

OPENAI 推出 GPT-4O：多模態 AI 的新時代

OPENAI

GOOGLE DEEPMIND 和 ISOMORPHIC LABS 發布 ALPHAFOOLD 3 AI 模型

GOOGLE

NVIDIA BLACKWELL 平台減少最多 25 倍成本和能源

NVIDIA

人工智慧技術月報 AI TRENDS

Artificial Intelligence Technology Monthly Report

目錄

精選文章

-
- Google DeepMind 和 Isomorphic Labs 發布 AlphaFold 3 AI 模型 9
 - NVIDIA Blackwell 平台減少最多 25 倍成本和能源 10
 - OpenAI 推出 GPT-4o：多模態 AI 的新時代 12
-

模型技術

-
- Microsoft Research 革新 LLM 效率與 FastGen 15
 - LotteON 的動態 A/B 測試革新了個性化推薦 16
 - 加強版的 VASIM 自動擴展：雲資源管理的遊戲改變者 17
 - MatterSim: 真實材料模擬的突破 18
 - 解鎖 AI 的力量：結合 RAG 和 Voyage AI 在 AWS 上的應用 19
 - 使用 Amazon Titan Text Premier 和 AWS CDK 構建生成 AI 應用程式 21
 - NVIDIA 和 Google DeepMind 合作推進 AI 創新 23
 - Google I/O 2024：揭示雙子時代與 AI 進展 25
 - 令人興奮的 Google Labs 更新：VideoFX、ImageFX 和 MusicFX 26
 - Veo 和 Imagen 3：Google 最新的生成式 AI 工具 27

● Google 的 Gemini 和 Gemma 模型的令人興奮的更新	28
● Google 推出 LearnLM 和 Illuminate：邁向負責任 AI 與增強學習的新突破	29
● RASCAL：革新存儲和檢索系統	30
● UAE 發表 Falcon 2 AI 模型，在科技領域瞄準高點	31
● 使用 ONNX Runtime 在 AWS Graviton 處理器上加速 NLP 推理	33
● 手術機器人的革命性進展：ORBIT-Surgical 框架	34
● Basecamp Research：利用AI編目地球生物多樣性	36
● Google I/O 2024：創新盛宴	38
● Google Gemini AI 模型的故事	39
● Microsoft Research 發表突破性創新	40
● LotteON 利用 Amazon SageMaker 和 MLOps 提升個人化推薦系統	41
● Microsoft Research Africa 的公平 AI 之旅	43
● Mixtral 8x22B 在 Amazon SageMaker JumpStart 上設立了機器學習的新標準	44
● WindSeer：以即時風力預測提升無人機能力	46
● 評估你的 LLM 聊天機器人：關鍵考量	47
● AWS 推出具備先進 RAG 和 Amazon Bedrock 的多模態助手	48
● Amazon SageMaker 提供高效的多租戶 LoRA 服務	49
● NVIDIA 在 Green500 中以能源效率超高的超級電腦稱霸	51
● NVIDIA 超級增強 RTX AI 電腦的新性能優化	53
● Azure AI Studio 中的新生成式 AI 評估和監控工具	55
● 推出 JAIS：Azure 上以阿拉伯語為中心的大型語言模型	57
● OpenAI 與 Reddit 合作進行即時數據整合	58
● GigaPath：數位病理學的突破	60

● 加速在 Amazon SageMaker 上使用專家並行進行 Mixtral 8x7B 的預訓練	62
● Atera 的 AI 演進：革新 IT 管理	64
● NVIDIA 在 COMPUTEX 最佳選擇獎大放異彩	65

資訊安全

● HiddenLayer Model Scanner：提升 AI 模型安全性	68
-----------------------------------------	----

應用

● 釋放生成式 AI 的能量：Verisk 的即時洞察引擎提升客戶支持	70
● 建立 AI/ML 卓越中心：企業的遊戲規則改變者	72
● AWS DeepRacer：人人都能接觸的機器學習入門	73
● 使用 Amazon SageMaker Canvas 和 JumpStart 進行無程式碼 LLM 微調，提升客戶互動	75
● AWS 引入生成式 AI 用於臨床報告摘要	76
● NVIDIA 和 Recursion 加速藥物發現與 AI 超級電腦	77
● AWS 引入混合人機工作流程於生成式 AI 應用	79
● Cerence 使用生成式 AI 轉變車內體驗	81
● NVIDIA 與日本合作推動生成式 AI 基礎建設	83
● 釋放 Google AI 在 Android 上的威力	85
● 3 種利用 Gemini for Google Workspace 提升生產力的新方法	86

● 介紹 Google 的 LearnLM：教育中的生成式 AI 革命	87
● 四位藝術家與 Imagen 2：通過 AI 重新想像奇幻世界	88
● 使用 Amazon Bedrock 自動化考試創建	89
● Amazon Ads 推出由 Amazon SageMaker 支持的 AI 生成圖像功能	91
● 機器人的未來：AI 正在改變產業	92
● Mozilla Firefox 增加 AI 驅動的 RTX Video 以提升串流品質	94
● Google Cloud AI 如何幫助發掘英格蘭的下一個足球明星	96
● Microsoft 在 CHI 2024 大放異彩：引領以人為本的設計	97
● GAAD 2024：Google 發表 AI 驅動的無障礙功能	98
● Google AI：Gemini 透過 Workspace 整合提升教育	100
● 智能自動化崛起為戰略差異化因素	102
● IBM 和 Tech Mahindra 發布可信賴的 AI 與 Watsonx	103
● 利用 AWS 建立高效能 AI 工作流程：提示鏈結與人機互動指南	104
● Google 支持中東歐 AI 發展的新舉措	105
● 探索 AI 的未來：介紹 AI Parabellum	106
● AI 是否正在改變賭博業？	107
● NVIDIA 和 Dell 發布突破性 AI 創新	108
● Microsoft 推出具備先進 AI 功能的 Copilot+ 電腦	109
● 使用 Amazon Bedrock 提升新聞業的生成式 AI	111
● NVIDIA 與 Microsoft 聯手加速 AI 發展	113
● NVIDIA 五月 Studio 驅動程式革新內容創作	115
● 加速 GenAI 應用程式開發與增強追蹤和除錯	117

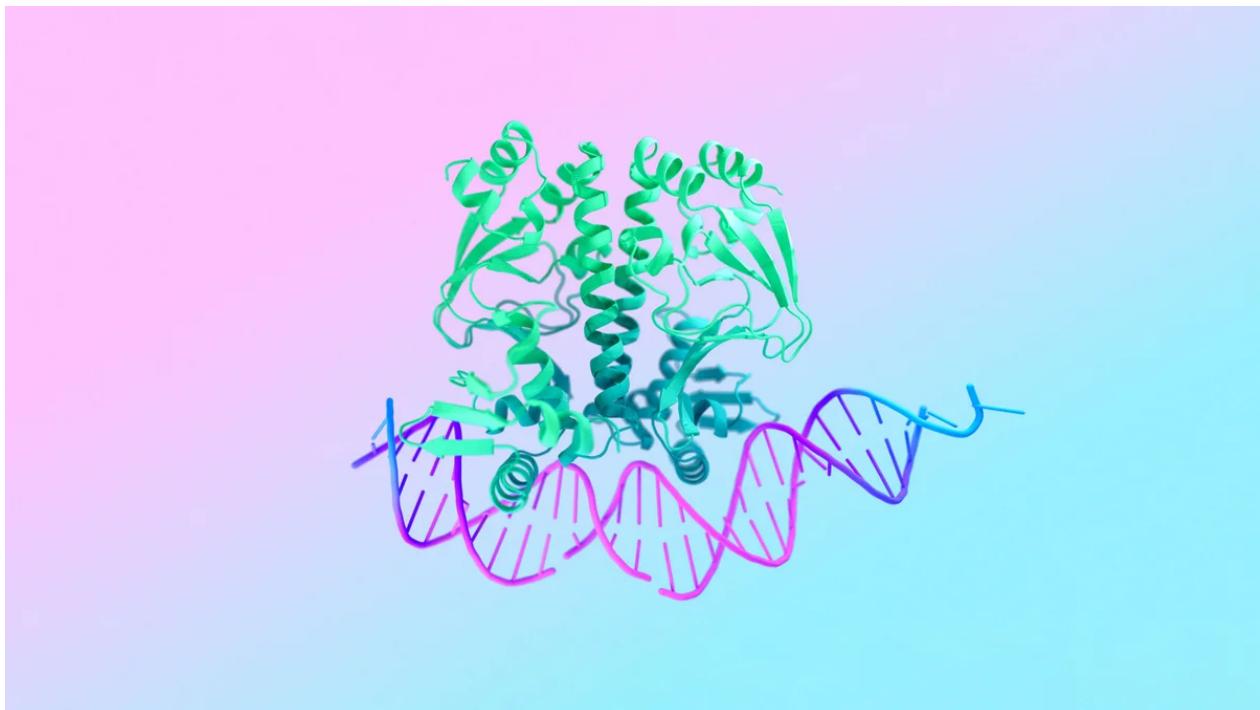
● 釋放 AI 潛力：Azure AI MaaS 與 ClearML、Arize AI 和 Dataloop 的整合提升 LLM 模型的可及性	118
● Microsoft 與 Hugging Face 提升生成式 AI 合作	119
● Azure AI Studio 與變革性的AI開發	120
● Microsoft 公布 AI 應用程式範本以簡化 AI 開發	121
● 使用視覺語言模型和 Amazon Bedrock 自動化時尚產品描述	122
● SoftServe 和 Continental 透過 OpenUSD 和生成式 AI 推動數位化	124
● 三項改變職場的 Google AI 功能在 Android 上	125
● 為人們設計 AI：與 Abigail Sellen 的見解	127
● 革命化醫療保健：雲端運算和人工智慧的影響	128
● Azure AI Studio：協作 AI 開發的新時代	129

