 | 分解擴散模型 | 零樣本語音合成  
CompeteAI | 大型語言模型 | PRISE | 機器人技術 | 決策效率 | 醫療 | 氣候科學

2024-07-22



## 微軟在 ICML 2024 展示機器學習創新

在國際機器學習會議 (ICML 2024) 上，微軟揭示了機器學習的重大進展，這些進展有望改變各個領域。其中一個亮點是 NaturalSpeech 3，這是一個開創性的文本轉語音系統，利用新型的分解擴散模型進行零樣本語音合成。這項技術通過有效建模複雜的語音模式，提升了機器之間的溝通能力。

此外，研究人員還介紹了 CompeteAI，該項目探討大型語言模型代理如何通過競爭環境加速社會科學研究。另一個值得注意的創新是 PRISE，這是一種在機器人技術中學習時間行動抽象的方法，能提高決策效率和訓練數據的使用效果。

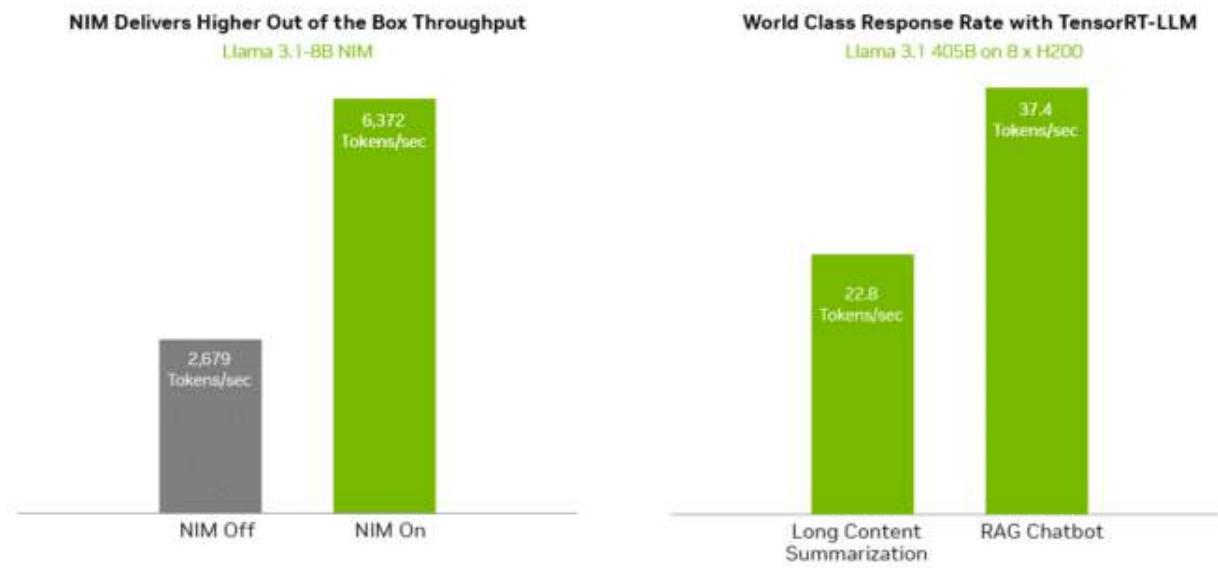
這些貢獻彰顯了微軟致力於利用機器學習，不僅用於自動化，還用於解決醫療和氣候科學等關鍵挑戰。隨著機器學習的不斷發展，這些創新為可以惠及整個社會的實際應用開闢了新的途徑。

[閱讀更多](#)

# NVIDIA AI Foundry：為企業量身打造 生成式 AI

**NVIDIA AI Foundry** | 生成式 AI | 客製化模型 | **DGX Cloud** | **NVIDIA NeMo** | 企業創新 | 基礎設施

2024-07-23



Llama3.1-BB-instruct, 1x H100 300MiB input and output token length: 1,000. Concurrent client requests: 200. NIM ON: BF16, TFTT ~1s, ITL ~20ms. NIM OFF: BF16, TFTT ~4s, ITL ~65ms.

Llama3.1-409B-instruct, 8 x H200 8XM. RAG Chatbot: 2,048 token input, 128 token output, 37.4 tokens per second, FPR, batch size 1. Long Content Summarization: 120,000 token input, 2,048 token output, 22.8 tokens per second, FPR, batch size 1.

## NVIDIA AI Foundry：為企業量身打造生成式 AI

NVIDIA AI Foundry 正在幫助企業創造符合其獨特產業需求的客製化生成式 AI 模型。透過加速計算與先進軟體工具的結合，企業可以開發並部署增強其 AI 施行的模型。就像半導體晶圓廠生產晶片一樣，NVIDIA AI Foundry 提供企業所需的基礎設施，讓他們打造自己的 AI 解決方案，並利用如 DGX Cloud 和 NVIDIA NeMo 軟體等資源。

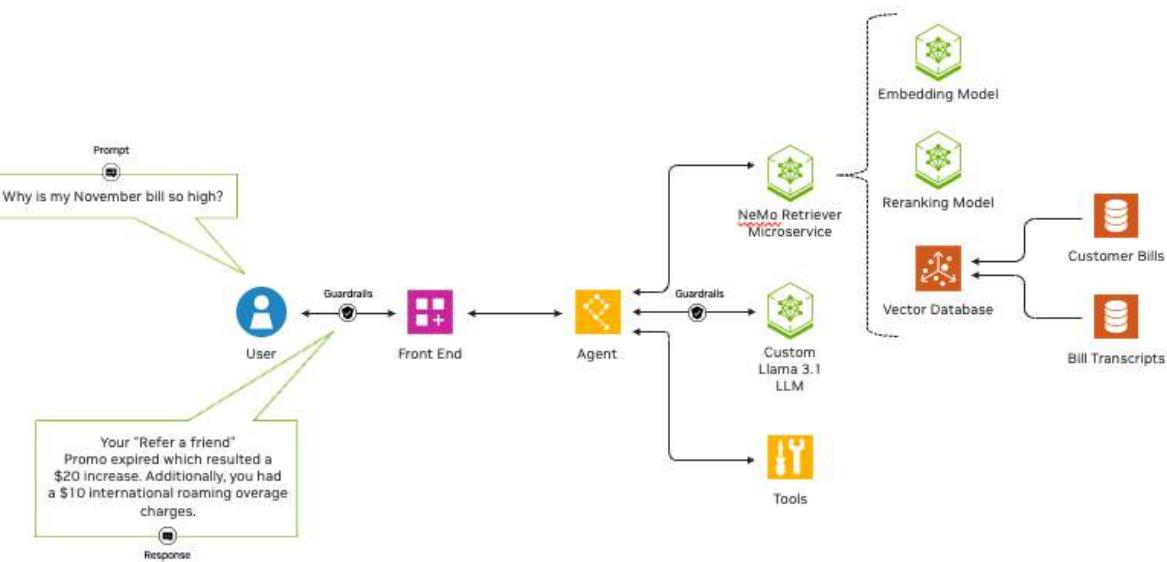
在 Capital One 和現代汽車等業界領導者的支持下，企業正利用客製化模型來獲得競爭優勢。該平台促進各種模型的客製化，包括全新的 Llama 3.1 系列，確保這些模型與特定產業需求緊密對接。此外，NVIDIA 的專家指導和廣泛的夥伴生態系統協助企業將這些模型無縫整合到現有的工作流程中，推動創新與運營效率。

[閱讀更多](#)

# NVIDIA 推出 NeMo Retriever 微服務以提升 AI 效率

NVIDIA | NeMo Retriever | AI | 大型語言模型 | 信息檢索

2024-07-23



## NVIDIA 推出 NeMo Retriever 微服務以提升 AI 效率

NVIDIA 發布了其全新的 NeMo Retriever NIM 推理微服務，旨在提升大型語言模型 (LLMs) 在企業應用中的準確性和速度。這些微服務使組織能夠高效地訪問和利用專有數據，改善像聊天機器人和安全分析工具等 AI 應用的性能。

NeMo Retriever 包含兩種類型的模型：將多樣化數據轉換為數值向量以便快速檢索的嵌入模型，以及根據相關性精煉這些數據的重新排序模型。這兩者結合起來顯著提升了信息檢索的準確性，與其他系統相比，錯誤回答減少了 30%。

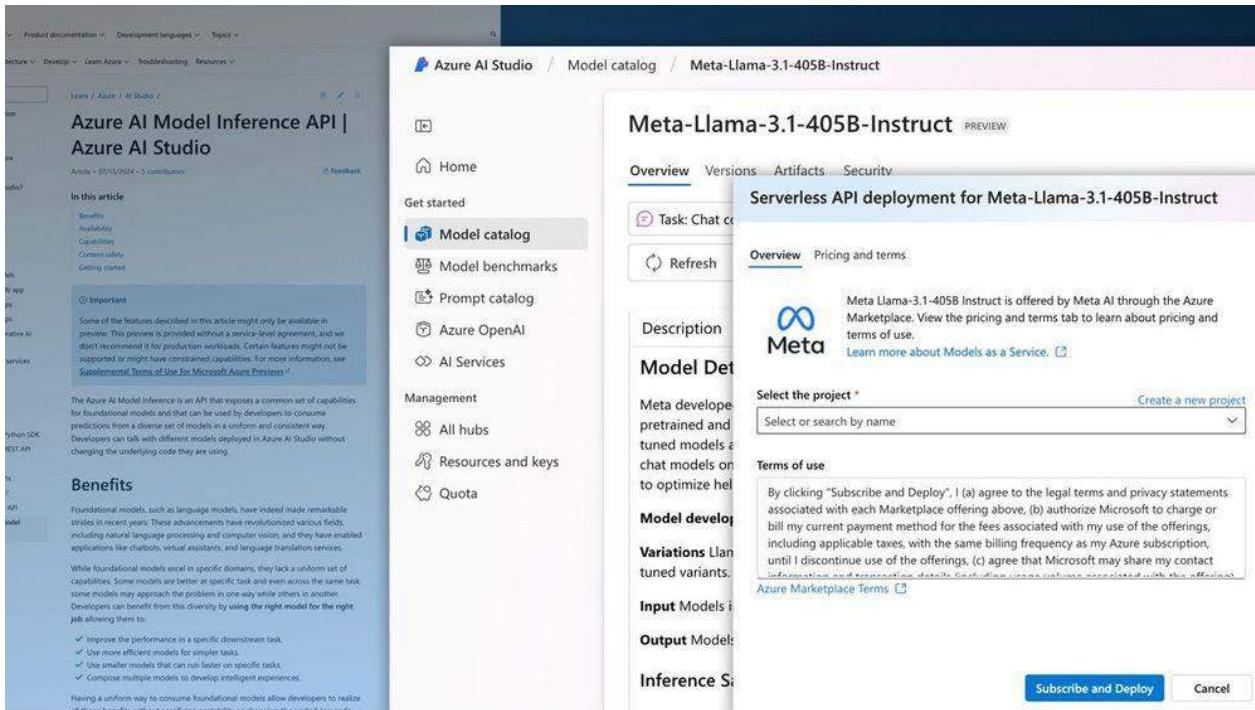
這些微服務現在已與多個平台整合，包括 DataStax 和 Cohesity，使企業能夠無縫地增強其數據驅動的 AI 解決方案。這項創新為各行各業的更智能、更具反應能力的 AI 技術鋪平了道路，促進了更有效的客戶互動和數據分析。

[閱讀更多](#)

# Meta 在 Azure AI 上推出 Llama 3.1 模型：徹底改變 AI 能力

**Llama 3.1 | AI 模型 | Azure AI | 數據生成 | 客戶支持 | 醫療保健 | 法律服務 | 教育 | 金融**

2024-07-23



## Meta 在 Azure AI 上推出 Llama 3.1 模型：徹底改變 AI 能力

Meta 已經推出其下一代 AI 模型 Llama 3.1 405B，現在可以通過 Azure AI 的 Models-as-a-Service 來訪問。這個模型及其較小的版本—Llama 3.1 8B 和 70B—能夠讓開發者快速測試和實施，使用像是 Azure AI prompt flow 和 OpenAI 等熱門 AI 工具。

Llama 3.1 405B 以其合成數據生成和模型蒸餾能力脫穎而出，讓較大的模型可以作為“小老師”來指導較小且更專業的“學生”模型。這種方法對於在客戶支持、醫療保健、法律服務、教育和金融等各行各業量身定製 AI 解決方案至關重要。

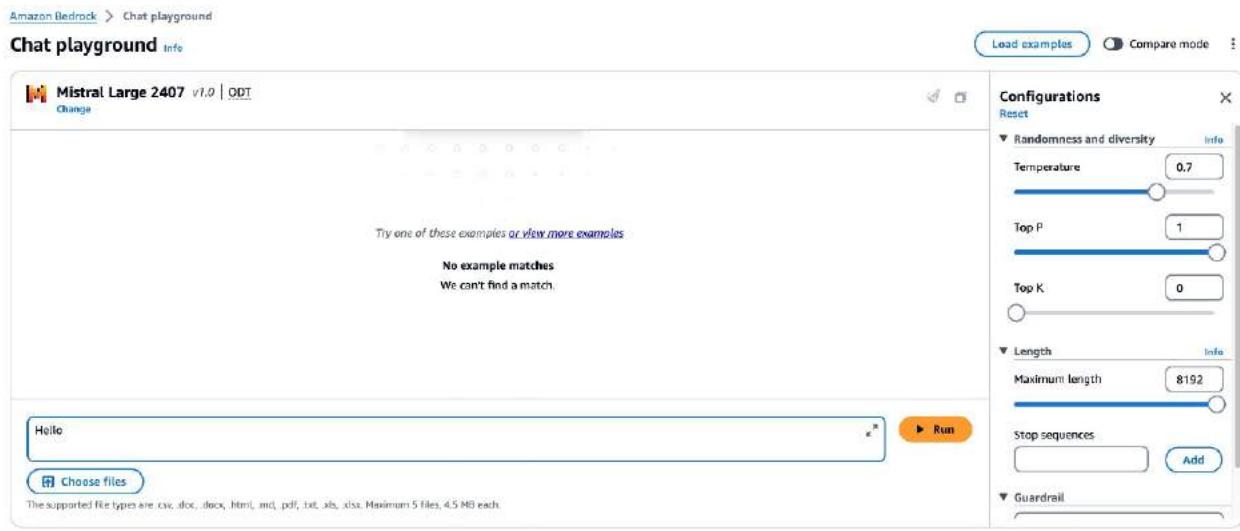
增強的功能包括支持長上下文長度和多語言應用，使得 Llama 3.1 模型適用於各種用例，同時保持效率和性能。通過 Azure AI，開發者可以安全有效地部署這些模型，推動各行各業的創新。

[閱讀更多](#)

# Mistral Large 2 在 Amazon Bedrock 上推出

**Mistral Large 2** | 大型語言模型 | 多語言能力 | 編碼 | 開發者

2024-07-24



## Mistral Large 2 在 Amazon Bedrock 上推出

Mistral AI 推出了其最新的基礎模型 Mistral Large 2，現在可在 Amazon Bedrock 上使用。這款先進的大型語言模型 (LLM) 在多語言能力、推理、數學和編碼任務等方面帶來了顯著的提升。Mistral Large 2 支援數十種語言，包括英語、西班牙語、中文和印地語，使其在全球應用中具有多樣性。

一個顯著的特點是其增強的上下文窗口，達到 128,000 個標記，能夠處理長文本或複雜的代碼，而不會失去上下文。此外，它可以以原生 JSON 格式輸出回應，簡化了開發者的數據整合。