01 精選文章

Google DeepMind 和 Isomorphic Labs 發布 AlphaFold 3 AI 模型

[Google DeepMind](#) [Isomorphic Labs](#) [AlphaFold 3](#) [AI模型](#) [分子結構預測](#) [生物學](#) [藥物發現](#) [基因組學](#) [作物抗逆性](#) [蛋白質結構](#)

2024-05-08



Google DeepMind 和 Isomorphic Labs 推出了他們最新的 AI 奇蹟，AlphaFold 3，這個模型可以以驚人的準確度預測所有生命分子的結構和相互作用。這個模型深入探究蛋白質、DNA、RNA 和配體的複雜舞蹈，展示了比現有方法提高 50% 的精確度。

AlphaFold 3 有望革新我們對生物學和藥物發現的理解。科學家可以通過新推出的 AlphaFold 伺服器免費訪問其功能。這個平台允許研究人員輕鬆地建模複雜的分子結構，大大加快了科學工作流程。

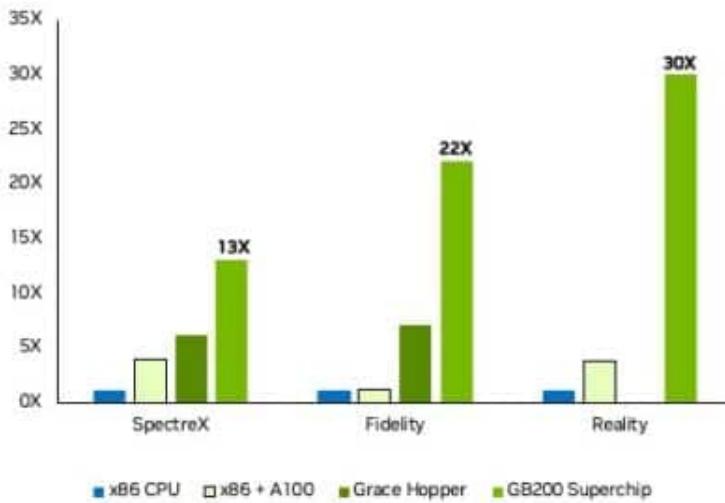
基於 AlphaFold 2 的成功，AlphaFold 3 將其預測能力擴展到蛋白質之外，涵蓋更廣泛的生物分子，可能改變基因組學、作物抗逆性和藥物設計等領域。通過讓這個工具在全球範圍內可及，Google DeepMind 旨在加速發現並促進對生命分子基礎的更深入理解。

[閱讀更多](#)

NVIDIA Blackwell 平台減少最多 25 倍 成本和能源

NVIDIA	Blackwell 平台	科學計算	人工智慧	量子計算	數據分析	生成式 AI	雙精度計算
GB200 Grace Blackwell 超級晶片	電路模擬	流體動力學	Quantum-X800 InfiniBand 網路平台				
數據傳輸	能源消耗						

2024-05-13



NVIDIA 推出了突破性的 Blackwell 平台，這是科學計算領域中的一個奇蹟。這個平台在人工智能、量子計算和數據分析等領域帶來重大進展。

NVIDIA Blackwell 平台支援生成式 AI 用於萬億參數模型，與以前的架構相比，成本和能源消耗最多可減少 25 倍。這意味著 AI 工作負載可以更高效地處理，影響藥物發現和融合能源研究等領域。

一個突出的特點是其在雙精度 (FP64) 計算中的增強性能，這對於基於物理的模擬至關重要。這些模擬對從飛機到藥物的產品設計和開發至關重要。

此外，該平台的 GB200 Grace Blackwell 超級晶片結合了 CPU 和 GPU，用於更快的電路模擬和流體動力學計算。與傳統 CPU 系統相比，速度提升可達 22 倍。

另外，NVIDIA Quantum-X800 InfiniBand 網路平台顯著提升數據傳輸速度並降低能源消耗，使其成為現代科學研究的基石。

有了這些創新，Blackwell 平台將加速科學發現並推動全球可持續計算的努力。

[閱讀更多](#)

OpenAI 推出 GPT-4o：多模態 AI 的新時代

模型技術 | 多模態 | 人工智慧 | 自然語言處理

2024-05-13



OpenAI 推出其最新旗艦模型 GPT-4o，該模型整合了文本、音頻和視覺輸入和輸出。這一創新模型旨在提升機器交互的自然性和流暢性。

主要特點和功能

多模態整合 GPT-4o，其中 'o' 代表 'omni'，支持更廣泛的輸入和輸出模式。用戶可以輸入任意組合的文本、音頻和圖像，並接收這些形式的輸出。該模型的響應時間極快，平均為 320 毫秒，接近人類對話速度。

開創性進展 GPT-4o 通過單一神經網絡處理所有輸入和輸出，實現了前所未有的飛躍。這種統一方法保留了上下文和細微差別，克服了以前使用不同管道的模型的局限性。它在複雜任務中的性能得到了提高，如歌曲和諧、實時翻譯和生成表達性音頻輸出。

多功能和性能 該模型在各種基準測試中表現出色，尤其是在非英語語言、編碼任務和一般知識推理方面。它在音頻和翻譯任務中超越了以前的最先進模型，展示了優越的多語言、音頻和視覺能力。

行業觀點 Superintelligent 創始人兼 CEO Nathaniel Whittemore 指出，GPT-4o 作為真正的多模態模型意義重大，而不僅僅是文本系統的附加。他強調了廣泛用例的潛力，儘管他承認在評估如此先進產品的實際影響方面存在初期挑戰。

結論 GPT-4o 樹立了多模態 AI 的新標準，提升了交互的自然性並擴大了 AI 應用的範圍。其在單一模型中整合文本、音頻和視覺處理標誌著 AI 技術的一個里程碑，承諾提供更多包容性和多功能的解決方案。

[閱讀更多](#)

02 模型技術

Microsoft Research 革新 LLM 效率與 FastGen

Microsoft Research | FastGen | LLM | 記憶體管理 | KV 快取 | 自適應壓縮 | ICLR 2024 | 人工智慧

2024-05-08



在大型語言模型（LLMs）的領域中，高效管理記憶體至關重要。Microsoft Research 推出的 FastGen 是一項突破性的創新，旨在優化關鍵值（KV）快取系統。KV 快取對於儲存先前計算的數據至關重要，使 LLMs 能更快地生成回應。然而，它也會消耗驚人的記憶體，有時一個操作就能佔用高達 320 GB 的記憶體。

FastGen 的目標是在不影響性能的情況下將記憶體使用量減半。怎麼做到的呢？透過一種稱為自適應 KV 快取壓縮的方法。通過對各種 LLM 模組進行詳細的剖析，FastGen 能夠識別並丟棄不必要的數據，將記憶體資源集中在最相關的信息上。這種方法確保 LLMs 保持快速且高效。

這項創新在第12屆國際學習表徵會議（ICLR 2024）上展示，承諾將使先進的 AI 工具變得更易於使用且更具成本效益。FastGen 代表了人工智慧領域的一大進步，開啟了更廣泛應用和更高效資源使用的大門。

[閱讀更多](#)

LotteON 的動態 A/B 測試革新了個性化推薦

動態 A/B 測試 | 個性化推薦 | 多臂機器人 | MAB 演算法 | Thompson 取樣 | AWS | Amazon SageMaker | AWS Lambda | 點擊率 | 模型性能

2024-05-09

The image displays three vertical screenshots of the LotteON mobile application. Each screenshot shows a different variation of a product recommendation screen. The top navigation bar includes 'ON' (with a dropdown arrow), a search bar, and a red notification badge. Below the navigation are tabs: '특가' (Sales), '백화점' (Department Store), '장보기' (Grocery), '추천' (Recommendation), '선물하기' (Gift), '포함비주간' (Incl. Non-Weekend), 'DIOR' (Designer), and '이벤' (Event). The first two screenshots show a '고객님을 위해 준비한 오늘의 쇼핑 추천' (Recommendations for you today) section with various products like Vitamix blenders. The third screenshot shows a '최근 본 상품과 비슷한 상품' (Similar products to recently viewed items) section with beauty products like Dior lipsticks.

LotteON 的動態 A/B 測試革新了個性化推薦

LotteON 通過實施動態 A/B 測試系統，在提升客戶體驗方面邁出了重要一步。這種創新的方法實時評估用戶互動，例如產品點擊，從各種選項中動態選擇最有效的推薦模型。

動態 A/B 測試使用多臂機器人 (MAB) 演算法，通過根據每個模型的成功率調整曝光率來優化模型性能。具體來說，LotteON 採用了 MAB 演算法的 Thompson 取樣變體。此方法利用歷史數據來調整和選擇能夠提高點擊率 (CTR) 的模型。

通過利用 AWS 服務，包括 Amazon SageMaker 和 AWS Lambda，LotteON 簡化了過程，減少了傳統 A/B 測試所需的時間和資源。這使得 LotteON 的數據科學家能夠更多地專注於模型開發，減少操作上的負擔。

此轉變為動態 A/B 測試使實時、數據驅動的決策成為可能，顯著改善了推薦的準確性和客戶滿意度，同時提供了有價值的模型性能洞察。

[閱讀更多](#)

加強版的 VASIM 自動擴展：雲資源管理的遊戲改變者

VASIM | 自動擴展 | 雲端運算 | 演算法 | Microsoft Research | 資源管理 | 優化

2024-05-13



Microsoft Research 推出了一個創新的工具，稱為 Vertical Autoscaling Simulator Toolkit (VASIM)，旨在徹底改變自動擴展演算法的開發和優化方式。自動擴展對於雲端運算至關重要，能夠根據需求動態調整計算資源，從而確保高效能和成本管理。

VASIM 的特別之處

VASIM 是一個尖端的模擬工具，在著名的 IEEE International Conference on Data Engineering (ICDE 2024) 上發表。與現有工具不同，VASIM 專注於自動擴展演算法中的關鍵推薦元件。它根據使用數據建議最佳的擴展行動，有效平衡效能和成本。這種獨特的方法使開發者能夠快速迭代，提升演算法效能，同時減少資源浪費。

VASIM 的主要特點

1. 用戶友好的介面：VASIM 簡化了自動擴展政策的評估。使用者可以上傳歷史數據、定義政策並快速運行模擬。
2. 資源效率：通過消除實時測試的需求，VASIM 加速了開發週期並降低成本。
3. 多目標優化：VASIM 應用 Pareto 優化技術來平衡關鍵指標如餘量和 CPU 使用率。
4. 多功能性：該工具允許在各種工作負載上測試演算法，證明其在不同情境中的可靠性。

未來展望

Microsoft 計劃擴展 VASIM，包括記憶體和網路頻寬等其他資源，提供一個全面的系統性能優化工具。

VASIM 有望成為開發者尋求高效優化雲資源管理的必備工具。

[閱讀更多](#)

MatterSim: 真實材料模擬的突破

MatterSim | 深度學習 | 材料模擬 | 生成模型 | 分子動力學 | 奈米技術 | 儲能 | 可持續性

2024-05-13



在尋求對奈米電子學、儲能以及醫療保健至關重要的先進材料時，預測材料屬性在創建材料之前是一個重大挑戰。Microsoft Research 推出了 MatterSim，一個深度學習模型，正面解決這種複雜性。

MatterSim 利用深度學習準確地模擬和預測材料在廣泛元素、溫度和壓力下的屬性。與傳統方法不同，MatterSim 從量子力學的基本原理出發，涵蓋0到5000開爾文和高達1000萬個大氣壓的條件。這使它能夠處理各種材料—金屬、氧化物、硫化物、鹵化物—在不同狀態下如晶體、非晶態固體和液體。

MatterSim 的創新在於其結合了主動學習、生成模型和分子動力學模擬。這使得能量、原子力和應力的預測極其準確，顯著超越以往的模型準確度。

該模型在特定設計任務中的適應性，如以顯著減少數據來優化水的性質模擬，強調了其高效性。MatterSim 也跨越了原子模型和現實世界測量之間的鴻溝，提供了無需複雜模擬的強大預測。

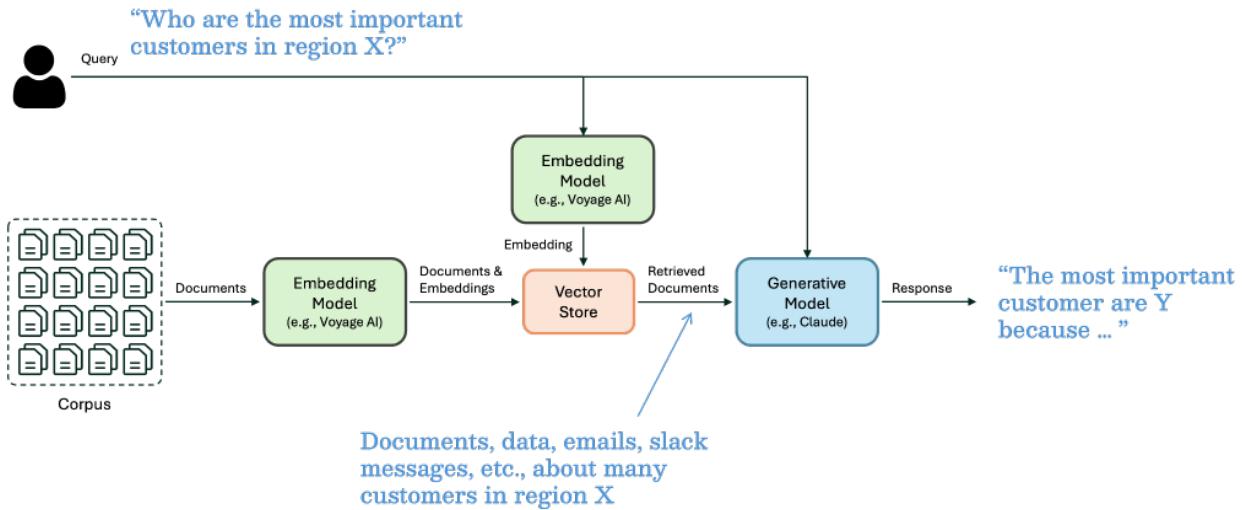
展望未來，MatterSim 旨在通過與生成式 AI 和強化學習的整合，革新可持續性、儲能和奈米技術等領域，開創材料設計和開發的新時代。

[閱讀更多](#)

解鎖 AI 的力量：結合 RAG 和 Voyage AI 在 AWS 上的應用

RAG | Voyage AI | Amazon SageMaker JumpStart | Anthropic Claude 3 | 嵌入模型 | Amazon Bedrock | Amazon OpenSearch Service

2024-05-14



在一項突破性的發展中，將檢索增強生成（RAG）架構與 Voyage AI 的嵌入模型整合到 Amazon SageMaker JumpStart 和 Anthropic Claude 3 模型上，使得 AI 更加智慧且具有上下文理解能力。以下是這些創新如何改變 AI 應用領域的方式。

RAG 結合了兩個世界的最佳優勢：大型數據庫的龐大信息資源和 AI 模型的生成能力。這種混合方法在生成回應時動態檢索相關數據，使 AI 模型能夠生成精確且具有上下文豐富的輸出。

嵌入模型是這個架構的核心，它將大量文本數據轉換為緊湊的數值表示。這些模型顯著提高了數據檢索的準確性，確保 RAG 的生成組件能接收到最相關的信息。

Voyage AI 提供最先進的嵌入模型，這些模型可在 Amazon SageMaker JumpStart 上獲得。像 voyage-large-2 這樣的模型優化了檢索質量，而像 voyage-code-2 這樣的領域特定模型在專業領域（如代碼檢索）中表現卓越。

這種強大的組合在包括 Amazon Bedrock 和 Amazon OpenSearch Service 在內的 AWS 服務上實施，創建了一個強大的基礎設施來部署和擴展 AI 解決方案。通過 SageMaker JumpStart，這些模型變得易於訪問，使企業能夠快速將先進的 AI 能力整合到他們的應用中。

這種技術協同標誌著在使 AI 在各種應用中更可靠和有用方面的一次重大飛躍，從通用查詢到法律和金融數據檢索等專業領域。隨著這些尖端工具在我們手中，AI 的未來看起來更加光明。

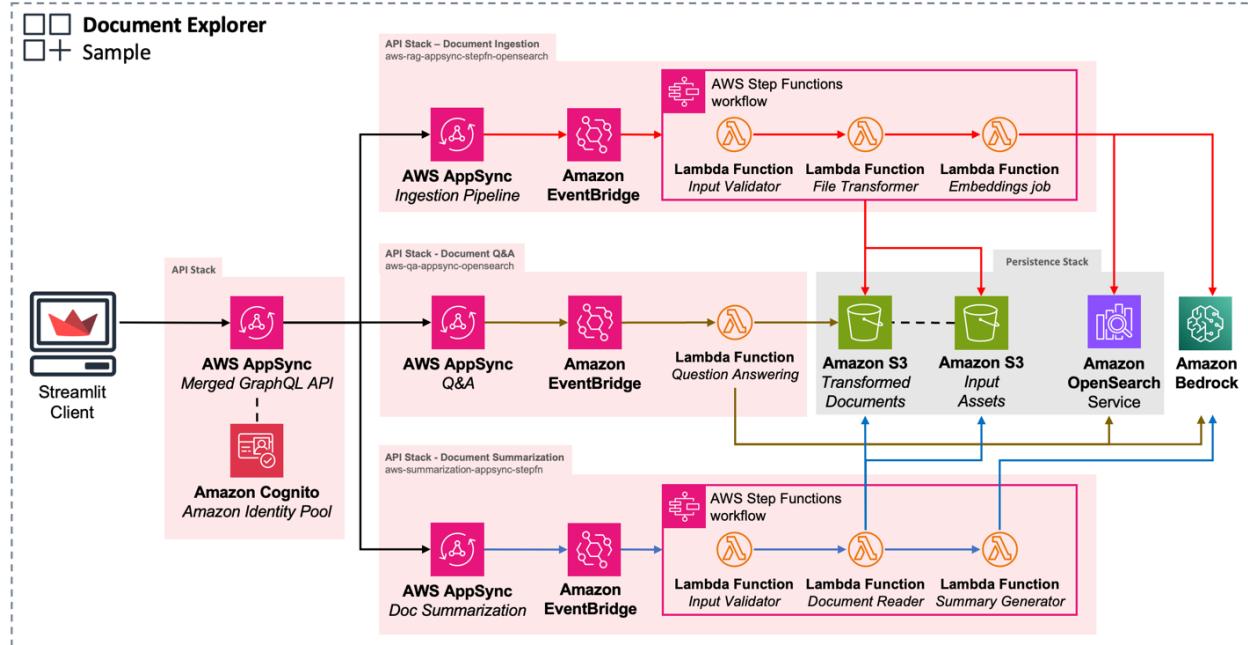
閱讀更多

使用 Amazon Titan Text Premier 和 AWS CDK 構建生成 AI 應用程式

Amazon Titan Text Premier **AWS CDK** **Amazon Bedrock** **生成 AI** **基礎模型** **企業級任務** **RAG**