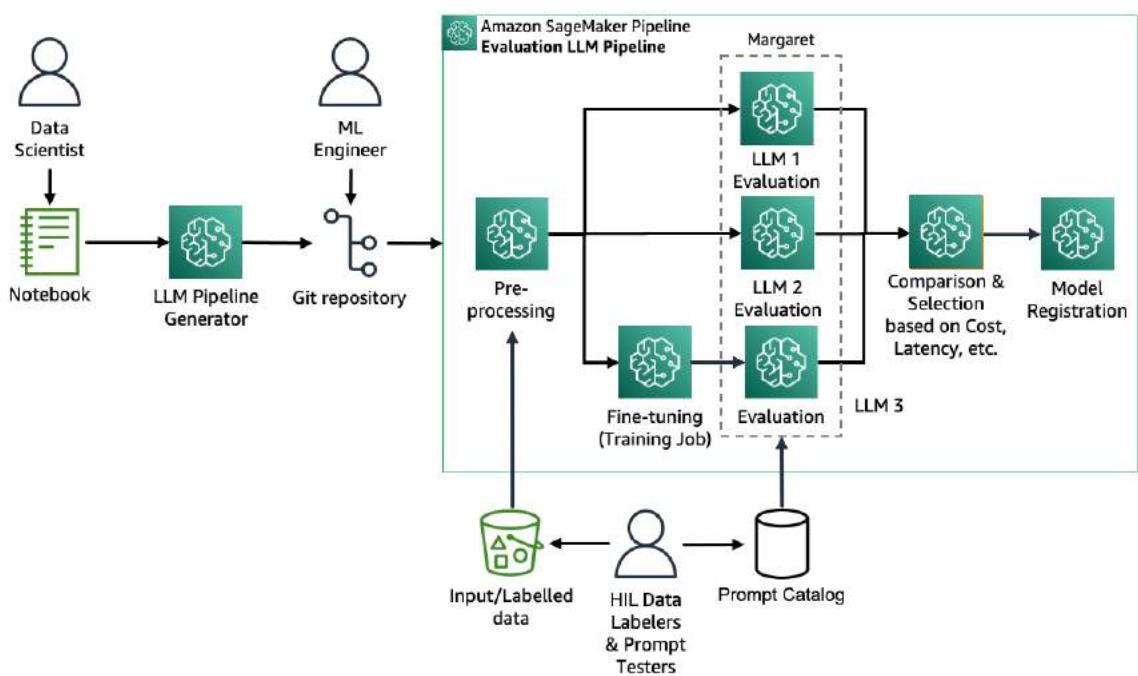
Mistral Large 2 的設計旨在降低信息檢索中的不準確性，並在必要時承認其限制。它接受了超過 80 種程式語言的訓練，可以協助開發者進行代碼生成、除錯以及許多與編碼相關的任務，為各個領域的創新提供強大的工具。

[閱讀更多](#)

# 利用 AWS SageMaker Pipelines 和 MLflow 解鎖大型語言模型的客製化

AWS | SageMaker | MLflow | 大型語言模型 | 微調 | 自訂 | 自然語言處理 | 實驗追蹤 | 部署

2024-07-24



## 利用 AWS SageMaker Pipelines 和 MLflow 解鎖大型語言模型的客製化

亞馬遜網路服務 (AWS) 推出了創新的工具，以簡化使用 Amazon SageMaker 和 MLflow 自訂大型語言模型 (LLMs)。隨著LLMs在自然語言處理方面的優越表現，其在特定任務中的有效性可能會有所不同。AWS 通過允許使用者透過提示工程和檢索增強生成 (Retrieval Augmented Generation, RAG) 等技術來進行 LLM 的微調來解決這個問題。

這個過程涉及選擇合適的預訓練基礎模型，並通過各種數據集和超參數進行迭代，以優化性能。通過整合 MLflow，AWS 簡化了實驗結果的追蹤、模型版本控制和部署。使用者可以有效地同時運行多個實驗，視覺化性能，並記錄相關的元數據。

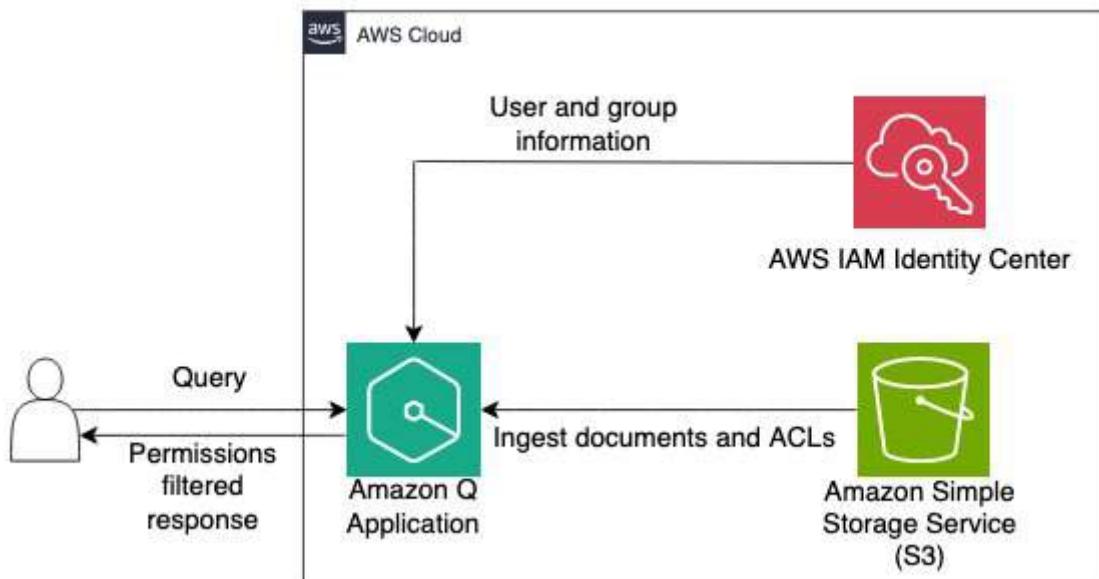
此外，Amazon SageMaker Pipelines 使這些微調過程的編排變得更加便捷，促進了大規模實驗的進行。這種組合讓資料科學家能夠識別最佳的模型並自信地進行部署，最終提升 LLM 在特定應用中的能力。

[閱讀更多](#)

# Amazon Q S3 連接器：用生成式 AI 解鎖洞察力

Amazon Q S3 連接器 生成式 AI 數據分析 ACL 信息檢索 數據安全

2024-07-24



## Amazon Q S3 連接器：用生成式 AI 解鎖洞察力

Amazon 推出了 Amazon Q S3 連接器，這是一款尖端工具，旨在利用生成式人工智慧 (AI) 的力量進行企業數據分析。這個完全管理的助手可以與存儲在 Amazon Simple Storage Service (S3) 中的信息互動，使用戶能夠提問、接收摘要以及根據他們公司數據生成內容。

Amazon Q S3 連接器整合了各種數據來源，使組織能夠安全地索引和檢索相關信息。通過使用存取控制列表 (ACL)，它確保用戶僅能訪問他們被授權查看的數據，從而實現精確和安全的查詢。

利用這項技術，用戶可以迅速發現來自各種內容類型的關鍵洞察，從博客到用戶指南以及案例研究。這項創新不僅提升了信息檢索的效率，還賦予企業以信心做出明智的決策，最終解鎖其數據庫的全部潛力。

[閱讀更多](#)

# Salesforce 利用 Amazon SageMaker 加強程式碼生成

[Salesforce](#) [Amazon SageMaker](#) 程式碼生成 LLM Einstein 平台 AI 解決方案

2024-07-24

EINSTEIN SIDEBAR

I want to create an Apex class called ContactSum. Create a method named getSumOfContactRecords that retrieves the number of contact records linked to a specific Account record. The method should accept accountId as the parameter. Follow security best practices and ensure the code runs in the user mode

Clear Ask

Apex Copy code

```
public with sharing class ContactSum {  
    public static Integer getSumOfContactRecords(Id acc  
        if (accountId == null) {  
            return 0;  
        }  
        List<Contact> contacts = [SELECT Id FROM Contac  
        return contacts.size();  
    }  
}
```

Upvote Downvote

Salesforce 利用 Amazon SageMaker 加強程式碼生成

Salesforce 最近升級了其 Einstein 平台的程式碼生成能力，透過運用 Amazon SageMaker 來實現。這項增強主要針對提升 Salesforce 的 CodeGen 的性能，這是一個開源的大型語言模型（LLM），設計用來將自然語言轉換成像 Python 這樣的程式語言。

透過使用 SageMaker，Salesforce 改善了延遲和吞吐量，實現了每分鐘處理的程式碼請求數量驚人地提高了 6,500%。SageMaker 促成這項進展的關鍵功能包括專門為大型模型推理設計的深度學習容器、動態批量策略以優化 GPU 資源的使用，以及高效路由以平衡多個實例之間的工作負載。

這些增強使得使用 Salesforce 的 EinsteinGPT 的開發者能夠即時接收程式碼建議，更有效地簡化其編碼任務，最終提高 Salesforce 生態系統內的生產力和使用者體驗。這次合作突顯了整合先進 AI 解決方案以支持業務需求的潛力。

[閱讀更多](#)

# 探索使用 Adobe 和 NVIDIA RTX 的 AI 協助創意

Adobe | NVIDIA | RTX | AI | 創意 | 生成式 AI | Photoshop | Premiere Pro | After Effects | 音質 | 視覺效果

2024-07-24



## 探索使用 Adobe 和 NVIDIA RTX 的 AI 協助創意

Adobe 與 NVIDIA 合作，透過 NVIDIA RTX 技術增強創意工作流程，提供 AI 幫助工具。用戶可以利用超過 100 種創新的功能，運用生成式 AI 的能力，簡化在各種 Adobe 應用程式中的創作過程。

Adobe Firefly 是一系列生成式 AI 模型，允許創作者根據描述性提示生成圖像和設計。例如，在 Adobe Photoshop 中，生成填充工具使用戶能夠輕鬆地向圖像添加內容，而參考圖像功能則幫助根據上傳的視覺資料來精練輸出。

在視頻編輯方面，Adobe Premiere Pro 的 AI 驅動增強語音工具能顯著提升音質，讓對話聽起來更加專業。與此同時，Adobe After Effects 引入了先進的物件隔離技術，實現無縫的視覺效果。

這些尖端工具不僅提升了生產力，還激發了創新的靈感，使得高級藝術創作變得前所未有的容易。

[閱讀更多](#)

# Mistral Large 2：AI 模型競爭的新選手

[Mistral Large 2](#) [AI 模型](#) [編程語言](#) [上下文窗口](#) [參數](#) [資源消耗](#) [信息生成](#) [研究許可證](#)

2024-07-25



## Mistral Large 2：AI 模型競爭的新選手

Mistral AI 最近推出了其 Mistral Large 2 (ML2) 模型，將自己定位為對抗來自業界巨頭如 OpenAI 和 Meta 的大型模型的有力競爭者。值得注意的是，ML2 支援 128,000 個 token 的上下文窗口，並在超過 80 種編程語言及多種人類語言中表現出色。

ML2 以 1230 億個參數高效設計，在消耗資源方面顯著低於更大型的對手，實現了驚人的性能。這種效率促進了更快的反應生成，對於實際應用至關重要。

Mistral 也針對常見的 AI 挑戰進行了改進，提升了 ML2 避免生成誤導性資訊的能力，並改善了其遵循指令的能力。這款模型在 Mistral 研究許可證下可供研究和非商業用途免費使用，成為大型語言模型領域的一項有前景的進展。

[閱讀更多](#)

# Amazon SageMaker 強化生成式 AI 模型的自動擴展

Amazon SageMaker | 生成式 AI | 自動擴展 | 大型語言模型 | 吞吐量 | 基礎設施成本 | 即時令牌串流

2024-07-25

## Amazon SageMaker 強化生成式 AI 模型的自動擴展

Amazon 最近宣布對 SageMaker 推論進行強化，推出更快速的生成式 AI 模型自動擴展功能。這項新能力使自動擴展能在不到一分鐘內做出反應，顯著減少在 AI 應用需求波動時的延遲。

由於生成式 AI 模型，特別是大型語言模型（LLMs），經常需要大量的處理時間，因此這種反應速度至關重要。這項強化採用了新的指標——`ConcurrentRequestsPerModel` 和 `ConcurrentRequestsPerCopy`—使得能夠精確監控同時請求的數量。這意味著當需求激增時，額外的資源會迅速分配，從而在控制成本的同時保持性能。

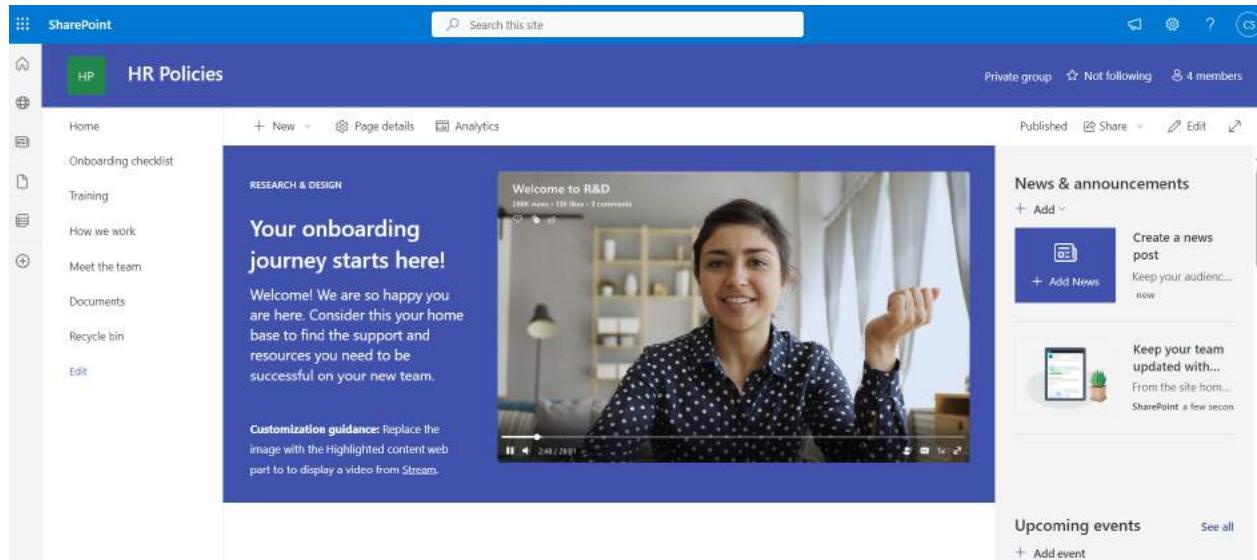
此外，SageMaker 的優化工具包能將吞吐量提高一倍，並將基礎設施成本降低約 50%。引入的即時令牌串流功能也增強了用戶體驗，通過最小化感知延遲，使得像對話 AI 這樣的解決方案更具回應性。這項創新不僅簡化了生成式 AI 模型的部署，還確保它們能高效運行，實時適應變化的工作負載。

[閱讀更多](#)

# 透過 Amazon Q Business 解鎖 SharePoint 洞察力

**Amazon Q Business | SharePoint | AI | 自然語言 | 數據整合 | 團隊協作 | 知識共享 | 生產力**

2024-07-25



## 透過 Amazon Q Business 解鎖 SharePoint 洞察力

Amazon 推出了一個強大的工具，名為 Amazon Q Business，旨在提升組織訪問和利用存儲在 SharePoint Online 中的數據的方式。這個 AI 驅動的助手允許用戶以自然語言提問，通過綜合來自不同 SharePoint 網站的信息，生成快速且準確的回應。

這項創新的核心是 SharePoint Online 連接器，能夠無縫整合來自不同系統的內容。這使得用戶可以加速研究、簡化內容創建，並自動化日常任務。例如，員工可以詢問人力資源政策或薪資細節，並立即獲得直接來自 SharePoint 的答案。

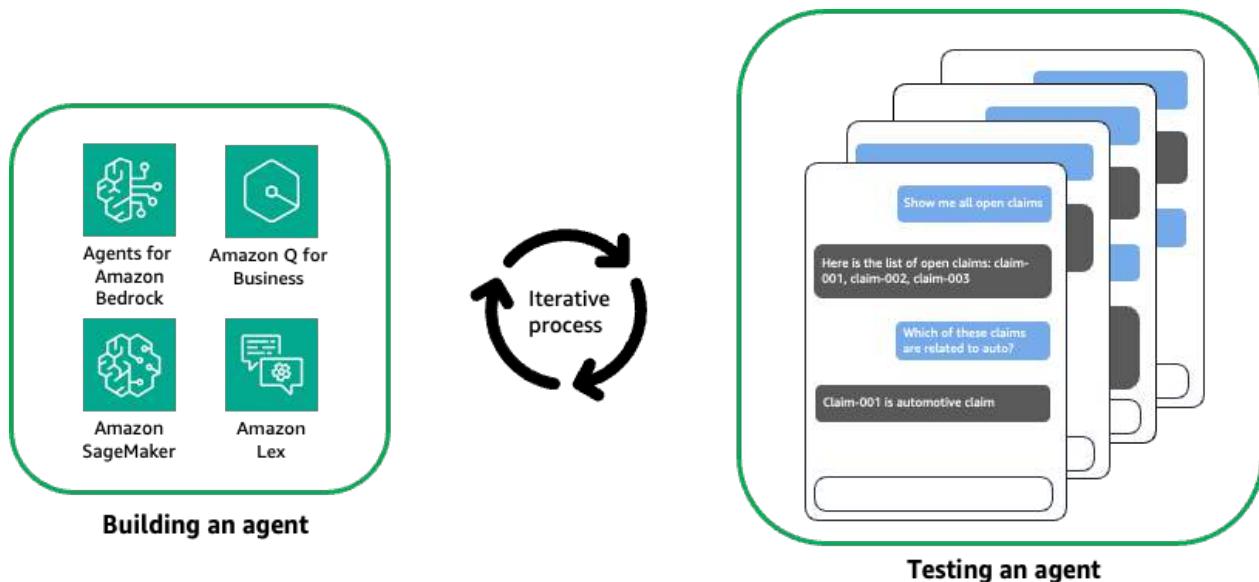
透過利用生成式 AI，Amazon Q Business 改變了團隊協作的方式，使知識共享和決策變得更高效。這項進步不僅賦能員工，還幫助組織最大化其 SharePoint 投資，推動生產力和創新。

[閱讀更多](#)

# Amazon Bedrock 推出對話式 AI 測試的代理評估

**Amazon Bedrock** | 對話式 AI | 代理評估 | 自動化測試 | CI/CD | 用戶體驗 | 客戶支援 | 資訊檢索

2024-07-25



## Amazon Bedrock 推出對話式 AI 測試的代理評估

Amazon 推出了代理評估 (Agent Evaluation)，這是一個開源解決方案，旨在提高大規模對話式 AI 代理的測試和驗證。隨著對話式 AI 的普及，確保用戶互動的可靠性和一致性至關重要。傳統的評估方法常常因這些 AI 系統的複雜性而無法滿足需求。

代理評估與 Amazon Bedrock 無縫整合，該平台提供高效能基礎模型的訪問。這個工具允許開發者創建全面的測試計畫，涵蓋多輪對話並實時驗證 AI 的回應。功能包括對話的協調、自動化測試整合到 CI/CD 管道中，以及詳細的性能洞察。

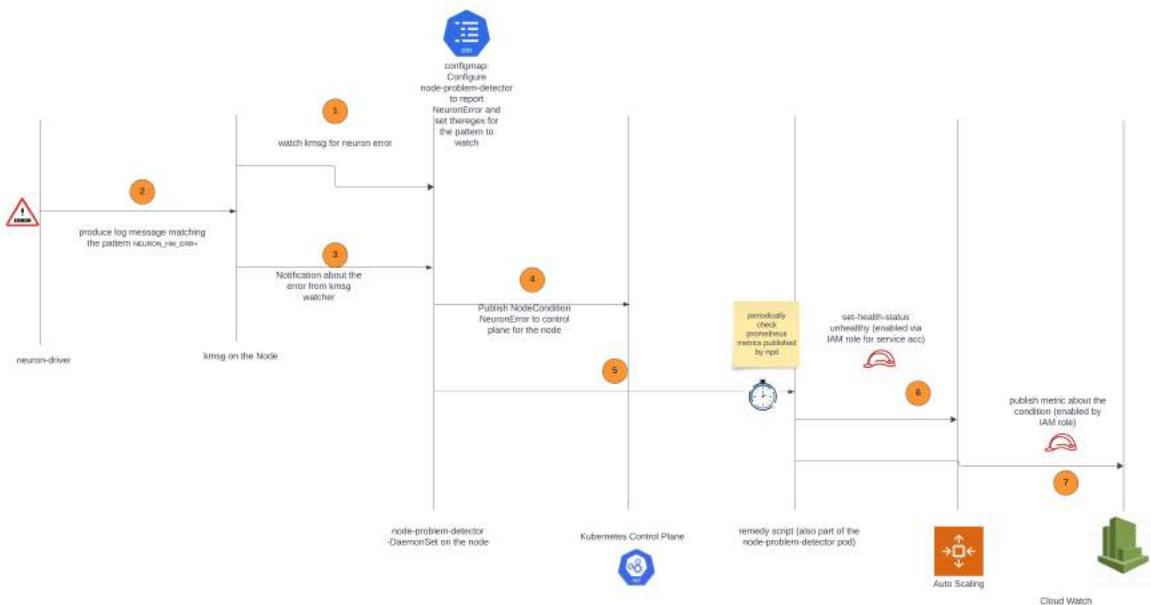
藉由自動化測試過程，組織可以顯著提高對話代理的可靠性，確保它們在到達用戶之前如預期般運作。這項創新有望精簡開發工作流程，並在各種應用中提升整體用戶體驗，從客戶支援到資訊檢索。

[閱讀更多](#)

# AWS 推出自動節點問題檢測與恢復功能，專為 EKS 叢集設計

**AWS | EKS | 節點問題檢測 | 自動恢復 | 機器學習 | AI**

2024-07-25



## AWS 推出自動節點問題檢測與恢復功能，專為 EKS 叢集設計

Amazon Web Services (AWS) 推出了一項創新的解決方案，旨在提升 Amazon Elastic Kubernetes Service (EKS) 叢集中機器學習訓練過程的可靠性。這個新的節點問題檢測器和恢復 DaemonSet 專門針對 AWS Trainium 和 AWS Inferentia 節點，提供主動的健康監測和自動恢復機制。

這個尖端工具持續監控 Neuron 裝置的日誌以檢查錯誤。在檢測到故障時，它會迅速將受影響的工作節點標記為不健康並啟動替換。這種自動化的回應最小化了停機時間，並降低了硬體故障可能造成的效果，確保訓練工作流程不間斷。

此解決方案適用於受管理和自我管理的節點群組，顯著提高了機器學習訓練環境的容錯能力，為 AI 和機器學習專案的更穩健及高效運行鋪平了道路。隨著這一發展，組織可以更專注於創新，而非處理硬體問題。

[閱讀更多](#)

# 微軟研究揭示 Trace : AI 優化的新時代

Trace | AI 優化 | 微軟研究 | 史丹佛大學 | 自我適應 | 動態適應 | 超參數調整 | 機器人控制

2024-07-25



## 微軟研究揭示 Trace : AI 優化的新時代

微軟研究與史丹佛大學合作推出了 Trace，這是一個創新的框架，旨在比以往更有效地優化 AI 系統。Trace 將 AI 系統視為計算圖，允許使用通用反向傳播方法進行端到端的優化。

Trace 的主要特點包括：

1. 動態適應：它能夠響應 AI 系統不斷演變的特性，根據各種輸入和反饋進行調整。
2. 多功能應用：Trace 可以優化 AI 系統中的不同參數，提升超參數調整和機器人控制等任務，通常能比傳統方法更快地達成結果。
3. 自我適應的代理：這個框架能夠建立從經驗中學習的 AI 代理，無需大量手動編碼。

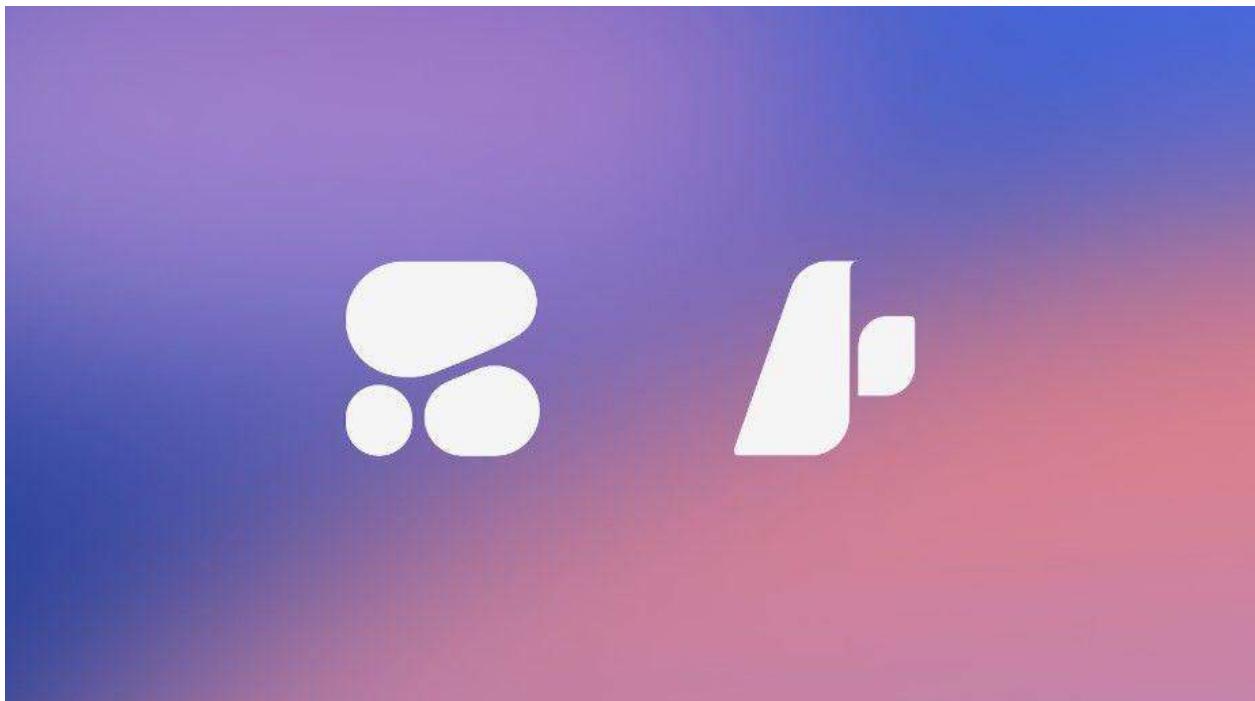
隨著 Trace 的實施，微軟旨在簡化複雜 AI 系統的開發，推動人工智慧在各種應用中所能達成的邊界。

[閱讀更多](#)

# Cohere Rerank 3 強化了 Azure AI 的搜尋功能

**Cohere** **Rerank 3** **Azure AI** 搜尋功能 **AI 模型** 語意重排 上下文長度 多語言支持 數位助理  
搜尋準確率

2024-07-25



## Cohere Rerank 3 強化了 Azure AI 的搜尋功能

Cohere 推出了 Rerank 3，這是一個先進的 AI 模型，現在已在 Azure AI 上提供，預計將徹底改變搜尋操作。該模型在語意重排方面表現出色，顯著提高了搜尋結果的相關性。Rerank 3 具備 4,000 字的上下文長度和支持超過 100 種語言的能力，能有效處理多方面及半結構化數據，非常適合各種應用，從 IT 服務管理到客戶支援。

開發者只需幾行程式碼，即可將 Rerank 3 無縫整合進現有系統，利用 Azure 強大的基礎設施來提升性能和成本效率。值得注意的是，像 Atomicwork 等公司報告顯示，利用 Rerank 3 在他們的數位助理中，搜尋準確率提高了 20% 以上。這項整合為企業在多樣化的場景中提升搜尋功能鋪平了道路，驅動了生產力並改善了用戶體驗。

[閱讀更多](#)

# Galileo 發佈生成式 AI 模型的幻覺指數

生成式 AI | 幻覺指數 | 大型語言模型 | 開源模型 | 評估

2024-07-29



## Galileo 發佈生成式 AI 模型的幻覺指數

在 AI 社群的一個重要進展中，Galileo 發佈了他們的新幻覺指數，對來自 OpenAI、Anthropic、Google 和 Meta 等頂尖公司的 22 款領先生成式 AI 大型語言模型（LLMs）進行評估。今年的指數引入了 11 款新模型，展示了開源和閉源 LLM 的快速演變。

該指數利用一種名為上下文遵循的專有指標，評估不同輸入長度下的輸出準確性。研究結果顯示，Anthropic 的 Claude 3.5 Sonnet 是整體最佳表現者，而 Google 的 Gemini 1.5 Flash 在成本效益方面表現優異。值得注意的是，Alibaba 的 Qwen2-72B-Instruct 脫穎而出，成為領先的開源模型。

該指數突顯了新興趨勢，例如開源和閉源模型之間的表現差距正在縮小，並建議較小的模型有時可以超越較大的模型。這項評估對於在複雜的生成式 AI 環境中尋找平衡成本和可靠性的企業來說，提供了一個關鍵資源。

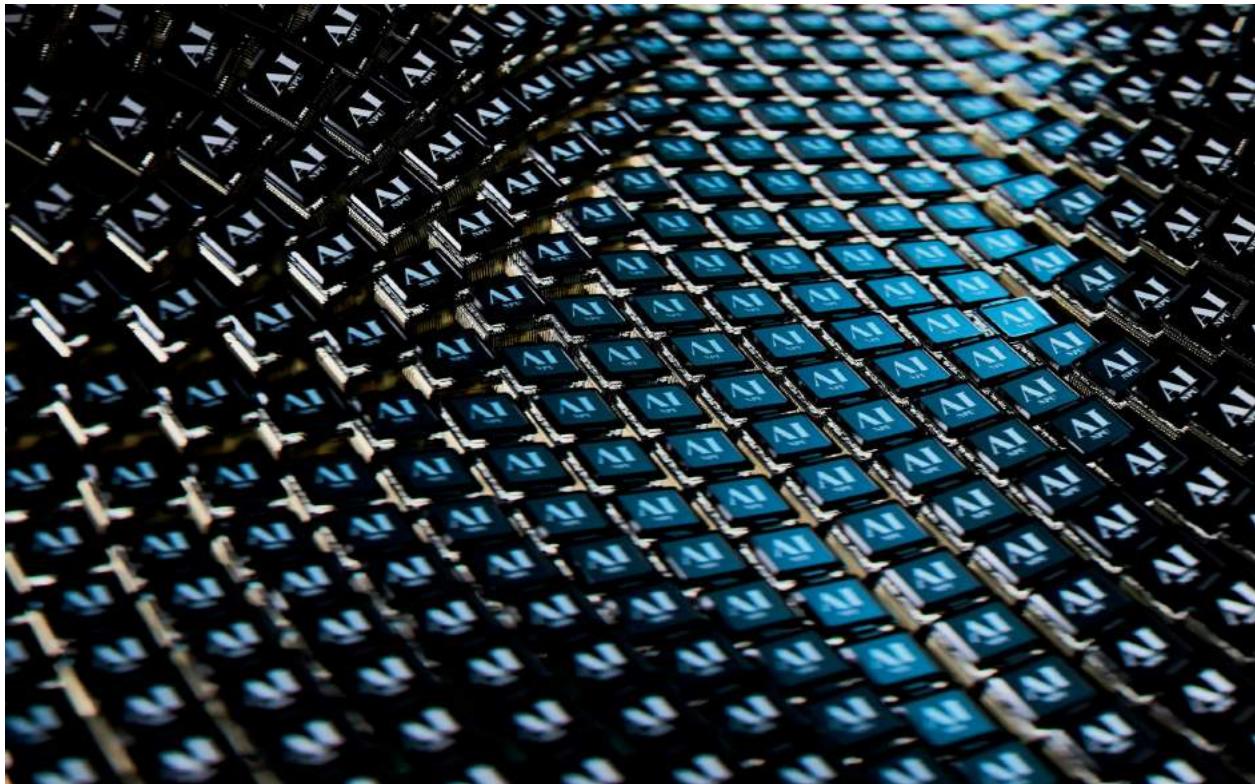
123

閱讀更多

# AI 開發成本上升

AI | 開發成本 | GPU | 計算資源 | 數據中心

2024-07-29



## AI 開發成本上升

隨著人工智慧（AI）技術的不斷演進，像 Microsoft、Alphabet 和 Meta 等大型公司正在面對與其開發相關的高昂費用。儘管他們報告來自 AI 驅動雲端服務的營收顯著增長，但財務壓力卻是實實在在的。根據 Bloomberg 的說法，這種現象被稱為「巨大的金錢黑洞」，強調了追求更先進 AI 模型（例如需要龐大計算資源的大型語言模型）的高成本。

這些費用的一個關鍵因素是對專用 AI 晶片的需求，特別是圖形處理單元（GPU）。Nvidia 的 H100 晶片，對於 AI 訓練至關重要，已成為熱銷商品，價格約為 \$30,000 美元。此外，各公司正大舉投資於定製處理器以減少對供應商的依賴，同時也在管理支持這些複雜系統所需的龐大數據中心。