聊天助手 | 文檔摘要 | 問答

2024-05-14



Amazon 推出了其在 AI 領域的最新創新——Amazon Titan Text Premier，這是 Amazon Titan 大型語言模型 (LLMs) 家族的一個強大補充。透過 Amazon Bedrock (一個完全託管的服務) 提供，這個模型旨在簡化生成 AI 應用程式的開發。

Amazon Bedrock 提供來自頂尖 AI 公司的一系列高效能基礎模型，這些模型透過單一 API 即可訪問。這使開發人員更容易將先進的 AI 功能整合到他們的應用程式中，同時注重安全性和負責責任的 AI 實踐。

Amazon Titan Text Premier 在多種與文本相關的任務中表現優異，包括摘要生成、文本生成、分類、問答和資訊提取。它特別針對企業級任務進行了優化，例如檢索增強生成 (RAG) 和基於代理的應用程式。

為了幫助加速開發，Amazon 提供了 AWS Generative AI CDK Constructs，這些構件透過可重用的基礎設施模式簡化了部署過程。介紹了兩個示例應用程式：1. Document Explorer，處理數據攝取、文檔摘要和問答。2. Amazon Bedrock Agent and Custom Knowledge Base，一個使用 RAG 進行文獻相關查詢的聊天助手。

Amazon Titan Text Premier 現已在美國東部（北維吉尼亞）地區提供。請關注 Amazon Titan 產品頁面以獲取更新，並查看 AWS Generative AI CDK Constructs GitHub 存儲庫以獲取更多詳細信息。

[閱讀更多](#)

NVIDIA 和 Google DeepMind 合作推進 AI 創新

NVIDIA | Google DeepMind | LLM | Gemma 2 | PaliGemma | VLM | NIM | RAPIDS cuDF | Google
Colab | Tensor Core | Genkit | Firebase | AI

2024-05-14



Explore What's Next in AI With the Best of GTC

Watch On Demand

NVIDIA 和 Google DeepMind 已經聯手推進大規模語言模型 (LLM) 技術。這次合作預計將在生成式 AI 領域產生重大影響，這種技術使用能夠處理文本、圖像和聲音的模型。

介紹 Gemma 2 和 PaliGemma 模型

這次合作的核心是兩個新的模型：Gemma 2 和 PaliGemma。Gemma 2 是原始 Gemma 模型的進化版本，擁有全新的架構，旨在廣泛應用，提供卓越的性能和效率。

PaliGemma 是一個開放式的視覺-語言模型 (VLM)，專為結合視覺和文本數據的任務而設計。它在圖像標註、視覺問答和物體檢測等活動中表現出色。基於 SigLIP 視覺模型和 Gemma 語言模型，PaliGemma 承諾提供頂級的微調性能。

這兩個模型都受益於 NVIDIA NIM 推理微服務，促進大規模 AI 部署。開發者可以通過 API 目錄訪問 PaliGemma，NVIDIA NGC 和 GitHub 上的容器發布即將推出。

在 Google Colab 上加速數據分析

另一個令人興奮的發展是，開源 GPU 數據框架庫 RAPIDS cuDF 現已在 Google Colab 上默認支持。這種整合允許用戶無需更改代碼，即可使用 NVIDIA L4 Tensor Core GPUs 將 Python 工作流程加速多達 50 倍。

使用 Genkit 提升應用開發

NVIDIA 和 Google 還推出了 Firebase Genkit，使開發者能夠輕鬆將生成式 AI 模型整合到網頁和移動應用中。這個工具支持 JavaScript，使其對廣大開發者來說更加易於使用。

這些合作強調了 AI 領域的持續創新，為科技世界設定了性能和可及性的新的基準。

[閱讀更多](#)

Google I/O 2024：揭示雙子時代與 AI 進展

Google I/O | AI | Gemini | 多模態 | Trillium TPU | Search | Photos | 雙子時代

2024-05-14



在 Google I/O 2024 大會上，Sundar Pichai 為我們介紹了「雙子時代」，這標誌著人工智慧能力的新世代。核心是 Gemini 模型，設計為原生多模態，這意味著它可以理解和處理文字、圖片、影片、程式碼等多種形式的資料。這種靈活性允許任何輸入轉換為任何輸出。

其中一項突破是 Gemini 1.5 Pro，它能夠處理多達 100 萬個 tokens，相當於數百頁的文字或數小時的音訊。這個擴展的上下文視窗已經被開發者用於諸如除錯程式碼和生成新見解等任務。

Gemini 的能力也在改變 Google 的產品。例如，Google Search 中新的 AI 總覽提供更詳細和上下文感知的回應。在 Google Photos 中，「詢問照片」功能允許進行深入搜索，例如識別汽車的車牌號或追蹤您孩子的游泳進展。

此外，Gemini 1.5 Flash 的推出提供了一個更快、更具成本效益的模型，優化於需要低延遲的任務。Google 還在擴展其 AI 基礎設施，推出新的 Trillium TPU，提供 4.7 倍的計算性能改進。

這些創新正在推動 AI 的邊界，使科技對每個人更有幫助、更易於接觸。

[閱讀更多](#)

令人興奮的 Google Labs 更新： VideoFX、ImageFX 和 MusicFX

[Google Labs](#) [VideoFX](#) [ImageFX](#) [MusicFX](#) [AI 創意](#) [視頻生成](#) [圖像生成](#) [音樂生成](#)

2024-05-14



來自 Google Labs 的令人興奮的更新：VideoFX、ImageFX 和 MusicFX

Google Labs 正在通過三個突破性的工具推動 AI 創意的界限：VideoFX、ImageFX 和 MusicFX。

VideoFX：想像一下，只需一個文字提示就能將你的想法轉換成視頻。這在 VideoFX 的幫助下成為可能，這是 Google DeepMind 最新的視頻生成模型 Veo 驅動的。它捕捉情感細微差別和電影效果，提供一個分鏡模式來構建場景和添加音樂。最初在美國推出，你可以加入候補名單以獲得使用權。

ImageFX：自二月份推出以來，ImageFX 已經成為生成視覺藝術和重現記憶的熱門工具。新的更新增加了編輯控制，允許用戶通過刷過它們來修改特定元素。此外，它還結合了 Imagen 3，這是 Google DeepMind 頂級的圖像生成模型，增強了照片真實感和文本精確度。

MusicFX：釋放你內在的 DJ，利用 MusicFX 的新 DJ 模式。這一功能與藝術家 Jacob Collier 一起展示，讓你可以通過混合不同的流派和樂器來混合節拍，開啟音樂創作的新途徑。

這些工具現在在超過 110 個國家和 37 種語言中可用，旨在使創意過程變得更直觀和便捷。訪問 [labs.google](#) 探索這些創新並開始創作吧！

[閱讀更多](#)

Veo 和 Imagen 3：Google 最新的生成式 AI 工具

Google **Veo** **Imagen 3** **生成式 AI** **影片生成** **文字轉圖片** **數位浮水印** **創意媒體** **負責任的開發**

2024-05-14



在 2024 年的 Google I/O 大會上，Google 令人興奮地展示了兩個突破性的生成式 AI 工具：Veo 和 Imagen 3。這些創新將重新定義創意媒體的格局。

Veo：這個先進的影片生成模型可以創建出具有電影感的 1080p 影片。Veo 能夠理解複雜的指示，使其能夠生成高度詳細且符合用戶願景的影片。只需一個簡單的提示，就可以創建出縮時攝影或鳥瞰鏡頭——Veo 使這一切成為可能。

Imagen 3：作為 Google 最高品質的文字轉圖片模型，Imagen 3 能夠生成具有驚人細節且幾乎無瑕疵的照片級真實圖像。無論您想要一張逼真的狼照片，還是一幅場景的藝術素描，Imagen 3 都能精準地實現。

這些工具在設計時考慮到了負責任的開發，並納入了數位浮水印等安全措施，以確保 AI 生成內容的可信度。目前，Veo 和 Imagen 3 已經在私人預覽中提供，正在開創一個創意的新時代。

[閱讀更多](#)

Google 的 Gemini 和 Gemma 模型的令人興奮的更新

Google | Gemini | Gemma | 模型更新 | 編碼 | 翻譯 | 推理 | API | 視覺語言 | 圖像字幕 | 視覺問答
GPU | TPU

2024-05-14



Google 已經對其 Gemini 1.5 Pro 模型進行了重大更新，並推出了新的 1.5 Flash 模型以及兩個新的 Gemma 模型。這些進步旨在提升各種使用情境下的性能，例如編碼、翻譯和推理。

Gemini 1.5 Pro 已經進行了質量改進，使其在處理複雜任務時更加多功能。Gemini 1.5 Flash 是一個精簡版，針對需要快速回應的任務進行了優化。兩款模型都支持 100 萬標記的上下文窗口，並且可以整合文本、圖像、音頻和視頻。

開發人員現在可以通過 Gemini API 訪問如視頻幀提取和平行函數調用的新功能。此外，Gemini 1.5 Pro 將在私人預覽中提供 200 萬標記的上下文窗口。

Gemma 模型也在擴展。PaliGemma 專為視覺語言任務設計，如圖像字幕和視覺問答。Gemma 2 引入了一個強大的 270 億參數模型，旨在提高 GPU 或單個 TPU 主機上的效率。

這些創新使 Google 的 AI 生態系統對開發人員和研究人員來說更加強大和易於訪問。

[閱讀更多](#)

Google 推出 LearnLM 和 Illuminate：邁向負責任 AI 與增強學習的新突破

Google LearnLM Illuminate 負責任 AI 增強學習 Gemini 模型 學習科學 Search YouTube
Google Classroom 互動教育 Gemini 1.5 Pro 長文脈能力 AI 驅動 SynthID 數位內容真實性

2024-05-14



Google 推出 LearnLM 和 Illuminate：邁向負責任 AI 與增強學習的新突破

Google 持續推動負責任 AI 的界限，推出新防護措施與創新工具，旨在使學習更具吸引力且更易於接觸。

介紹 LearnLM Google 宣布了 LearnLM，一個全新針對學習進行微調的 AI 模型家族。LearnLM 基於 Gemini 模型，結合了研究支持的學習科學，創造個性化且自適應的學習體驗。這項尖端技術嵌入於多種 Google 產品中，如 Search、YouTube 和 Google Classroom，有助於增強學習指導與互動教育影片。

Illuminate：使知識變得易於接觸 另一項突破是 Illuminate，利用 Gemini 1.5 Pro 的長文脈能力，將複雜的研究論文轉化為簡短的音頻對話。Illuminate 生成的 AI 驅動對話總結了關鍵見解，並允許用戶提出後續問題，使高深知識更易於理解。

致力於安全性與可靠性 Google 亦在提升其 AI 模型，透過 AI 輔助紅隊測試與數千名安全專家的反饋。此外，最初用於圖像和音頻的 SynthID 技術，現已擴展至文字和視頻，確保數位內容的真實性。

透過這些發展，Google 旨在負責任地利用 AI 的變革力量，造福社會並推動科學發現。

[閱讀更多](#)

RASCAL：革新存儲和檢索系統

RASCAL | 自動化存儲和檢索系統 | 機器人系統 | 可擴展性 | 高可用性 | 數據中心存儲需求

2024-05-14



Microsoft Research 已經發佈了 RASCAL，一種創新的機器人系統，旨在重新定義自動化存儲和檢索系統 (ASRS)。這項尖端技術承諾具有可擴展性和高可用性，以解決當前 ASRS 框架中的常見挑戰。

RASCAL 是一種無線機器人，運行在間隔均勻的平行貨架和水平軌道上。其模組化設計增強了可擴展性和冗餘性，使其成為管理垂直存儲系統中的小型物品的理想選擇。RASCAL 的靈活性源於數據中心存儲需求的啟發，使其能夠適應各種其他應用。

RASCAL 的主要特點包括：

- 可尋址性：任何機器人都可以訪問貨架上的任何物品。
- 可擴展性：可以輕鬆添加或移除機器人和貨架，且停機時間最少。
- 可用性：單個機器人故障對系統的影響很小。
- 可維護性：不需要專業培訓即可添加或移除機器人。

RASCAL 獨特的軌道系統支持水平和垂直移動，實現獨立和靈活的操作。這一機器人技術的突破強調了在數位時代高效存儲解決方案的重要性。

[閱讀更多](#)

UAE 發表 Falcon 2 AI 模型，在科技領域瞄準高點

UAE | Falcon 2 | AI 模型 | 生成型 AI | Microsoft | 投資 | 教育 | 醫療 | 媒體

2024-05-15



阿拉伯聯合大公國 (UAE) 正在通過阿布達比的科技創新研究所 (Technology Innovation Institute, TII) 介紹的 Falcon 2 系列來提升其人工智慧的競爭力。這款開源生成型 AI 模型將 UAE 定位為全球 AI 領域中的重要競爭者。

Falcon 2 系列包括兩個主要模型：文字模型 Falcon 2 11B 和能夠從圖像生成文字描述的 Falcon 2 11B VLM。此項發展突顯了 UAE 將先進 AI 技術整合到各個領域的承諾。

UAE 的 AI 野心還得到戰略投資和合作的支持，包括 Microsoft 知名的 15 億美元投資。這一舉動強調了該國利用 AI 推動創新和經濟增長的重點。

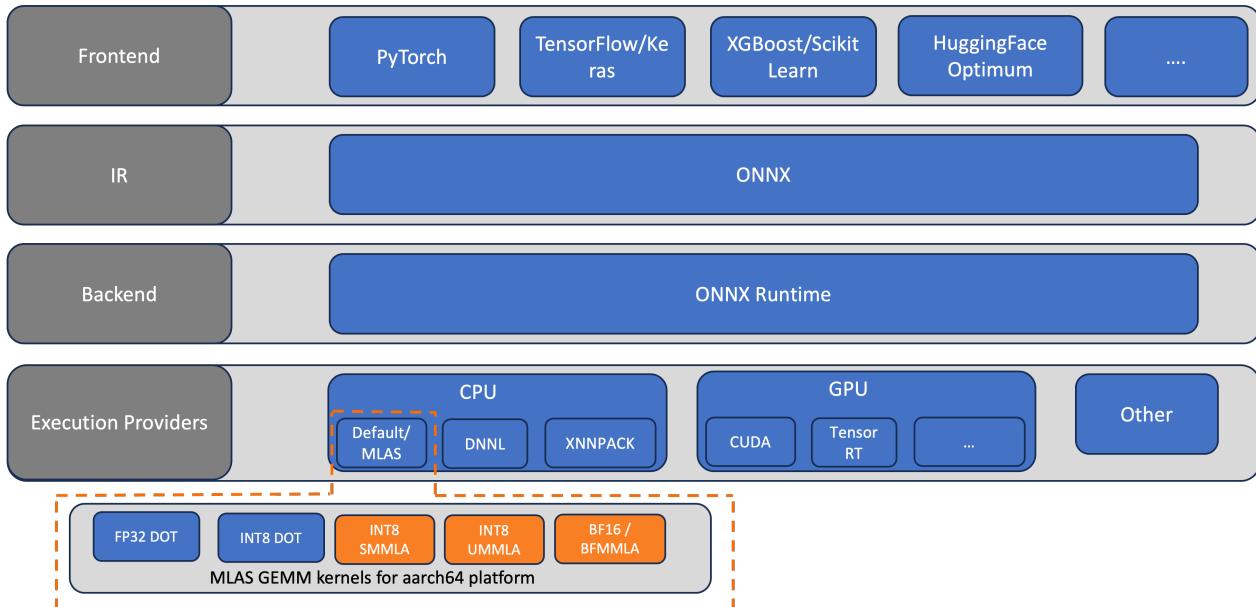
此外，UAE 最近發布的「生成型 AI」指南展示了其在教育、醫療和媒體領域推動 AI 採用的承諾，展示了 100 個面向多元化受眾的實際 AI 用例。

敬請期待，UAE 將繼續在快速發展的科技領域中成為關鍵角色。

使用 ONNX Runtime 在 AWS Graviton 處理器上加速 NLP 推理

NLP | ONNX Runtime | AWS | Graviton3 | 自然語言處理 | 機器學習 | 加速 | 性能 | BERT | RoBERTa
GPT2

2024-05-15



Amazon Web Services (AWS) 已經為自然語言處理 (NLP) 任務引入了一項令人印象深刻的升級，這是透過 ONNX Runtime 在他們的 Graviton3 處理器上實現的。ONNX 是一個開源的機器學習框架，現在透過優化的運行時引擎支持在這些處理器上增強性能。

為什麼這很重要：Graviton3 處理器配備了特殊的指令，例如 bfloat16 和矩陣乘法 (MMLA)，使 NLP 模型的處理速度提高了多達 65% (針對 fp32) 和 30% (針對 int8 量化模型)。這使得文本數據的分析變得更加快速和高效。

主要改進：
 - 優化的核心：ONNX Runtime 利用了新的 bfloat16 和 int8 加速核心，大大提升了推理性能。
 - 基準測試結果：使用像 BERT、RoBERTa 和 GPT2 這樣的模型進行測試，顯示出顯著的吞吐量和延遲改善。

開始使用：要受益於這些優化，用戶可以在 ONNX Runtime 配置中啟用新的快速數學核心，此功能從 1.17.0 版本開始可用。詳細的說明和基準測試腳本可用來指導用戶完成設置過程。

這項升級承諾了更快速和更高效的 NLP 模型部署，充分利用 AWS Graviton3 處理器的潛力。

[閱讀更多](#)

手術機器人的革命性進展：ORBIT-Surgical 框架

ORBIT-Surgical | NVIDIA | 手術機器人 | 強化學習 | 模擬框架 | 手術訓練 | GPU | Omniverse | 數位雙胞胎 | 開源代碼

2024-05-15



Explore What's Next in AI With the Best of GTC

Watch On Demand

NVIDIA 與頂尖的學術研究人員合作，推出了 ORBIT-Surgical，一個開創性的模擬框架，將重新定義手術機器人的培訓。該框架使用 NVIDIA Isaac Sim 和 NVIDIA Omniverse 平台開發，這項創新有望提升手術團隊的能力，並減輕外科醫生的認知負擔。

ORBIT-Surgical 支援十多種腹腔鏡手術所需的複雜操作，例如精確夾持和傳遞針頭。基於物理框架，它利用 NVIDIA 最先進的 GPU 和 Omniverse 進行擬真渲染。這使得強化學習和模仿學習演算法的訓練過程顯著加快和更高效。

該框架具有基準任務，包括單手和雙手操作，如插入分流器和處理縫合針。通過 da Vinci Research Kit 的展示，ORBIT-Surgical 已經證明在模擬中訓練數位雙胞胎可以有效地轉移到真實實驗室環境中的物理機器人。

這種創新方法不僅加速了學習速度，還生成高保真度的合成數據，提高了 AI 模型在實際手術應用中的訓練效果。ORBIT-Surgical 的開源代碼現已在 GitHub 上提供，標誌著機器人輔助手術向前邁出了一大步。

[閱讀更多](#)