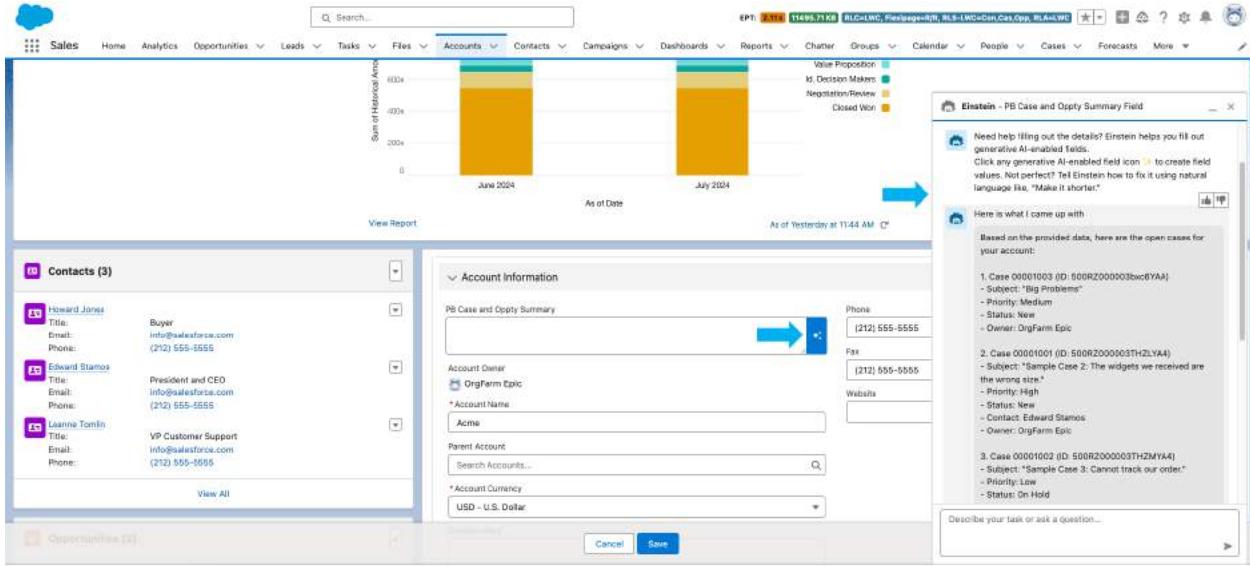
總的來說，隨著成本上升，業界正在探索創新技術，以創造更高效的 AI 模型和分擔計算負載，旨在克服這些財務挑戰，並在 AI 革命中持續進步。

[閱讀更多](#)

# 利用 Amazon Bedrock 和 Salesforce 的生成式 AI 應用程式

**Amazon Bedrock** **Salesforce** **生成式 AI** **大型語言模型** **數據隱私** **自訂模型** **Einstein Trust**  
**Layer** **自動化流程** **客戶互動**

2024-07-29



## 利用 Amazon Bedrock 和 Salesforce 的生成式 AI 應用程式

Amazon Web Services (AWS) 推出了 Salesforce 與 Amazon Bedrock 之間的革命性整合，使企業能夠使用自訂的大型語言模型 (LLMs) 來構建生成式 AI 應用程式。這項創新讓用戶能夠利用 AI 模型的力量，例如 Anthropic 的 Claude，來總結客戶案例並直接在 Salesforce 中提升運營效率。

這項整合透過 Einstein Trust Layer 運行，確保負責任的 AI 實踐與數據隱私。透過利用 Amazon Bedrock 的全托管服務，組織可以根據其獨特需求自訂 AI 模型，而無需承擔管理基礎設施的負擔。

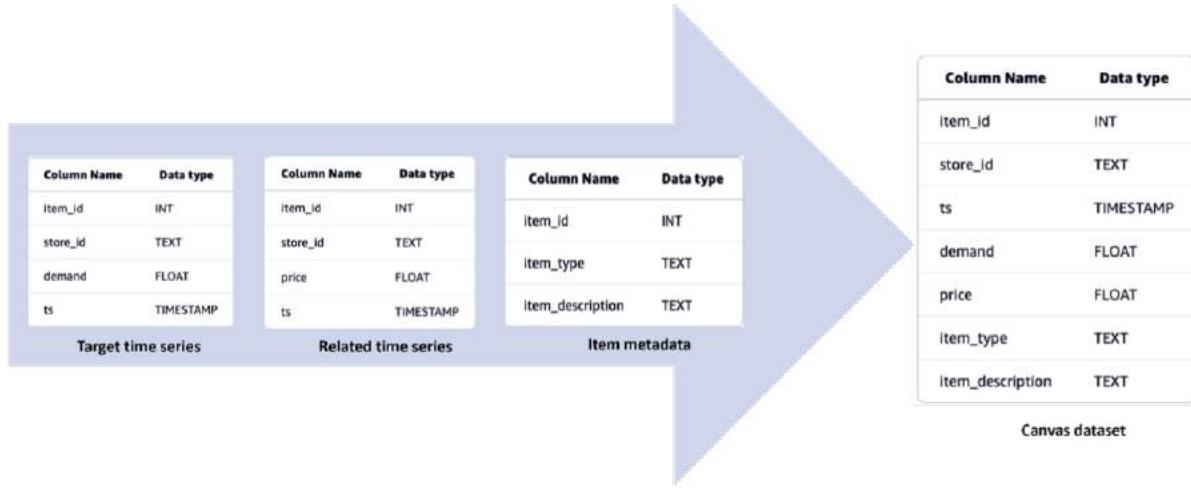
Salesforce 的 Einstein Model Builder 現在進一步增強了這項整合，讓用戶能夠創建和部署自己的模型，促進更直觀的用戶體驗。這項合作使企業能夠利用生成式 AI 技術自動化流程、個性化客戶互動，並最終在各部門推動更智能的決策。

[閱讀更多](#)

# AWS 從 Amazon Forecast 過渡到 SageMaker Canvas

**AWS** **Amazon Forecast** **SageMaker Canvas** 機器學習 預測 數據管理 模型透明度 時間序列  
預測

2024-07-29



## AWS 從 Amazon Forecast 過渡到 SageMaker Canvas

Amazon Web Services (AWS) 宣布將於 2024 年 7 月 29 日停止新的客戶訪問 Amazon Forecast。不過，現有用戶仍可繼續使用該服務，而 AWS 將專注於增強安全性和性能，但不會引入新功能。

為了簡化預測過程，AWS 鼓勵用戶轉向 Amazon SageMaker Canvas，這是一個低程式碼/無程式碼的機器學習模型建構與部署工具。SageMaker Canvas 提供了顯著的優勢，包括更快的模型建構速度——可達 50% 的提速——以及具成本效益的預測。它還提供了一個用戶友好的介面，無需廣泛的機器學習專業知識即可使用。

SageMaker Canvas 中的新功能促進了更好的數據管理和模型透明度，使用戶能夠輕鬆整合各種數據集並進行假設分析。AWS 承諾通過這個平台提供先進的時間序列預測能力，增強預測建模的可及性和效率。

[閱讀更多](#)

# 將 Amazon Q Business 連接到 Microsoft SharePoint Online：利用生成式 AI 提升洞察力

**Amazon Q Business | Microsoft SharePoint Online | 生成式 AI | 數據存儲 | 信息檢索 | 生產力 | 探索洞察**

2024-07-29

[Home](#) > [App registrations](#) > AnyCompanyQ4BApp1

API / Permissions name	Type	Description	Admin consent req...	Status
GroupMember.Read.All	Delegated	Read group memberships	Yes	Granted for [redacted]
Sites.Selected	Delegated	Access selected Sites, on behalf of the sign...	No	Granted for [redacted]
User.Read.All	Delegated	Read all users' full profiles	Yes	Granted for [redacted]

## 將 Amazon Q Business 連接到 Microsoft SharePoint Online：利用生成式 AI 提升洞察力

最近，Amazon 推出了將生成式 AI 助手 Amazon Q Business 與 Microsoft SharePoint Online（用於存儲和組織數據的平台）連接的方法。這項整合使員工能夠利用其組織的數據，輕鬆提問、生成摘要和提取有價值的洞察。

此解決方案利用了 Amazon Q Business Connectors，通過最小權限訪問來尊重現有的用戶權限和訪問控制。通過使用 AWS Secrets Manager 管理的安全憑證，僅使用用戶被授權訪問的數據，確保遵守組織政策。

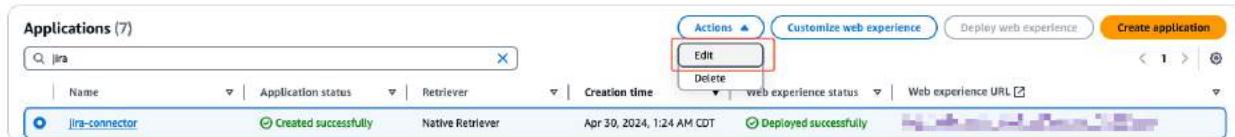
設置過程涉及創建一個連接到 SharePoint Online 的 Amazon Q Business 應用程序，從而實現內容的高效索引和爬取。這項創新不僅簡化了信息檢索，還顯著提升了生產力，促進了組織內的更佳決策和知識分享。

[閱讀更多](#)

# Amazon Q Business 與 Atlassian Jira 整合以提升生產力

Amazon Q Business | Atlassian Jira | 生成式 AI | 生產力 | 整合 | 客戶支持 | 專案管理

2024-07-29



## Amazon Q Business 與 Atlassian Jira 整合以提升生產力

Amazon 推出了其生成式 AI 助手 Amazon Q Business 與 Atlassian Jira 之間的強大整合，顯著改善了客戶支持和專案管理的工作流程。這項整合使團隊能夠以更直觀和高效的方式從 Jira 的廣泛數據中提取有價值的見解。

使用者可以進行自然語言查詢，以查找任務、問題或專案相關資訊，而無需特定的關鍵字。系統能夠總結相關內容，並對模糊的查詢尋求澄清，確保使用者收到針對其需求量身定製的準確結果。

Amazon Q Business 擁有預建的連接器，能夠無縫同步來自多個來源的數據，包括 Jira，使使用者能夠安全地訪問來自其組織資料庫的摘要和答案。透過這些能力，團隊可以增強協作，簡化決策過程，最終提升其軟體運營的生產力。

[閱讀更多](#)

# NVIDIA 在 SIGGRAPH 2024 的 AI 助手 願景

**NVIDIA | AI 助手 | 生成式 AI | 微服務 | DGX Cloud | 機器人技術 | OpenUSD**

2024-07-29



## NVIDIA 在 SIGGRAPH 2024 的 AI 助手願景

在最近的 SIGGRAPH 會議上，NVIDIA 的 CEO 黃仁勳分享了對未來人工智慧的引人入勝的願景，並表示「每個人都將擁有一個 AI 助手」。這反映出在各個領域將人工智慧融入日常工作流程的重要轉變。黃仁勳強調，根植於視覺計算的生成式 AI 將提升人類的生產力，同時推動更具能源效率的計算解決方案。

NVIDIA 揭示了其新的 NIM 微服務套件，旨在滿足從 3D 建模到機器人技術等多樣化的應用，並通過在 DGX Cloud 上整合 Hugging Face 的 Inference-as-a-Service 進一步加強。此外，新的 OpenUSD NIM 微服務是專為生成式物理 AI 應用設計，包括在人形機器人技術方面的進展。

最終，黃仁勳相信 AI 將改變各行各業——無論是在天氣預測、創意合作，還是機器人技術的演進——這預示著未來 AI 與人類創造力的融合將驅動創新。

[閱讀更多](#)

# WPP 與 NVIDIA Omniverse 點燃可口可樂的生成式 AI 內容革命

可口可樂 | WPP | NVIDIA Omniverse | 生成式 AI | 數位行銷 | 3D 廣告資產

2024-07-29



## WPP 與 NVIDIA Omniverse 點燃可口可樂的生成式 AI 內容革命

在一項令人振奮的發展中，可口可樂公司透過與 WPP 和 NVIDIA Omniverse 的合作，利用生成式 AI 來增強其全球市場行銷工作。這項夥伴關係使可口可樂能夠迅速對創意活動進行迭代，確保在超過 100 個市場中的品牌真實性。

這項創新核心是 NVIDIA 的 NIM 微服務，包括 USD Search 和 USD Code，這些工具促進了與當地文化相契合的 3D 廣告資產的創建。通過運用這些工具，可口可樂可以快速生成個性化的影像和行銷內容，解決全球品牌一致性所面臨的挑戰。

此外，WPP 新推出的 Production Studio 整合了這些功能，簡化了多語言內容的創建過程。這種變革性的方法不僅簡化了廣告流程，還使創意人員能夠高效地產出高品質的作品，在數位行銷領域邁出了重要的一步。

[閱讀更多](#)

# NVIDIA 發表 fVDB 以增強數位建模

NVIDIA | fVDB | 深度學習框架 | 虛擬表徵 | AI | 自駕車 | 氣候科學 | 智慧城市 | 神經輻射場 | 激光雷達  
數位雙胞胎

2024-07-29



## NVIDIA 發表 fVDB 以增強數位建模

在近期的 SIGGRAPH 會議上，NVIDIA 介紹了 fVDB，這是一個開創性的深度學習框架，旨在創建適合 AI 的虛擬表徵，重現我們的世界。利用已建立的 OpenVDB 庫，該庫能夠模擬如煙霧和雲朵等體積數據，fVDB 使得自駕車、氣候科學和智慧城市等應用得以加強，讓機器能有效理解和導航 3D 空間。

這個創新框架使用神經輻射場和激光雷達等技術，捕捉大量高解析度的現實世界數據，並將其轉換為廣闊的實時渲染環境。fVDB 顯著提高了操作效率，提供四倍於之前模型的空間範圍，並且處理速度提升至 3.5 倍。

此外，其多種 AI 操作符允許企業為各種應用構建精密的神經網絡，從城市規劃到災難管理等。隨著行業越來越依賴數位雙胞胎，fVDB 成為前所未有的規模上 harnessing 空間智慧的關鍵工具。

[閱讀更多](#)

# NVIDIA 利用先進的生成式 AI 技術強化 數位行銷

NVIDIA | 生成式 AI | 數位行銷 | OpenUSD | NIM 微服務 | 數位雙胞胎 | 客戶參與度 | 個性化體驗

2024-07-29



## NVIDIA 利用先進的生成式 AI 技術強化數位行銷

NVIDIA 在數位行銷的革命前沿，推出強大的生成式 AI 工具，如 OpenUSD 和 NVIDIA NIM 微服務。這些創新使創意機構能夠高效地生產與品牌一致的視覺內容。透過 USD Search NIM 微服務，開發者可以訪問豐富的品牌授權資產庫，簡化行銷材料的組裝過程。

NVIDIA Omniverse 平台正掀起熱潮，允許創建數位雙胞胎——即產品的真實 3D 表現。這項技術不僅提升了客戶參與度和個性化體驗，還使品牌能夠迅速調整內容以適應不斷變化的消費者偏好。

像 Monks 和 Collective World 這樣的公司正在利用這些工具快速提供高品質、可定制的產品。透過將生成式 AI 整合進創意流程，他們正在為廣告設立新的標準，並提升數位行銷活動的整體效率。這些進展讓 NVIDIA 為更具動態性和個性化的行銷環境鋪平了道路。

[閱讀更多](#)

# NVIDIA 揭示創新數位人類技術以提升客戶互動

NVIDIA 數位人類 生成式AI 客戶互動 遠端會議 Maxine James

2024-07-29



## NVIDIA 揭示創新數位人類技術以提升客戶互動

NVIDIA 近期推出突破性的數位人類技術，提升企業與客戶之間的互動。在 SIGGRAPH 大會上，他們展示了「James」，這是一個由 NVIDIA ACE 驅動的互動數位人類，旨在進行即時、情感豐富的對話。James 使用生成式 AI 提供具上下文準確性的回應，讓客戶服務變得更加個性化和有效。

此外，NVIDIA Maxine 正在其創造逼真 2D 和 3D 虛擬形象的能力顛覆遠端會議。Maxine 3D 將 2D 影片輸入轉換為動態 3D 表現，而 Audio2Face-2D 則根據音訊為靜態圖像添加動畫，讓數位人類在對話中栩栩如生。

這項技術正在各行各業中廣泛採用，提升了娛樂和酒店等行業的客戶服務體驗。HTC 和 Reply 等公司正利用這些工具創造沉浸式和反應靈敏的數位品牌大使，突破虛擬客戶互動的界限。

[閱讀更多](#)

# Hugging Face 推出基於 NVIDIA NIM 的推論即服務

**Hugging Face** | **NVIDIA NIM** | 推論即服務 | AI 模型 | 微服務 | **Llama 3** | **Mistral AI** | **DGX Cloud**

2024-07-29



## Hugging Face 推出基於 NVIDIA NIM 的推論即服務

Hugging Face 新推出了一項基於 NVIDIA 的 NIM 微服務的推論即服務解決方案，使開發者能比以往更有效率地訪問熱門 AI 模型。這項服務將令牌處理效率提升至五倍，促進了像 Llama 3 和 Mistral AI 等先進模型在 NVIDIA DGX Cloud 平台上的快速部署。

透過這項提供，開發者可以輕鬆原型設計並部署托管於 Hugging Face Hub 的開源 AI 模型。無伺服器的推論能力最小化了基礎設施需求，同時優化性能，讓用戶只需幾個點擊即可實驗和測試尖端模型。

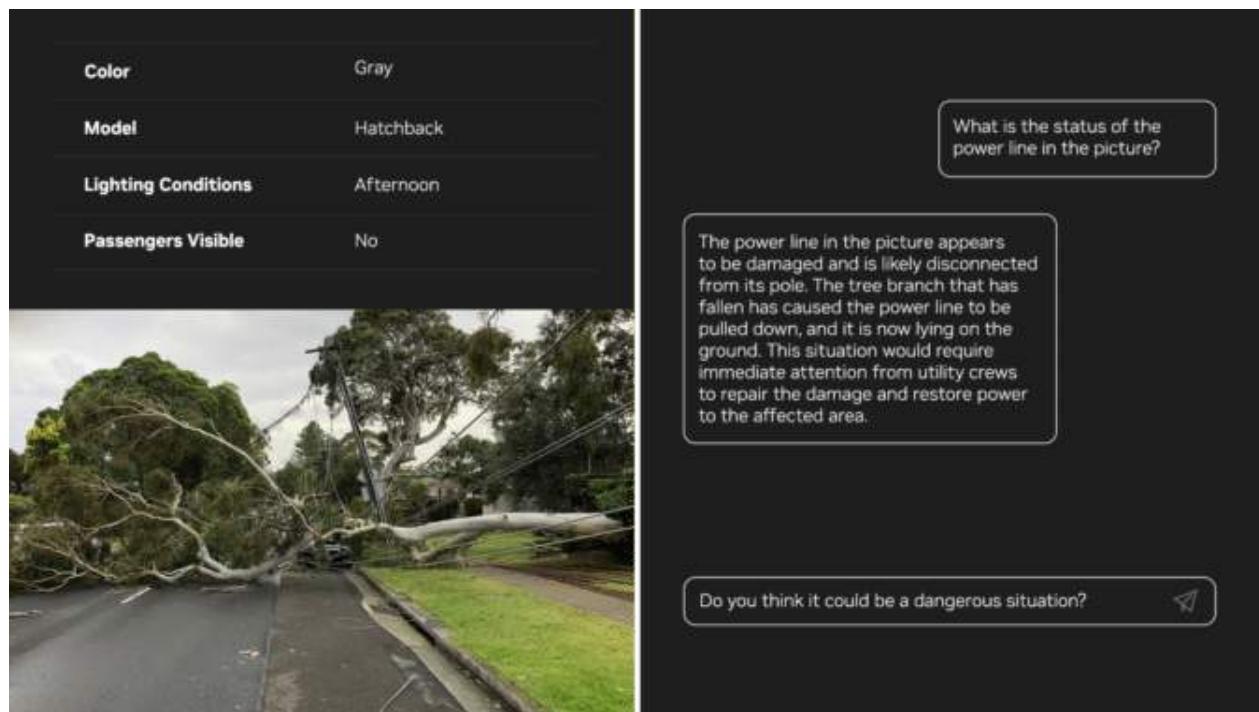
NVIDIA NIM 包含針對推論優化的 AI 微服務，增強了底層的 DGX Cloud 基礎設施。這一進展使開發者幾乎能瞬時訪問強大的計算資源，加速了從開發到市場就緒應用的轉變。總的來說，這一合作為各行各業的 AI 開發和部署鋪平了更高效的道路。

[閱讀更多](#)

# NVIDIA 推出 NIM 微服務以支援實體環境中的生成式 AI

**NVIDIA | NIM 微服務 | 生成式 AI | 視覺 AI 代理 | 數據生成 | 工業自動化**

2024-07-29



## NVIDIA 推出 NIM 微服務以支援實體環境中的生成式 AI

NVIDIA 最近推出了其 NIM 微服務，這是一項在生成式實體 AI 領域的突破性進展，旨在提升數位環境的能力。在 SIGGRAPH 大會上宣布的這些微服務，使開發者能夠訓練實體機器，更有效地完成複雜任務。

新的 NVIDIA Metropolis 參考工作流程允許創建互動式視覺 AI 代理，可用於製造業、醫療保健和智慧城市等各個領域。這些代理使用視覺語言模型 (VLMs) 來解讀現實世界的數據，促進了操作的改善——例如在意大利巴勒莫的交通管理，實時分析交通數據以便更好地管理道路。

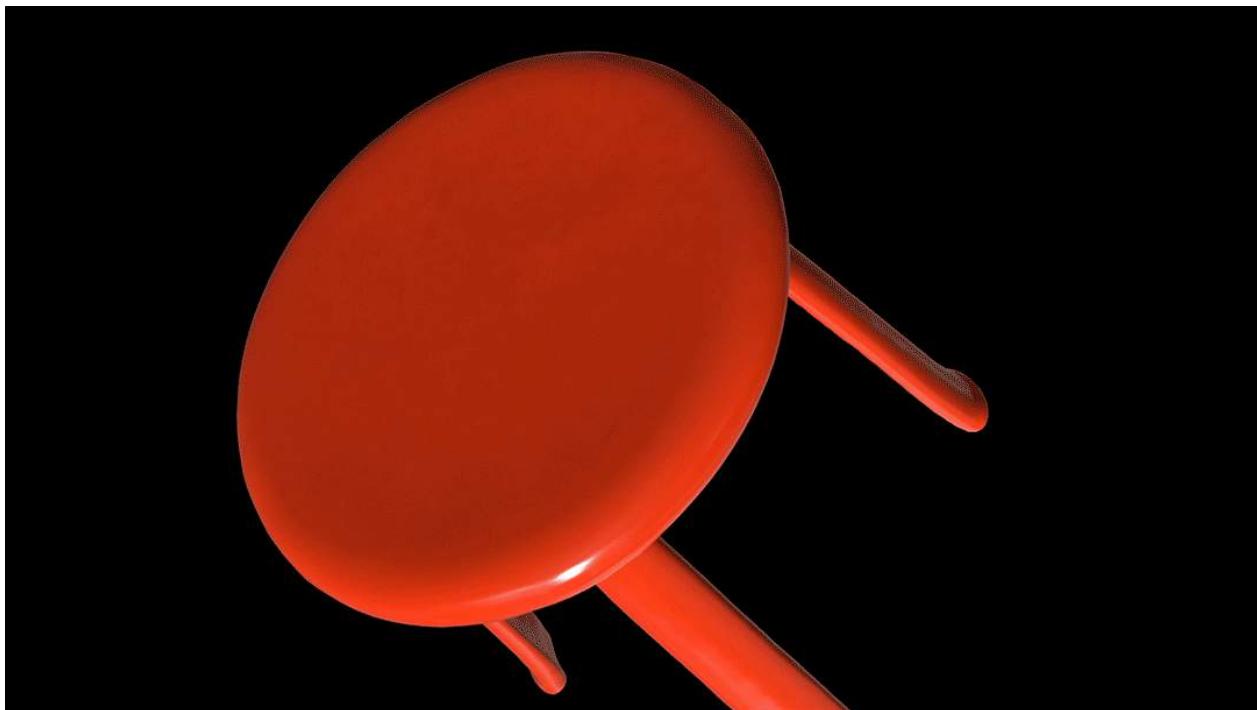
這項技術還支援合成數據的生成，這對於訓練 AI 模型至關重要，因為它不再依賴昂貴的現實世界數據集。這一創新有望改變工業自動化，使流程更安全、更有效，同時縮短模擬與現實之間的差距。

[閱讀更多](#)

# Shutterstock 與 Getty Images 透過生成式 AI 強化創意工作流程

生成式 AI | 3D 服務 | 創意工作流程 | Getty Images | Shutterstock | 圖像質量 | 藝術家 | 設計師 | 生產力 | 創造力

2024-07-29



Shutterstock 與 Getty Images 透過生成式 AI 強化創意工作流程

Shutterstock 推出了其生成式 3D 服務，目前處於商業測試階段，讓設計師能夠使用文字或圖片提示來創建 3D 資產和 360 HDRi 背景。這個創新的平台允許創作者快速原型，僅需幾秒鐘便能生成詳細的 3D 物件，準備好以各種流行格式進行編輯。

與此同時，Getty Images 已經升級了其生成式 AI 服務，使其速度提高了兩倍並改善了圖像質量。用戶現在可以操作圖像構圖和相機設置，例如景深，以生成客製化的視覺效果。該服務還允許整合品牌特定數據以製作量身訂做的圖片。

這兩項服務都利用了 NVIDIA 的 Edify，一種多模態生成式 AI 架構，為藝術家和設計師提升生產力和創造力。隨著這些進步，創意產業將有望在效率和輸出質量上實現顯著提升。

[閱讀更多](#)

# Google 擴展 AI 驅動的野火邊界追蹤工具至歐洲和非洲

Google | AI | 野火邊界追蹤 | 氣候變遷 | 衛星影像 | 公共安全

2024-07-29



## Google 擴展 AI 驅動的野火邊界追蹤工具至歐洲和非洲

為了應對氣候變遷導致的野火威脅日益增加，Google 已經增強其野火邊界追蹤工具，擴展至歐洲和非洲的 15 個新國家。這個創新的工具運用人工智慧和衛星影像，提供有關野火邊界的實時資訊，幫助社區在緊急情況下保持掌握情勢。

新加入的國家包括安道爾、波士尼亞和赫塞哥維納、克羅埃西亞、塞浦路斯、法國、希臘、意大利、肯尼亞、摩納哥、黑山、葡萄牙、盧旺達、斯洛文尼亞、西班牙和土耳其。透過 Google 搜尋和地圖，用戶可以接收到多語言的警報和安全提示，確保當地居民和旅客能夠獲取重要資訊。

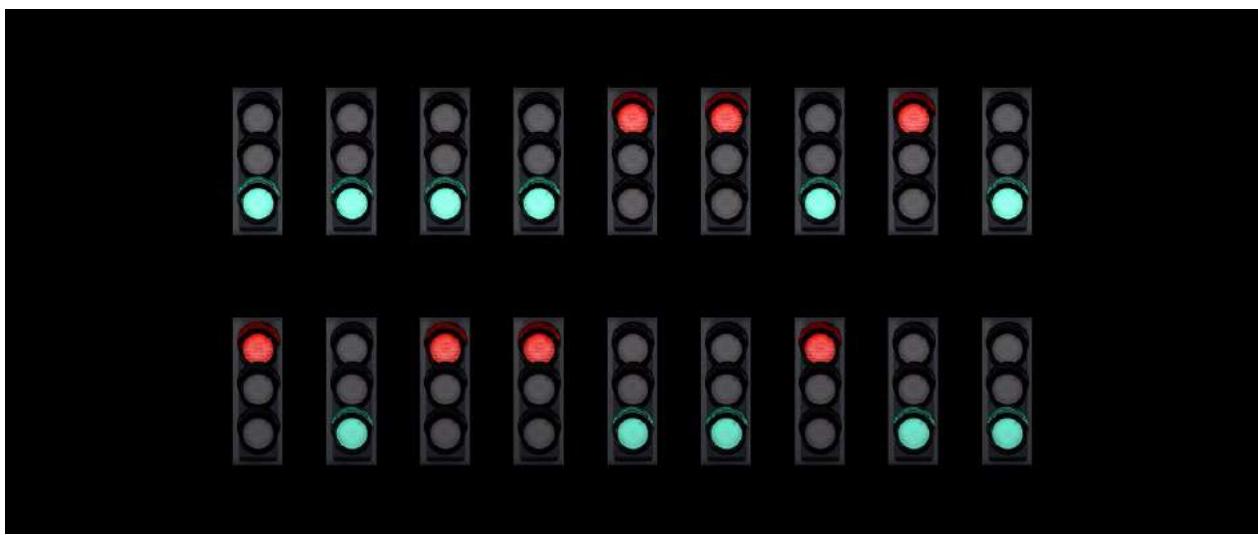
這項倡議旨在透過提供準確、可靠的野火數據來增進公共安全，輔助地方政府的努力，並提高社區對氣候變遷帶來日益嚴峻挑戰的韌性。

[閱讀更多](#)

# 谷歌的綠燈計畫：利用 AI 解決交通排放問題

谷歌 AI 交通排放 綠燈計畫 城市交通 數據分析

2024-07-29



## 谷歌的綠燈計畫：利用 AI 解決交通排放問題

谷歌的研究團隊啟動了一項名為綠燈計畫的創新倡議，旨在減少城市交叉路口的停停走走交通，降低燃料排放。透過利用谷歌地圖的龐大數據，該計畫分析交通流量並建議最佳的紅綠燈時機。

目前在多個城市的 70 多個交叉路口運行，綠燈計畫有潛力減少 30% 的停車次數，並將排放量降低多達 10%。實施過程相當簡單，城市不需額外投資硬體或軟體；工程師可在短短五分鐘內根據 AI 生成的建議採取行動。

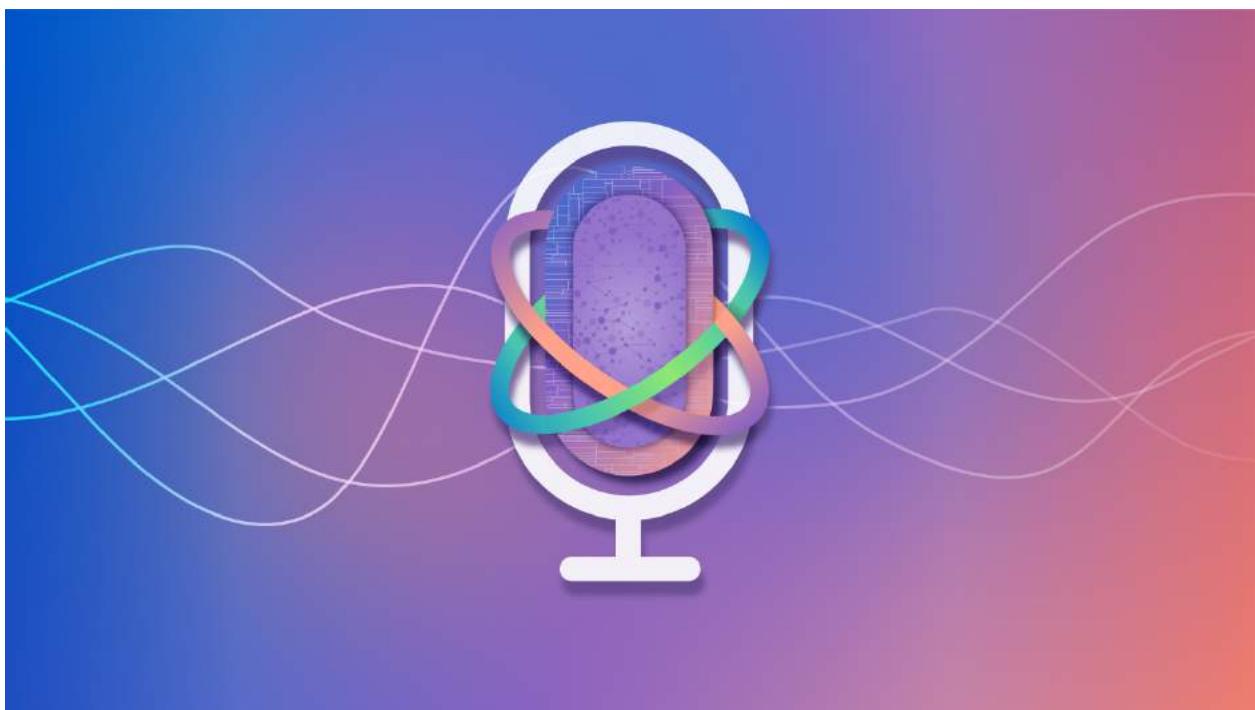
綠燈計畫有志於擴展至數百座城市和數千個交叉路口，不僅旨在提高交通效率，還希望能創造更順暢的駕駛體驗，同時對環境做出正面貢獻。