Basecamp Research：利用AI編目地球生物多樣性

Basecamp Research | AI | 生物多樣性 | NVIDIA | 深度學習 | 醫學 | 環境保護 | 基因編輯 | 道德數據管理

2024-05-15



Explore What's Next in AI With the Best of GTC

Watch On Demand

Basecamp Research：利用AI編目地球生物多樣性

Basecamp Research 正在進行一項前所未有的任務，規模空前地編目地球的生物多樣性。在人工智慧（AI）的幫助下，他們正在推進醫學和環境保護等領域。在最近一集的 AI Podcast 中，這一集是在 NVIDIA GTC 全球 AI 會議上現場錄製的，Basecamp 的首席技術官 Phil Lorenz 討論了他們的創新方法。

該公司系統地收集與全球自然公園合作的生物多樣性數據。然後他們使用深度學習（一種 AI 技術）來分析這些數據，用於預測蛋白質結構和基因編輯等應用。Lorenz 還強調在他們的工作中，進行道德數據管理的重要性。

協助 Basecamp Research 的關鍵技術進步之一是 NVIDIA 的 AI 硬體和軟體，這增強了他們處理大型數據集並迅速獲得有意義見解的能力。這些突破可能為生物學和醫學帶來重大進展，強調了 AI 在理解和保護我們星球的生物多樣性方面的變革潛力。

[閱讀更多](#)

Google I/O 2024：創新盛宴

[Google I/O 2024](#) [AI](#) [Gemini 1.5 Flash](#) [Gemini 1.5 Pro](#) [Project Astra](#) [Trillium TPU](#) [Imagen 3](#)

[Veo](#) [YouTube Shorts](#) [Google Workspace](#) [Gemini Live](#) [AI 助手](#) [搜尋](#) [Gmail](#) [Docs](#)

2024-05-15



Google I/O 2024 充滿了令人眼花撩亂的公告，展示了 Google 最新的技術進步。以下是一些突出的亮點：

人工智慧突破 Google 介紹了 Gemini 1.5 Flash，一個快速且高效的 AI 模型。同時，增強版 Gemini 1.5 Pro 現在支持 200 萬個 token 的上下文窗口，方便開發者使用。還有未來 AI 助手的願景 Project Astra，以及新款 Trillium TPU，相較於前代，提供了 4.7 倍的性能提升和 67% 的能效改善。

先進的影像與影片生成 Google 的頂級影像生成模型 Imagen 3 現在能創造出更少瑕疵的真實感照片。最新的影片生成模型 Veo 則能製作超過一分鐘的高品質影片，不久將在 YouTube Shorts 上線。

增強的用戶體驗 Gemini 1.5 Pro 也將進入 Google Workspace，幫助總結電子郵件和分析數據。新功能 Gemini Live 支持與 AI 進行自然的語音對話。

搜尋與 Workspace 創新 搜尋功能將有 AI 概述來提升，使複雜的查詢變得更簡單。在 Gmail 和 Docs 中，新 AI 功能將幫助多語言寫作和組織附件。

這些進展突顯了 Google 致力於推動 AI 和以用戶為中心的技術的決心。

[閱讀更多](#)

Google Gemini AI 模型的故事

[Google Gemini](#) | [AI 模型](#) | [多模態](#) | [DeepMind](#) | [Google Research](#) | [Brain 團隊](#) | [NASA Gemini 計畫](#)

2024-05-15



Google Gemini AI 模型的故事

你是否曾經好奇 Google 是如何為其突破性的技術命名的呢？讓我們深入了解一下 Gemini AI 模型的引人入勝的起源故事。

在 Google I/O 2023 上首次提到，Gemini 是 Google 第一個原生的多模態 AI 模型，這意味著它可以同時處理文本、圖片和其他類型的數據。但是，它的名字是怎麼來的呢？

該專案最初被暱稱為 "Titan"，以土星最大的衛星命名。然而，Gemini 的聯合技術負責人 Jeff Dean 覺得另一個宇宙名稱更為合適。"Gemini" 在拉丁語中意為 "雙子"，這個名字十分理想，因為它反映了兩個強大團隊的合作：DeepMind 和 Google Research 的 Brain 團隊。

這個名字也向 NASA 的 Gemini 計畫致敬，該計畫是阿波羅登月任務的前驅。像 NASA 的努力一樣，Google 的 Gemini 旨在將我們帶入 AI 能力的新時代。

在十二月，Google 推出了多種尺寸的 Gemini 1.0——包括 Ultra、Pro 和 Nano。這個模型已經通過 Google 產品觸及了數十億人，證明了 Gemini 時代才剛剛開始。

[閱讀更多](#)

Microsoft Research 發表突破性創新

Microsoft | LLM | AI | 監督微調 | 計算筆記本 | 嵌入式系統 | 多語言模型 | AI 公平性 | MS MARCO | 自然科學

2024-05-15



Microsoft Research 發表突破性創新

向大型語言模型注入新知識Microsoft 的研究人員採用了監督微調 (Supervised Fine-Tuning, SFT) 來應對向大型語言模型 (LLMs) 更新領域外新知識的挑戰。他們實驗了基於 token 和基於事實的擴展策略，發現基於事實的擴展能夠更全面地涵蓋新資訊。這項進展有望顯著提升 LLM 在生成準確、即時回應方面的性能。

提升計算筆記本功能與 AI 結合隨著 ChatGPT 和 Copilot 等 AI 工具的整合，Microsoft 最新研究探討了這些進展對計算筆記本用戶體驗的影響。研究結果強調了在人類與筆記本以及人類與電腦互動中的新挑戰和機會，呼籲專注於改善用戶工作流程並建立 AI 工具的最佳實踐。

Jacdac：嵌入式系統原型設計的革命Microsoft 推出 Jacdac，一種基於服務的方法來設計嵌入式系統原型。這個創新框架將感測器和致動器與低成本微控制器配對，通過單線總線協議實現高效通信。Jacdac 簡化了開發過程並促進了嵌入式系統設計的模組化。

PARIKSHA：評估印度大語言模型Microsoft 新推出的評估平台 PARIKSHA，專注於多語言 LLM 的評估複雜性。通過對多種印度語言進行廣泛的人類和 LLM 評估，Pariksha 致力於促進公平透明的模型評估，推動創建具有文化相關性的 LLM 的進步。

推動 AI 公平性的脈絡化清單Microsoft 針對 AI 公平性清單的研究揭示了為特定使用案例脈絡化這些工具的重要性。研究表明，量身定製的清單可以增強 AI 實踐者對公平性的理解和實施，強調了在負責任的 AI 實踐中需要問責制和共享價值觀。

MS MARCO 網頁搜尋數據集Microsoft 公布了 MS MARCO 網頁搜尋數據集，提供了涵蓋大量查詢和語言的數百萬實際點擊標籤。這個數據集支持開發先進的信息檢索系統和新一代 AI 模型，有可能改變我們訪問和互動網頁信息的方式。

AI 在自然科學研究中的應用在最近的演講中，Bonnie Kruft 展示了 AI 驅動的突破如何加速科學發現。從開發新材料到加速藥物開發，AI 在科學中的作用正在縮短研究時間表，並在各個領域開啟新前沿。

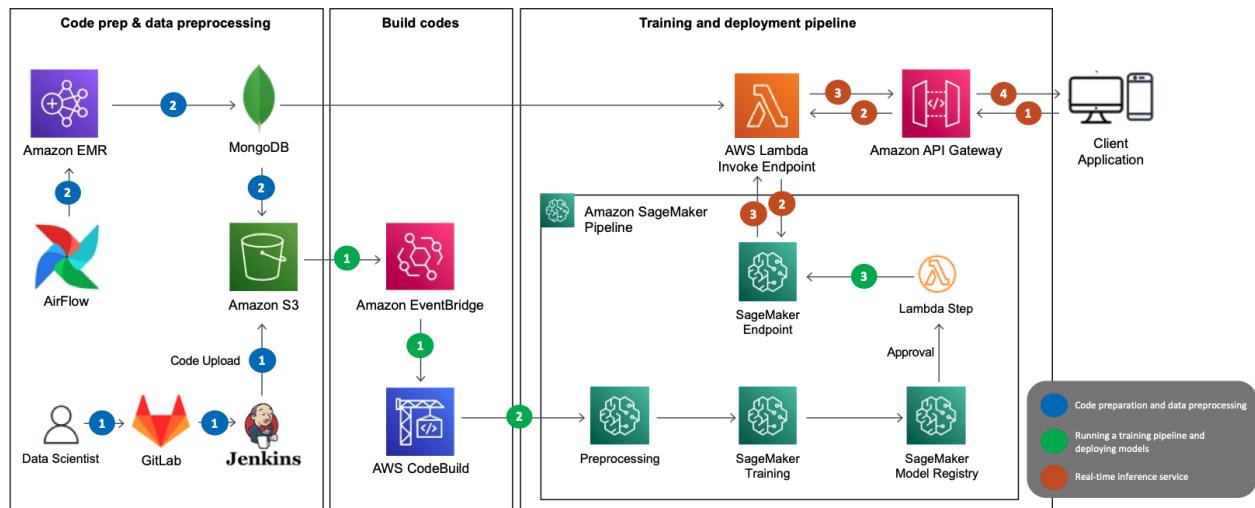
敬請期待更多來自 Microsoft Research 的洞察，他們將繼續推動科技和創新的邊界。

[閱讀更多](#)

LotteON 利用 Amazon SageMaker 和 MLOps 提升個人化推薦系統

LotteON | Amazon SageMaker | MLOps | 個人化推薦 | 深度學習 | 神經協作過濾 | AWS | 電商

2024-05-16



LotteON 利用 Amazon SageMaker 和 MLOps 提升個人化推薦系統

LotteON 為了創造一個量身定製的購物體驗，利用 Amazon SageMaker 和 MLOps 革新了其推薦系統。這個平台涵蓋了時尚、美妝和奢侈品等多種專門店，現在使用基於深度學習的推薦算法來更好地理解顧客的喜好。