[閱讀更多](#)

# 微軟研究發表 LongRoPE：語言模型能力的飛躍

**LongRoPE** | 語言模型 | 上下文窗口 | **RoPE** | 微軟 | **Phi-3** | 人工智慧

2024-07-29



## 微軟研究發表 LongRoPE：語言模型能力的飛躍

在2024年7月29日，微軟研究團隊推出了 LongRoPE，這是一種突破性的技術，旨在將大型語言模型（LLMs）的上下文窗口擴展到超過200萬個標記。這一進展使得模型能夠處理大規模的輸入——想像一下整個哈利波特系列——而不會影響性能。傳統上，LLMs的限制約為4000個標記，相當於只有10頁的文本。

LongRoPE 採用了名為 RoPE（旋轉位置嵌入）的技術，以管理變壓器模型中嵌入位置的複雜性。這一創新方法對現有模型結構的調整需求極小，便於集成到應用中。

LongRoPE 方法增強了微軟 Phi-3 系列語言模型的能力，使其能夠實現包括長上下文檢索和代碼除錯等實際應用。這項研究不僅推動了 LLM 性能的邊界，還為尋求先進人工智能能力的用戶提供了顯著的好處。

[閱讀更多](#)

# AI 革命：來自 NVIDIA 和 Meta 領導者的見解

人工智慧 | NVIDIA | Meta | AI Studio | 開源 AI | Llama 3.1 | 擴增實境 | 數位人類

2024-07-30



## AI 革命：來自 NVIDIA 和 Meta 領導者的見解

在 SIGGRAPH 2024 會議上，NVIDIA CEO 黃仁勳與 Meta CEO 馬克·祖克柏強調了未來每個企業都將人工智能(AI)整合進其運營的願景。祖克柏介紹了 AI Studio，一個新的平台，讓用戶能創建和分享 AI 角色，讓 AI 開發的門檻降低。黃仁勳預測，未來的環境中，AI 將像網站和社交媒體帳戶一樣普遍。

此次討論展示了 NVIDIA 的進展，包括「詹姆斯」這位數位人類，旨在提升客戶互動。兩位領導者都強調了開源 AI 的重要性，Meta 的 Llama 3.1 模型就是重大資源投資的範例。

展望未來，黃仁勳預見 AI 將發展到能夠處理複雜的互動，而祖克柏則看到了將 AI 與擴增實境結合的潛力。隨著各行各業的企業採用這些技術，SIGGRAPH 上的對話暗示著一個轉型時代的來臨，AI 將成為日常運營的重要組成部分。

[閱讀更多](#)

# JPMorgan 發表 LLM 套件：一款突破性的 AI 聊天機器人用於研究分析

JPMorgan | LLM 套件 | AI 聊天機器人 | 研究分析 | 金融業 | 數據安全 | 法規遵循 | 生成式 AI

2024-07-30



## JPMorgan 發表 LLM 套件：一款突破性的 AI 聊天機器人用於研究分析

JPMorgan Chase 推出了創新的生成式 AI 工具 LLM 套件，旨在協助公司內的研究分析師。這個先進的平台功能類似於流行的聊天機器人如 ChatGPT，但專為金融業量身打造。大約 50,000 名資產及財富管理部門的員工均可使用 LLM 套件，該工具能執行各種任務，如撰寫、生成創意及總結文件，徹底改變員工處理日常工作負載的方式。

值得注意的是，LLM 套件遵循嚴格的金融法規，確保敏感的客戶數據安全地保留在銀行的基礎設施內。與利用外部 AI 解決方案的競爭對手不同，JPMorgan 是在內部開發這項工具，進一步強調其對數據安全和法規遵循的承諾。雖然 LLM 套件的推出標誌著金融領域 AI 的一個重要里程碑，但這項技術的限制包括潛在的不準確性以及「幻覺」的風險，這是 AI 模型中常見的問題。

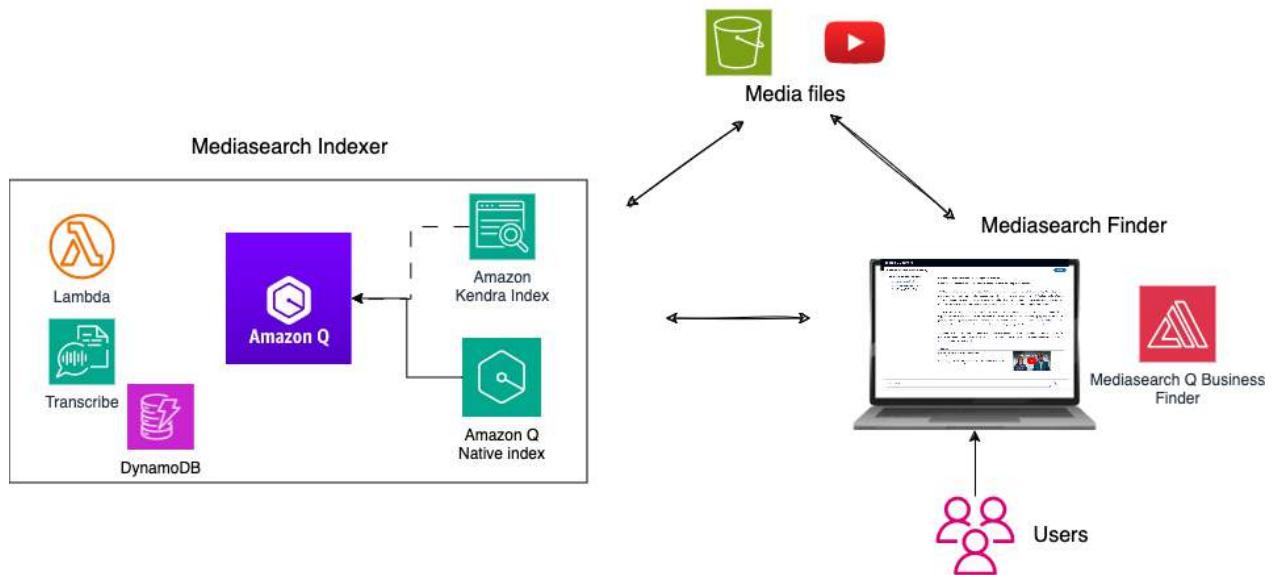
[閱讀更多](#)

# 透過 Amazon Q Business 和 Transcribe 強化媒體搜尋

**Amazon Q Business** | **Transcribe** | 媒體搜尋 | 自動轉錄 | 音訊 | 視頻 | 搜尋體驗 | 媒體資料

2024-07-30

## Mediasearch Q Business



### 透過 Amazon Q Business 和 Transcribe 強化媒體搜尋

因應對音訊和視頻內容日益增長的需求，Amazon 推出了 Mediasearch Q Business——一個旨在優化媒體搜尋體驗的強大工具。這個開源解決方案整合了 Amazon Q Business 和 Amazon Transcribe，讓在龐大的媒體資料中搜尋變得更加直觀和高效。

Mediasearch Q Business 透過自動轉錄儲存在 Amazon S3 或從 YouTube 獲取的音訊和視頻檔案，簡化了這一過程。用戶可以輕鬆查詢這些媒體內容，並直接在搜尋結果中訪問相關部分，這得益於內建的時間標記，指引播放。

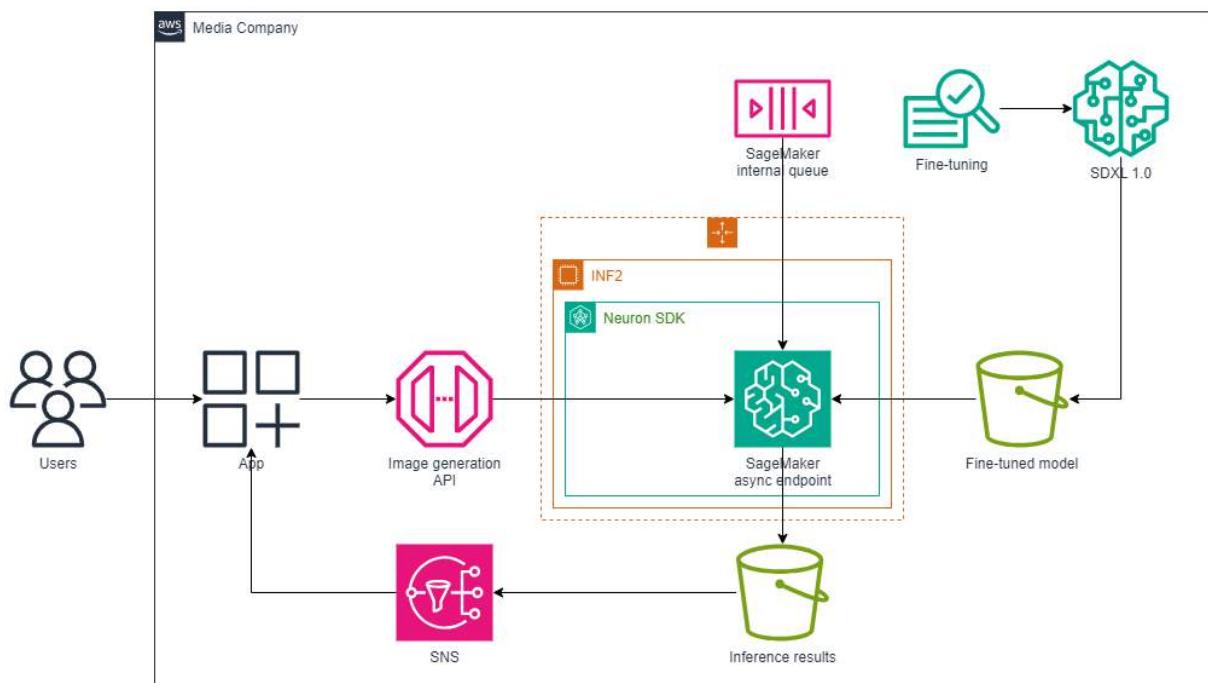
該解決方案由兩個主要組件組成：處理和轉錄媒體檔案的索引器，以及允許用戶搜尋和互動內容的查找器。這項創新不僅改善了組織管理媒體的方式，還通過基於轉錄數據提供快速、可靠的答案，顯著提升了用戶參與度。

[閱讀更多](#)

# Monks 使用 AWS 技術提升即時 AI 圖像生成

**AI 圖像生成 | AWS | Inferentia2 | SageMaker | 雲端技術**

2024-07-30



## Monks 使用 AWS 技術提升即時 AI 圖像生成

Monks 在 AI 驅動的圖像生成領域取得了重大進展，透過整合 AWS Inferentia2 晶片和 Amazon SageMaker，使處理速度提升了四倍。這些技術的創新應用使 Monks 能夠高效且具成本效益地應對即時擴散 AI 圖像生成的挑戰。

藉由利用 SageMaker 的非同步推理端點，Monks 能夠處理大量圖像生成請求，同時維持低延遲，並將每張圖像的成本降低 60%。AWS Inferentia2 晶片為深度學習任務提供了優化的性能，平均每小時處理近 28,000 張圖像。

這項解決方案不僅解決了擴展性的問題，還通過確保快速的圖像生成來提升用戶體驗—通常在 10 秒內完成。Monks 的實施展示了雲端技術如何革新 AI 應用，滿足高需求而不妥協效率或預算。

[閱讀更多](#)

# Amazon Bedrock 推出網頁爬蟲功能以增強知識庫

Amazon Bedrock | 網頁爬蟲 | 知識庫 | 生成式 AI | 數據收集

2024-07-30

## Review and create

### Step 1: Provide details

[Edit](#)

#### Knowledge base details

**Knowledge base name**  
knowledge-base-web-crawl-blog

**Knowledge base description**  
—

**Service role**  
AmazonBedrockExecutionRoleForKnowledgeBase\_4nmrs

#### Tags (0)

Key	Value
No tags to display	

### Step 2: Setup up data source

[Edit](#)

#### Data source: knowledge-base-quick-start-bf8g2-data-source

**Data source type**  
WEB

**Data source name**  
knowledge-base-quick-start-bf8g2-data-source

**Sync scope**  
Host only

**Authentication**  
No Authentication

### Step 3: Select embeddings model and configure vector store

[Edit](#)

#### Embeddings model

**Model**  
Titan Text Embeddings v2

**Vector dimensions**  
1024

#### Vector store

##### Quick create vector store - Recommended

We will create an Amazon OpenSearch Serverless vector store in your account on your behalf.

[Cancel](#)
[Previous](#)
[Create knowledge base](#)

## Amazon Bedrock 推出網頁爬蟲功能以增強知識庫

Amazon Bedrock 最近推出了一項新功能，將網頁爬蟲能力整合進其知識庫中。這項進步使得用戶能夠利用來自公共網站的最新資訊，顯著提升生成式 AI 應用的準確性和相關性。

這個網頁爬蟲能夠對面向公眾的網站進行索引，收集多樣的數據，以建立更全面的信息庫。用戶可以指定種子 URL，控制爬蟲速度，並應用包含和排除過濾器來精煉數據來源。這意味著公司可以設計應用程式，從豐富的在線內容中提取資訊，而無需進行大量的手動數據輸入。

透過利用這項功能，用戶可以高效地建立依賴即時網頁數據的應用程式，確保其 AI 模型獲取最新和最相關的資訊。這一舉措突顯了 Amazon 致力於增強生成式 AI 在各種商業應用中的能力。

[閱讀更多](#)

# Meta 與 NVIDIA 發表 AI Studio，打造個人化助手

Meta | NVIDIA | AI Studio | 個人化助手 | 開源 AI | Llama 3.1

2024-07-30



## Meta 與 NVIDIA 發表 AI Studio，打造個人化助手

在最近於 SIGGRAPH 2024 的討論中，Meta 執行長馬克·祖克柏（Mark Zuckerberg）和 NVIDIA 執行長黃仁勳（Jensen Huang）強調了開源 AI 的變革潛力，並介紹了 AI Studio。這個全新平台使用戶能夠創建、分享和發現 AI 角色，讓先進的 AI 技術對創作者和小型企業而言變得更加可及。

祖克柏預見，隨著每個企業擁有電子郵件和網站，他們不久後也將擁有個人化 AI 助手。他強調，隨著如 Llama 3.1—這個擁有 4050 億參數的開源模型—等技術的進步，AI 將越來越多地統合多樣化的內容類型，提升互動性和生產力。

展望未來，兩位領導人都看到了 AI 互動超越文字的演變，允許更流暢且豐富的體驗。這項合作旨在賦予個人數位助手的能力，潛在地改變我們日常生活中的各個領域，從教育到娛樂。

[閱讀更多](#)

# 蘋果選擇Google晶片進行AI開發，拋棄Nvidia

人工智慧 | Google TPUs | 蘋果 | Nvidia | AI模型訓練 | Siri | 電子郵件摘要工具

2024-07-31



## 蘋果選擇Google的晶片進行人工智慧發展，拋棄Nvidia

在一個戰略性轉變中，蘋果宣布決定使用Google的張量處理單元（TPUs）作為其人工智慧（AI）基礎設施的核心，逐步擺脫對Nvidia圖形處理單元（GPUs）的傳統依賴。考慮到Nvidia在市場上的主導地位，這一轉變可能會對AI領域產生重大影響，因為Nvidia佔據了包括Google和Amazon等大型公司在內的雲計算資源的80%。

蘋果的AI模型訓練將利用兩種類型的Google TPUs，部署了2080個TPUv5p晶片和8192個TPUv4處理器。這種做法不僅多樣化了蘋果的硬體依賴性，也暗示著未來有潛力開發出更複雜的AI模型。隨著蘋果向測試用戶推出AI驅動的功能，這一轉變似乎是朝著增強功能的明確步驟，特別是在Siri的能力以及電子郵件摘要工具方面。

[閱讀更多](#)

# 解鎖日本 LLM 與 AWS Trainium：人工智慧發展的新前沿

大型語言模型 | AWS Trainium | 日本 | 人工智慧 | 課程學習 | 雙語 LLM | 預訓練 | tsuzumi 模型  
客戶支援聊天機器人 | 多模態人工智慧

2024-07-31



## 解鎖日本 LLM 與 AWS Trainium：人工智慧發展的新前沿

AWS LLM 開發支援計畫正在推動日本的創新，賦能組織運用大型語言模型 (LLMs) 和基礎模型 (FMs)。值得注意的是，十五位參與者中有十二位使用 AWS Trainium 晶片來訓練他們的模型，促進了人工智能能力的重大進步。

理光 (Ricoh) 採用了課程學習策略開創了一個雙語 LLM，以強化日語處理能力；而 Stockmark 則專注於對他們的 LLM 進行預訓練，以減少幻覺——這是人工智慧輸出中的常見問題。同時，NTT 正在開發輕量級的 tsuzumi 模型，該模型在日語能力和多模態功能方面表現優異，適用於包括醫療保健在內的各種應用。

該計畫還見證了專業領域模型的創建，比如 KARAKURI 的客戶支援聊天機器人和 Sparticle 的多模態人工智慧解決方案。這些進展突顯了生成式人工智慧在日本的變革潛力，為未來各行業的創新鋪平了道路。



# AWS 發佈 ApplyGuardrail API 以增強 Amazon Bedrock 的內容管理功能

[ApplyGuardrail API](#) [Amazon Bedrock](#) [內容管理](#) [生成式 AI](#) [安全性](#) [道德標準](#) [合規性](#)

2024-07-31



AWS 發佈 ApplyGuardrail API 以增強 Amazon Bedrock 的內容管理功能

Amazon Web Services (AWS) 推出了 ApplyGuardrail API，旨在透過在輸入及輸出過程中應用嚴格的內容管理來改善生成式 AI 應用的安全性。此 API 幫助防止生成有害或偏見的內容，通過執行符合道德和法律標準的指導方針來實現。

ApplyGuardrail API 允許開發者即時評估使用者輸入和模型回應，在 Amazon Bedrock 的大型語言模型中維持合規性。它支持長上下文輸入並啟用串流輸出，對於對話式 AI 和即時字幕等應用特別有用。

此 API 具有彈性，允許針對多種用例定制多重防護措施，能夠過濾敏感信息、拒絕某些主題以及確保遵循內容政策。透過實施高效的分塊策略和即時評估，ApplyGuardrail API 有望顯著提升 AI 技術的負責任使用。

[閱讀更多](#)

# NVIDIA 展示即時生成 AI 於 3D 世界建構

**NVIDIA Edify** | **3D生成** | **AI技術** | **Omniverse** | 藝術家 | 沙漠景觀 | 資產生成 | **USD** | 沉浸式環境

2024-07-31



## NVIDIA 展示即時生成 AI 於 3D 世界建構

在最近的 SIGGRAPH 大會上，NVIDIA 的研究人員展示了其 NVIDIA Edify 技術的驚人能力，這是一個視覺生成 AI 模型。在一次現場演示中，他們在幾分鐘內創建了一個詳細的 3D 沙漠景觀，展示了 AI 如何能顯著協助藝術家建構沉浸式環境。透過利用 AI 代理，他們能夠快速且高效地生成和編輯資產，例如仙人掌和岩石。

Edify 模型，包括 Edify 3D 和 Edify 360 HDRi，使得用戶能夠從簡單的文字或圖片提示中創建高品質的 3D 資產和背景。這項創新使藝術家能專注於關鍵視覺元素，同時加快支持資產的生成。此

外，運用 Universal Scene Description (USD) 確保這些 3D 物件能無縫整合進 NVIDIA 的 Omniverse 平台，提升數位藝術與設計領域的創作流程和生產力。

[閱讀更多](#)

# Oracle Cloud Infrastructure 擴展 GPU 加速實例以支持 AI 和數位雙胞胎

Oracle Cloud Infrastructure | AI | 數位雙胞胎 | GPU 加速 | NVIDIA | 虛擬機 | 高效能

2024-07-31

The graphic is a promotional image for SIGGRAPH 2024, held in Denver from July 28 to August 1. It features two headshots side-by-side: Jensen Huang on the left and Mark Zuckerberg on the right. Both are smiling. Above them is the text "Jensen Huang and Mark Zuckerberg on AI Breakthroughs". Below their names are their respective titles: "NVIDIA" and "Meta". At the bottom left is a green button labeled "Watch Now". The top right corner displays the SIGGRAPH logo and the event details.

Oracle Cloud Infrastructure 擴展了針對 AI 和數位雙胞胎的 GPU 加速實例

Oracle Cloud Infrastructure (OCI) 已推出新的 NVIDIA GPU 加速實例，增強企業在生成式 AI 和數位雙胞胎技術方面的能力。這些新產品包括搭載 NVIDIA L40S GPU 的裸金屬實例，旨在高效處理複雜的工作負載。L40S GPU 在生成令牌方面表現卓越，速度遠超其前身 A100。

此外，OCI 計劃推出一款搭載單一 NVIDIA H100 Tensor Core GPU 的虛擬機，為小型高效能應用提供具成本效益的解決方案。這款虛擬機將提升各種 AI 任務的處理能力，提供顯著的性能改進。

此外，OCI 的基礎設施利用了 NVIDIA 的最新技術，如 BlueField-3 DPU，確保在網路、儲存和安全工作負載上實現最佳性能。這些進展使 OCI 成為企業在提升效率和創新方面的強大平台。

[閱讀更多](#)

# NVIDIA 加速 AI 發展，推出最新 RTX 創新

NVIDIA | AI | RTX | ChatRTX | Llama 3 | 光線追蹤 | 渲染 | 虛擬助理 | 遠端會議 | 數位媒體 | 沉浸式應用

2024-07-31



## NVIDIA 加速 AI 發展，推出最新 RTX 創新

在最近的 SIGGRAPH 會議上，NVIDIA 展示了在 AI 技術方面的突破性進展，這些進展提升了 PC 和工作站的內容創作與互動體驗。主要亮點包括全新的 ChatRTX 及對 Llama 3 的支援，旨在簡化 AI 驅動的生產力和開發。

NVIDIA 的 RTX 驅動工具實現了先進的光線追蹤和渲染，豐富了遊戲和虛擬實境中的圖形。由 NVIDIA ACE 驅動的虛擬助理“James”的推出，展現了情感智商客戶互動的潛力。

此外，NVIDIA Maxine AI 平台的更新也值得注意，增強了遠端會議的真實感，讓虛擬角色能夠進行即時互動。對於創作者而言，RTX Video HDR 和 Super Resolution 提升了影片品質，而像 Adobe 的 Substance 3D Modeler 和 Topaz AI 等工具則提供了更快速的工作流程和優質的輸出。

這些創新反映了 NVIDIA 對推進數位媒體 AI 邊界的承諾，並展望沉浸式應用的未來。

[閱讀更多](#)

# 微軟研究亮點：2024年7月29日當週

**SPOT** | 腦-機接口 | 生成式人工智慧 | 可微分因果發現 | 視覺想像 | 創意專業人士

2024-07-31



微軟研究亮點：2024年7月29日當週

微軟研究本週揭示了幾項突破性的研究，展示了各個領域的進展。

其中一項關鍵創新是 SPOT (Skeleton Posterior-guided Optimization)，這項技術增強了可微分因果發現，使研究人員能夠處理更大規模的數據集，並解決潛在混淆變數所帶來的挑戰。這個框架顯著提升了因果圖學習的表現。

另一項值得注意的研究專注於 基於腦電圖的腦-機接口 (BCIs)，探索視覺想像作為重度神經肌肉疾病患者的控制方法的有效性。研究結果顯示，短期的視覺想像可以產生更清晰的神經信號，這可能增強受影響個體的溝通能力。

此外，研究人員調查了創意專業人士在 生成式人工智慧 時代中角色的演變，揭示了這些技術如何增強工作流程和創意過程的見解。

這些研究反映了微軟研究對推動技術邊界和改善現實應用的承諾。

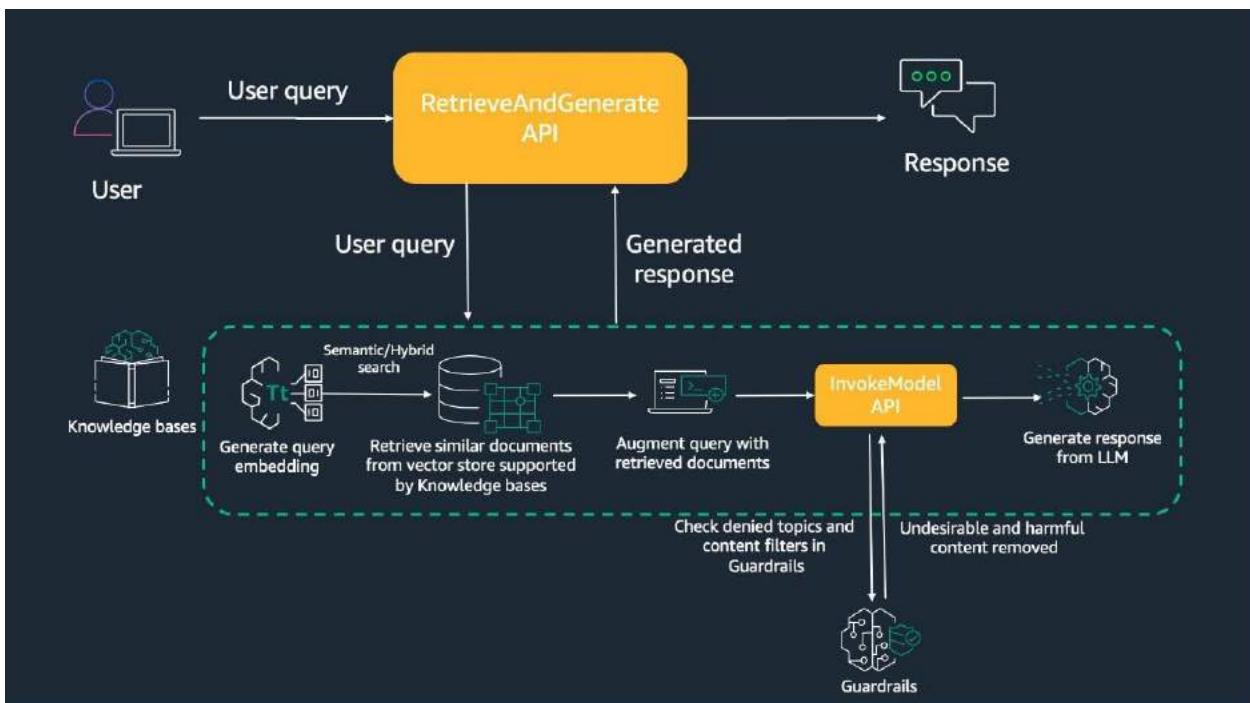
[閱讀更多](#)

# 03 資訊安全

# Amazon Bedrock 推出增強知識庫的安全防護措施

**Amazon Bedrock** | 生成式 AI | 安全防護措施 | 合規性 | 檢索增強生成 | 敏感信息 | AI 實踐

2024-07-03



## Amazon Bedrock 推出增強知識庫的安全防護措施

Amazon Bedrock 進行了一次重大升級，推出了增強生成式 AI 應用程序安全性和合規性的安全防護措施。這項功能與知識庫無縫整合，能透過檢索增強生成 (Retrieval Augmented Generation, RAG) 將基礎模型與企業數據連接。

安全防護措施允許用戶實施自訂的安全措施，以過濾有害內容並保護敏感信息，確保更加安全的使用體驗。例如，在法律環境中，安全防護措施可以防止檢索機密客戶詳情，而在金融服務中，它們則能防範未經授權的投資建議。同樣地，在電子商務的客戶支持中，安全防護措施可以阻止敏感個人信息的曝光。

這一發展不僅標準化了各種應用程序的安全控制，還符合負責任的 AI 實踐，使企業能夠更有信心和安全地利用 AI 的力量。

[閱讀更多](#)

# Mend.io 利用 Anthropic Claude 分析 CVE 數據

**Mend.io** **Anthropic Claude** **Amazon Bedrock** **CVE** 漏洞分析 **AI** 網路安全

2024-07-18



Mend.io 利用 Amazon Bedrock 上的 Anthropic Claude 來分析 CVE 數據

在網路安全領域的一項重大進展中，Mend.io 已經運用 Anthropic Claude 的能力，透過 Amazon Bedrock 簡化對常見漏洞與暴露（CVEs）的分析。這一創新方法使 Mend.io 能夠自動分類超過 70,000 種漏洞，此任務通常需要 200 天的專家手動工作。透過利用大型語言模型（LLMs），他們能高效提取有關攻擊需求的重要信息，這些信息通常在複雜且常常模稜兩可的 CVE 報告中。

在提示設計中整合 XML 標籤有助於提高分析的精確度。儘管面臨管理服務配額及細化提示等挑戰，Mend.io 在識別關鍵漏洞細節方面取得了 99.9883% 的驚人準確率。這一發展不僅優化了安全評估過程，也凸顯了 AI 在增強我們對不斷演變的網路威脅防禦方面的變革性影響。

[閱讀更多](#)

# AWS 強化對 PII 的數據保護，結合 Amazon Lex 和 CloudWatch Logs

**AWS PII 數據保護 Amazon Lex CloudWatch Logs 敏感數據 加密 訪問控制**

2024-07-23

▼ Slots (5) - optional [Info](#)

Information that a bot needs to fulfill the intent. The bot prompts for slots required for intent fulfillment, in priority order below.

Add slot

Slot	Prompt for slot	Slot type
1	Prompt for slot: Name Message: What is your name?	Slot type FullName
2	Prompt for slot: Address Message: What is your address?	Slot type Address
3	Prompt for slot: PhoneNumber Message: What is your phone number?	Slot type AMAZON.PhoneNumber
4	Prompt for slot: EmailAddress Message: What is your email address?	Slot type AMAZON.EmailAddress
5	Prompt for slot: AccountNumber Message: What is your account number?	Slot type AMAZON.AlphaNumeric

## AWS 強化對 PII 的數據保護，結合 Amazon Lex 和 CloudWatch Logs

為了顯著增強數據安全性，AWS 將先進功能整合進 Amazon Lex 和 CloudWatch Logs，旨在保護個人可識別資訊（PII）。這些新功能專注於在用戶互動過程中檢測和模糊化敏感數據。

Amazon Lex 現在採用插槽模糊化功能，在捕獲敏感資訊（如姓名和地址）時進行遮蔽，從而降低在日誌中暴露的風險。與此同時，CloudWatch Logs 也推出了數據保護政策，對 PII 進行審計和遮蔽，利用管理和自訂數據識別符提升安全性。

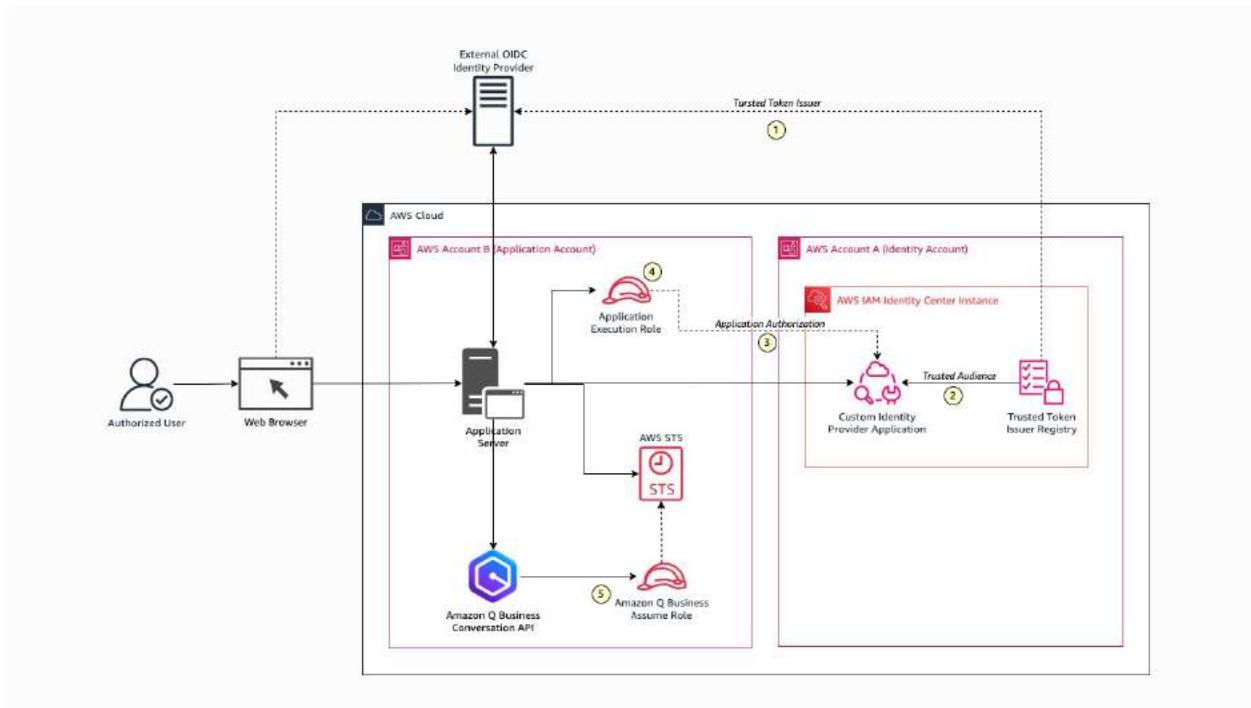
組織還可以利用 Amazon S3 安全存儲音訊錄音，確保加密和訪問控制措施到位。通過實施這些策略，企業能有效保護敏感數據，遵循法規標準，並在不斷上升的數據隱私擔憂中建立更大的消費者信任。