挑戰：

傳統的推薦系統專注於識別產品之間的關係。然而，LotteON 需要一個更先進的系統來分析每位顧客的獨特品味。他們引入了一種神經協作過濾 (NCF) 算法，能夠識別複雜的數據關係並提供實時的個人化推薦。

解決方案：

LotteON 建立了一個包含多種 AWS 服務的 MLOps 架構：

- Amazon SageMaker Pipelines：精簡了模型訓練和部署。
- Amazon EMR：加強了用戶-產品互動數據的預處理。
- AWS Lambda 和 API Gateway：支持實時推薦推斷。
- CI/CD Pipeline Integration：促進持續的模型更新和改進。

通過結合這些技術，LotteON 現在可以向用戶提供高度相關的推薦，提高他們的購物體驗並增加互動性。

這種創新的 AWS 工具使用方式不僅提升了服務質量，還為快速原型設計和未來模型的部署奠定了基礎，確保 LotteON 在個人化電商解決方案方面保持領先地位。

[閱讀更多](#)

Microsoft Research Africa 的公平 AI 之旅

Microsoft Research Africa | AI | 公平 | 多語言模型 | 社會公益 | 人機互動 | 民族誌研究 | 機器學習
非洲語言 | 肯亞

2024-05-16



在最近一集中，Microsoft Research Podcast 節目《What's Your Story》，Johannes Gehrke 探討了 Jacki O'Neill 的精彩職業生涯。O'Neill 是 Microsoft Research Africa 位於奈洛比的主管，她在人機互動、民族誌研究和機器學習方面有豐富的背景，她分享了自己的旅程以及研究實驗室正在進行的有影響力的工作。

Microsoft Research Africa，前身為 Microsoft Africa Research Institute (MARI)，專注於使 AI 更加公平和包容。O'Neill 的團隊正在應用他們的專業知識來開發針對不同語言和文化背景的技術。他們的一個主要項目包括加強多語言語言模型，以更好地服務那些在 AI 訓練數據中往往被低估的非洲語言。

O'Neill 從在印度工作到在非洲建立研究中心的旅程，突出她致力於利用技術促進社會公益。她的團隊還在探索 AI 的未來工作，幫助肯亞的小型企業整合 AI 工具，以提高生產力和效率。

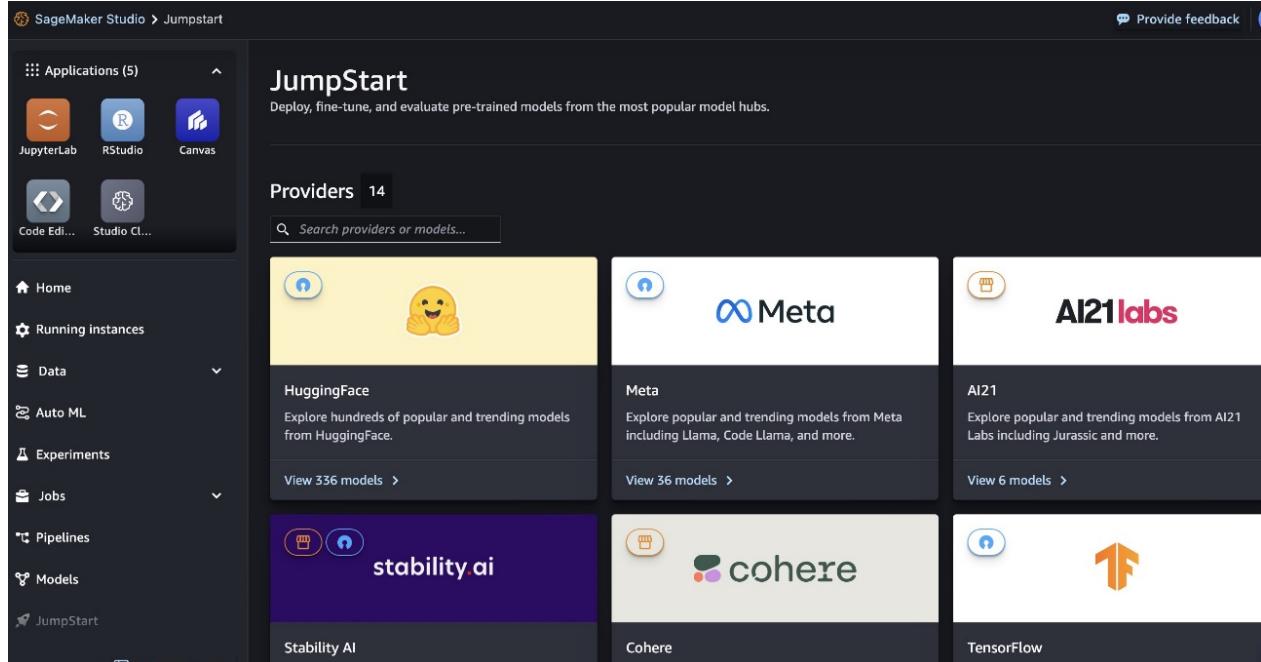
要了解更多有關他們開創性工作和推動 AI 全球公平的努力，您可以註冊 Microsoft Research Forum，該論壇特別討論了在通用 AI 時代的科學和技術話題。

[閱讀更多](#)

Mixtral 8x22B 在 Amazon SageMaker JumpStart 上設立了機器學習的新標準

Mixtral 8x22B **Amazon SageMaker JumpStart** 機器學習 大型語言模型 **SMoE** 多語言支持 編碼 數據安全

2024-05-17



Amazon 在其 SageMaker JumpStart 平台上推出了由 Mistral AI 開發的先進大型語言模型 (LLM)——Mixtral 8x22B。這項創新使用戶能夠通過單擊即可部署該模型進行推理任務，簡化了將先進機器學習能力整合到應用程序中的過程。

Mixtral 8x22B 的突出之處是什麼？

Mixtral 8x22B 是一種稀疏專家混合 (SMoE) 模型，具有 1,410 億個參數，其中在任何時刻只有 390 億個是活躍的。這種配置大大提高了成本效益，而不會犧牲性能。值得注意的是，該模型提供：

- 多語言支持：原生支持英語、法語、意大利語、德語和西班牙語。
- 高品質功能：在數學、編碼和函數調用方面表現出色，這對現代應用程序開發至關重要。
- 64,000 個 Token 的上下文窗口：增強了對大文檔的回憶和處理能力。

如何開始

SageMaker JumpStart 使機器學習從業者可以輕鬆發現和部署基礎模型。用戶可以通過 SageMaker Studio 介面或使用 SageMaker Python SDK 以編程方式訪問 Mixtral 8x22B。這個環境確保數據安全，並遵守 GDPR 和 HIPAA 等行業標準。

實際應用

在 SageMaker 上部署 Mixtral 8x22B 使用戶能夠處理各種任務，包括多語言翻譯、摘要和代碼生成。這些功能使其成為開發人員增強其 AI 驅動應用程序的寶貴工具。

訪問 SageMaker JumpStart 來探索和部署 Mixtral 8x22B，並在您的項目中利用先進機器學習的力量。

[閱讀更多](#)

WindSeer：以即時風力預測提升無人機能力

WindSeer | 微軟研究 | 無人機 | 即時風力預測 | sUAVs | 安全 | 環境監測 | 城市配送服務

2024-05-20



在一項引人入勝的發展中，Microsoft Research 推出了一項名為 WindSeer 的尖端技術，旨在顯著提升小型無人駕駛飛行器（sUAVs）的性能和安全性。這項創新由首席研究經理 Andrey Kolobov 領銜，解決了無人機操作中的一個關鍵挑戰——即時精確預測複雜地形上的風力模式。

傳統的風力預測方法常常忽略小型地形特徵，並在較大範圍內運行，這對於 sUAVs 的精確需求來說是不夠的。然而，WindSeer 提供了一種全新的解決方案。它可以利用有限的機載計算資源和最少的測量數據來估算無人機周圍依賴地形的風場。這項功能使無人機能夠更加安全高效地導航，並有可能延長其飛行時間和運營範圍。

WindSeer 的影響深遠。通過使無人機更好地理解和適應其環境，這項技術可以提升在農業、環境監測甚至城市配送服務等各個領域的應用。有了 WindSeer，未來的自動駕駛無人機操作看起來更加安全和節能，為更廣泛和更可靠的無人機技術應用鋪平了道路。

閱讀更多關於 WindSeer 及其突破性能力的資訊 [這裡](#)。

[閱讀更多](#)

評估你的 LLM 聊天機器人：關鍵考量

LLM | 聊天機器人 | 準確性 | 安全性 | 評估 | 微調模型 | 編碼 | 人工分析 | 越獄 | 提示注入 | CI/CD
Azure AI Studio | Azure PromptFlow

2024-05-20



在生成式 AI 的時代，許多企業正在部署檢索增強生成（RAG）聊天機器人。然而，評估它們的效能仍然是一個挑戰。你如何知道你的聊天機器人是否優秀？以下是 Microsoft 在評估大型語言模型（LLM）聊天機器人方面的一些見解。

確保準確性

在評估 LLM 聊天機器人時，出現兩個關鍵問題：它是否準確和安全？準確性可能很棘手，因為語言模型的輸出是概率性的——相同的輸入可能會產生不同但正確的輸出。常見的評估方法包括：