# Amazon Q Business 推出可信身份傳播功能

可信身份傳播 | AWS IAM | 用戶身份驗證 | API 安全性 | 數據隱私 | 生成式 AI

2024-07-30



Amazon Q Business 推出了創新的功能，增強了企業應用程式中 AI 驅動互動的安全性。這項功能稱為可信身份傳播，採用了 AWS IAM 身份中心來確保用戶身份在整個過程中得到準確驗證和保護。

透過整合可信身份傳播，應用程式可以通過外部身份提供者 (IdP) 進行用戶身份驗證，並安全地訪問 Amazon Q Business 的 API。這種方法使組織能夠根據用戶屬性（如群組成員身份）授權請求，同時防止未經授權訪問私有數據。

設置過程可以透過 AWS CloudFormation 模板自動化，讓開發者更容易實施這些安全措施。這項進步不僅改善了用戶體驗，還符合當今數據敏感環境中加強隱私控制的需求，使企業能夠自信地利用生成式 AI。

[閱讀更多](#)

# 04 應用

# 蘋果加入 OpenAI 董事會擔任觀察員

蘋果 | OpenAI | 董事會 | ChatGPT | iOS 18 | 人工智慧

2024-07-03



## 蘋果加入 OpenAI 董事會擔任觀察員

對於這兩家公司來說，這是一個重要的舉措，蘋果宣布將在 OpenAI 的董事會上擔任「觀察員」席次。這項安排將於今年稍晚生效，蘋果的資深專家 Phil Schiller 將代表這家科技巨頭。雖然 Schiller 將參加董事會會議，但他不會擁有表決權。

這次合作是在蘋果最近與 OpenAI 的夥伴關係之後，目的是將 ChatGPT 整合進即將推出的 iOS 18，作為蘋果智慧套件的一部分。有趣的是，這次合作並不涉及任何財務交易；蘋果認為 ChatGPT 在其平台上的曝光度極具價值。

透過這個觀察員的角色，蘋果期望能獲得有關 OpenAI 策略和決策過程的重要見解，進一步在快速發展的人工智慧領域鞏固自身的地位。Schiller 的參與尤為引人注目，因為他在蘋果的計畫中擁有多樣的經驗和貢獻。

[閱讀更多](#)

# 美國教育中的AI革命：中國應用程式 的影響

AI教育 | 中國應用程式 | 數據隱私 | 教育方法 | 人工智慧

2024-07-09



## 美國教育中的AI革命：中國應用程式的影響

最近，中國的AI教育應用程式如Question.AI和Gauth在美國市場取得了顯著進展，幫助學生更有效地解決作業。這些應用程式擁有創新的功能，允許用戶拍照其作業問題，並獲得即時的逐步解決方案，因此受到廣泛歡迎。Question.AI於2023年中推出，而Gauth最初是一個數學解題工具，自2020年開始運營後擴展了其服務。

這些工具的快速普及突顯了其背後先進的AI技術，該技術能夠個性化學習體驗。在這個日益依賴數位教育的後疫情時代，這些應用程式有望增強傳統的教學方法。然而，隨著它們進入美國市場，面臨著數據隱私和教育方法文化差異的挑戰。儘管存在這些困難，這些AI應用程式日益增長的影響力顯示出教育傳遞和體驗方式的變革性轉變。

[閱讀更多](#)

# AWS 強化負責任的生成式 AI 相關計畫

[生成式 AI](#) [安全性](#) [隱私措施](#) [透明度](#) [錯誤資訊](#) [政府合作](#) [健康照護](#) [氣候變遷](#)

2024-07-10



## AWS 強化負責任的生成式 AI 相關計畫

亞馬遜網路服務 ( AWS ) 持續優先考量安全及負責任的生成式 AI 發展，並宣布在其工具和實踐方面取得重要進展。主要創新包括為 Amazon Bedrock 推出的 Guardrails，該功能允許用戶針對其生成式 AI 應用實施量身訂做的安全和隱私措施，阻擋高達 85% 的有害內容。

此外，AWS 還推出了 AI 服務卡，這是一種清晰的文件，概述了他們 AI 服務的預期用途及限制，促進透明度。為了對抗錯誤資訊，所有由 Amazon Titan 圖像生成器生成的圖像現在都附有一個隱形的、抗篡改的水印。

AWS 也在促進科技公司與政府之間的合作，以增強 AI 的安全性和信任度。美國人工智慧安全研究所聯盟等倡議旨在開發確保 AI 技術安全部署的方法論。總體而言，AWS 致力於將生成式 AI 作為促進良善的力量，解決健康照護和氣候變遷等領域的挑戰。

[閱讀更多](#)

# 儘管高管信心十足，AI 採用仍面臨障礙

AI | 高管 | 技術挑戰 | 預算限制 | 投資回報率 | 軟體開發

2024-07-11



## 儘管高管信心十足，AI 採用仍面臨障礙

Zartis 最近的一項研究揭示了英國科技行業內的一個有趣悖論：雖然 85% 的高管認為他們的員工擁有強大的 AI 技能，但重重挑戰卻阻礙了廣泛採用。目前有高達 94% 的公司已經在實施某種形式的 AI，但預算限制 (41%)、AI 人才短缺 (38%) 和技術複雜性 (35%) 等障礙仍持續阻礙進展。

整合挑戰和對投資回報率 (ROI) 的擔憂也相當普遍，分別有 44% 和 42% 的高管將這些列為主要問題。儘管面臨這些障礙，93% 的組織仍在重金投資 AI，重點放在軟體開發上，顯示出對 AI 潛在長期利益的日益認識。

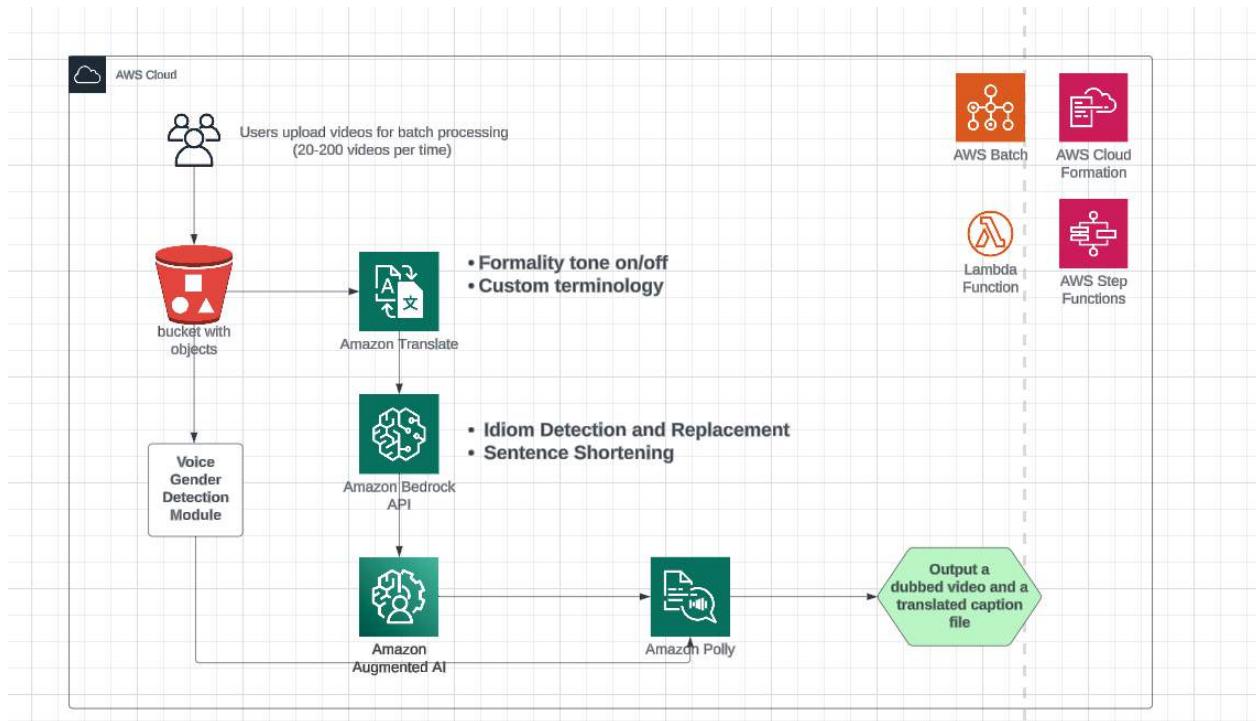
隨著 AI 生態的演變，英國科技高管渴望利用新技術，同時應對整合和成本效益的複雜性。

[閱讀更多](#)

# 利用 AWS 改變影片配音的遊戲規則

AWS 影片配音 Amazon Translate Amazon Bedrock Amazon Polly 媒體與娛樂 自動化

2024-07-15



## 利用 AWS 改變影片配音的遊戲規則

在媒體與娛樂產業中，亞馬遜推出了一項創新的影片自動配音解決方案，這項技術運用了 Amazon Translate、Amazon Bedrock 和 Amazon Polly 的能力，為傳統配音所面臨的高成本和耗時的挑戰提供了有效的解決方案。

這個過程首先由 Amazon Translate 提供快速且高品質的影片字幕翻譯。接著，Amazon Bedrock 進一步增強這些翻譯，整合了成語辨識等功能，以確保文化差異得以保留，並採用自動縮句演算法來改善音訊與影片之間的同步性。

此外，該解決方案還允許自訂術語和調整正式性語調，使其能夠適應不同類型的內容。這個全面的流程不僅簡化了配音過程，還大幅降低了成本，讓創作者能夠有效擴大其全球影響力。

[閱讀更多](#)

# Google Arts & Culture 推出四款創新遊戲，讓你在夏季探索藝術

**Google Arts & Culture** | 互動體驗 | AI | 藝術遊戲 | 音樂 | 藝術導覽 | 藝術作品 | 虛擬藝術空間

2024-07-17



Google Arts & Culture 推出四款創新遊戲，讓你在夏季探索藝術

今年夏天，Google Arts & Culture 推出四款令人興奮的互動體驗，旨在提升你與藝術和文化的互動。

1. 一聲音，兩畫面：這款遊戲結合了 AI 生成的音樂與藝術作品，挑戰玩家將正確的聲音與視覺作品匹配。
2. 一分鐘導覽：根據你的興趣，這些 AI 驅動的導覽提供關於主要藝術運動的快速見解，讓藝術欣賞變得更容易和有趣。
3. 藝術是什麼？：在這款多人遊戲中，玩家描繪著名的藝術作品，而其他人則競速猜測這些作品是什麼，融合了創意與競爭樂趣。
4. 每日藝廊：用每日新上架的藝術作品和家具來策劃自己的虛擬藝術空間，讓創意和發現持續進行。

所有體驗均可在 Google Arts & Culture 網站及應用程式的「遊玩」標籤中找到，並支援 Android 和 iOS 系統。快來探索吧！



# 教育中的生成式人工智能：來自家長和學生的見解

生成式人工智能 教學 個人化學習 即時回饋 深化理解

2024-07-18



## 教育中的生成式人工智能：來自家長和學生的見解

最近，Google的研究團隊與教育界的討論揭示了生成式人工智能如何正面影響學習體驗。以下是家長和學生分享的三個關鍵見解：

1. 即時回饋：生成式人工智能為家長提供了即時的孩子學業進展見解，使得及時介入和量身訂做的支持成為可能。
2. 深化理解：人工智能工具能透過提供補充資訊來增強課堂學習，讓學生能更全面地探索不熟悉的主題。
3. 個人化學習路徑：透過分析個別學習模式，生成式人工智能能為不同認知能力的學生量身定製教育內容，確保每個人都能獲得適當的支持。

雖然生成式人工智能在改變教育方面顯示出潛力，但強調了教師的基本角色。人工智能的設計是為了補充，而不是取代人類教育者，讓他們能更專注於培養課堂上的好奇心與參與感。

[閱讀更多](#)

# Google 與美國隊及 NBCUniversal 合作，為巴黎 2024 奧運會提供報導

**Google 巴黎 2024 奧運會 NBCUniversal AI 搜尋 Google Maps Gemini 運動 運動員**

2024-07-18



Google 與美國隊及 NBCUniversal 合作，為巴黎 2024 奧運會提供報導

在這項激動人心的合作中，Google、美國隊和 NBCUniversal 將共同提升巴黎 2024 奧運及殘奧會的體驗。整合 Google 的技術，如搜尋、Google Maps Platform 和 Gemini，將使觀眾能深入了解各項賽事，並探索美麗的巴黎市。

在轉播過程中，觀眾將受益於 Google 搜尋中的 AI 概述，提供快速的運動及運動員資訊。NBC 的「首席超級粉絲評論員」Leslie Jones 將利用 Gemini 與內容互動，揭開奧運會的迷人細節。美國奧運選手將透過 Google Lens 和 Google Maps 的沉浸式視圖，分享他們在巴黎的冒險，展現如凡爾賽宮和水上中心等標誌性地標。

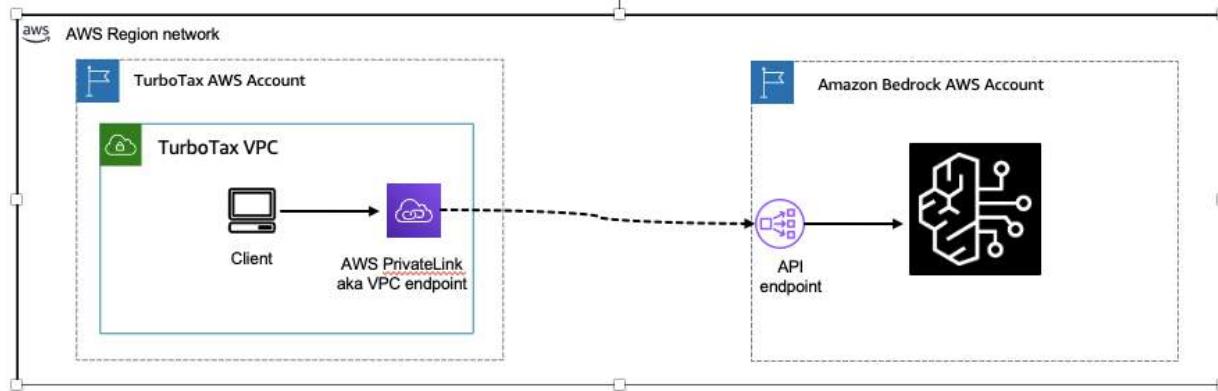
從 7 月 26 日的開幕典禮開始，敬請鎖定這項創新報導！

[閱讀更多](#)

# Intuit 利用 Amazon Bedrock 和 Claude 簡化 TurboTax 的報稅流程

Intuit | Amazon Bedrock | Claude | TurboTax | 報稅流程 | 稅務計算 | 用戶體驗 | 數據保護

2024-07-30



## Intuit 利用 Amazon Bedrock 和 Claude 簡化 TurboTax 的報稅流程

在簡化報稅體驗的重要進展中，Intuit 正在其 TurboTax 平台上運用 Amazon Bedrock 和 Anthropic 的 Claude 語言模型。隨著稅法規定日益複雜，Intuit 旨在讓數百萬用戶能更輕鬆理解稅務計算。

在 2024 年的報稅季，這項合作使 TurboTax 能夠提供清晰且具上下文的稅務計算解釋，並支持即時準確度檢查。透過整合 Claude，Intuit 的財務助手能夠幫助用戶解釋複雜的稅務概念，從而增強用戶對於報稅的信心。

這項夥伴關係強調可擴展性和安全性，使 Intuit 能夠在報稅季的高峰期管理用戶的激增，同時確保數據保護。這種創新方法不僅意在簡化報稅流程，還為未來財務管理解決方案的進步鋪平道路，彰顯了 Intuit 對於利用尖端技術以提升客戶體驗的承諾。

[閱讀更多](#)