- LLM 基於技術：這些方法評估連貫性、流暢性和相關性等指標。它們可以使用 Chain-of-Thought 或少量示例的方法進行調整。
- 微調模型：這些是 LLM 的專門版本，如 Vectara 的幻覺評估模型，專為特定用途而設計。
- 編碼 + 數學：這涉及將文字轉換為數學格式進行比較，例如使用餘弦相似度或 BLEU 分數。
- 人工分析：儘管自動化發展，但專家的人為評估仍然不可或缺。

確保安全性

聊天機器人面臨如越獄和提示注入等風險。紅隊測試——開發團隊模擬攻擊以發現漏洞——至關重要。這可以涉及使用像 RedEval 這樣的公共資料集或根據你的用戶群量身定製的資料集。

綜合考量

全面的評估框架應結合準確性和安全性評估。在 CI/CD 管道中自動化這些測試，確保你的聊天機器人在部署前達到質量和安全標準。使用像 Azure AI Studio 和 Azure PromptFlow 這樣的工具可以幫助建立一個健全的評估系統。

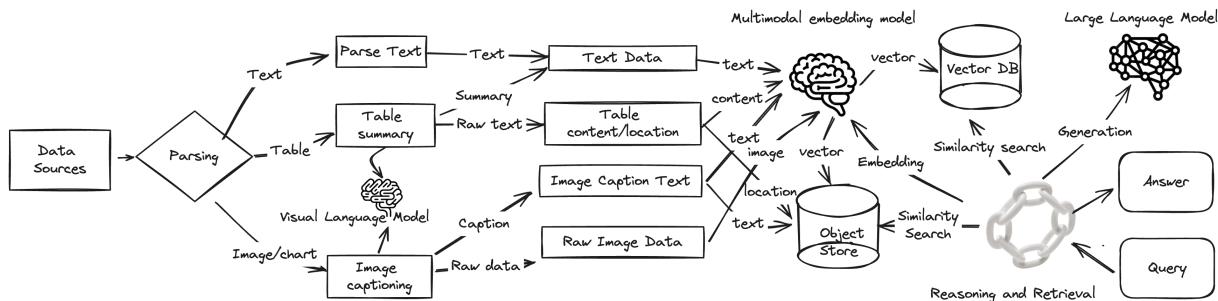
通過實施這些策略，團隊可以自信地部署並快速原型先進的聊天機器人，在快速發展的 AI 領域保持領先地位。

[閱讀更多](#)

AWS 推出具備先進 RAG 和 Amazon Bedrock 的多模態助手

AWS | RAG | Amazon Bedrock | 多模態 | 生成式 AI | 視覺語言模型 | 向量數據庫 | LangChain
Anthropic Claude 3

2024-05-21



Amazon Web Services (AWS) 最近推出了一種突破性的技術，通過稱為檢索增強生成 (Retrieval Augmented Generation , 簡稱 RAG) 的技術來提升語言模型的表現。這項進步被稱為多模態 RAG (mmRAG)，旨在解決傳統 RAG 模型在處理缺失內容、推理不匹配和多模態數據時的現有限制。

mmRAG 解決方案將文本、表格和圖像整合為統一的向量表示，實現更豐富的數據理解和檢索。利用 Amazon Bedrock 這項完全管理的服務，mmRAG 利用多種高效能基礎模型來構建安全、注重隱私的生成式 AI 應用。

mmRAG 的架構涉及從各種來源提取數據，使用視覺語言模型 (VLMs) 進行摘要，並將這些摘要嵌入到向量數據庫中。當被查詢時，系統會檢索最相關的向量以生成全面且精確的回應。這種創新方法顯著提高了系統處理和解釋多模態數據的能力，提供了一個更微妙和高效的信息檢索過程。

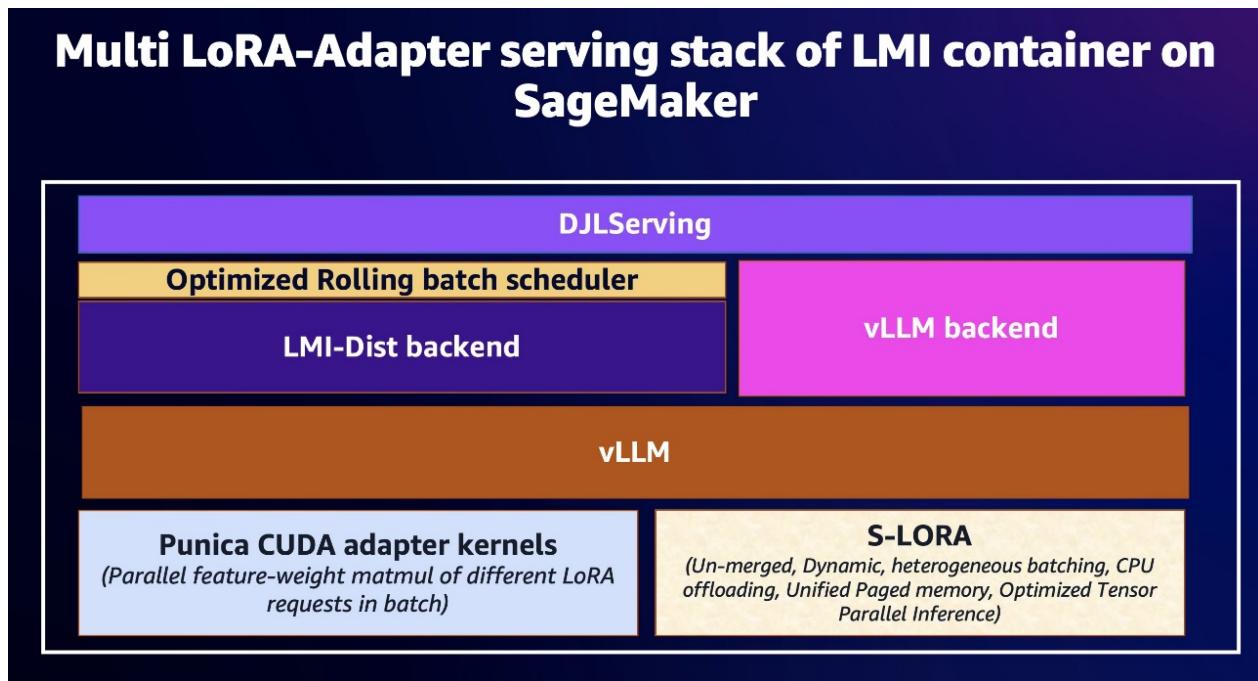
通過結合先進的 LangChain 能力和例如 Anthropic Claude 3 Sonnet 和 Haiku 等模型，AWS 旨在提供一個具有高級跨模態推理能力的強大生成式 AI 助手。這項技術有望徹底改變我們與複雜的多模態數據集互動和提取見解的方式。

[閱讀更多](#)

Amazon SageMaker 提供高效的多租戶 LoRA 服務

Amazon SageMaker | 低秩適應 | LoRA | 多租戶 | GPU | 高性能 | 微調模型 | 自訂AI模型 | 自然語言處理 | 內容生成

2024-05-21



在人工智慧 (AI) 不斷演變的世界中，生成型AI模型正在革新個人化和智慧化的體驗。企業正在利用這些模型來提升各個領域的服務，從自然語言處理到內容生成。

然而，挑戰在於如何有效地管理和部署這些模型，特別是在使用特定領域的術語進行微調時。這就是Amazon SageMaker最新創新發揮作用的地方。

低秩適應 (LoRA)，整合到Amazon SageMaker中，為部署多個自訂AI模型提供了一個具有成本效益的解決方案。LoRA允許企業針對特定需求 (如金融、醫療保健和客戶服務) 微調大型語言模型 (LLMs)，如BloombergGPT，而不需承擔通常與此類操作相關的高昂成本。

SageMaker的LoRA服務的主要特點包括：

- 合併LoRA：對於單一適配器案件而言，此方法直接修改基礎模型，確保零新增延遲。
- 未合併LoRA：此方法支持多個適配器，適合廣泛的使用案例，雖然延遲略高。

使用S-LoRA和Punica技術，SageMaker優化了GPU內存使用並提高了性能，允許數千個LoRA適配器在單個或多個GPU上高效運行。

通過這種創新方法，企業可以無縫管理多樣化的微調模型組合，優化成本並提高性能。這確保了為客戶提供可擴展、高性能和個人化的體驗，以滿足對專業化AI解決方案日益增長的需求。

如需了解更多詳情並探索此突破性解決方案，請訪問AWS Machine Learning Blog。

[閱讀更多](#)

NVIDIA 在 Green500 中以能源效率超高的超級電腦稱霸

NVIDIA | Green500 | 超級電腦 | 能源效率 | 高效能運算 | HPC | GH200 | GPU | 科學研究 | 可持續性

2024-05-21



Explore What's Next in AI With the Best of GTC

Watch On Demand

在能源效率計算技術方面取得重大突破，NVIDIA 橫掃最新的 Green500 排行榜，囊括前三名，並在前十名中佔據七個位置。這一榮譽突顯了 NVIDIA 高級技術，特別是在高效能運算 (HPC) 領域的日益影響。

這些頂尖系統的核心是 NVIDIA GH200 Grace Hopper 超級晶片。領軍的是德國 Forschungszentrum Jülich 的 JEDI 系統，實現了每瓦特 72.73 GFlops 的驚人成績。這一成就展示了 Grace Hopper 架構的卓越能源效率和計算能力。

NVIDIA 的加速計算技術利用 GPU 來優化吞吐量，大幅提升計算速度和能源效率。例如，GH200 晶片在混合精度 AI 任務上的能源效率比前幾代高出 1,000 倍以上。

這些進步對於科學研究至關重要，使得解決複雜挑戰的速度加快且更具可持續性。NVIDIA 的創新技術持續突破界限，讓高效能運算更具可用性和環保性。

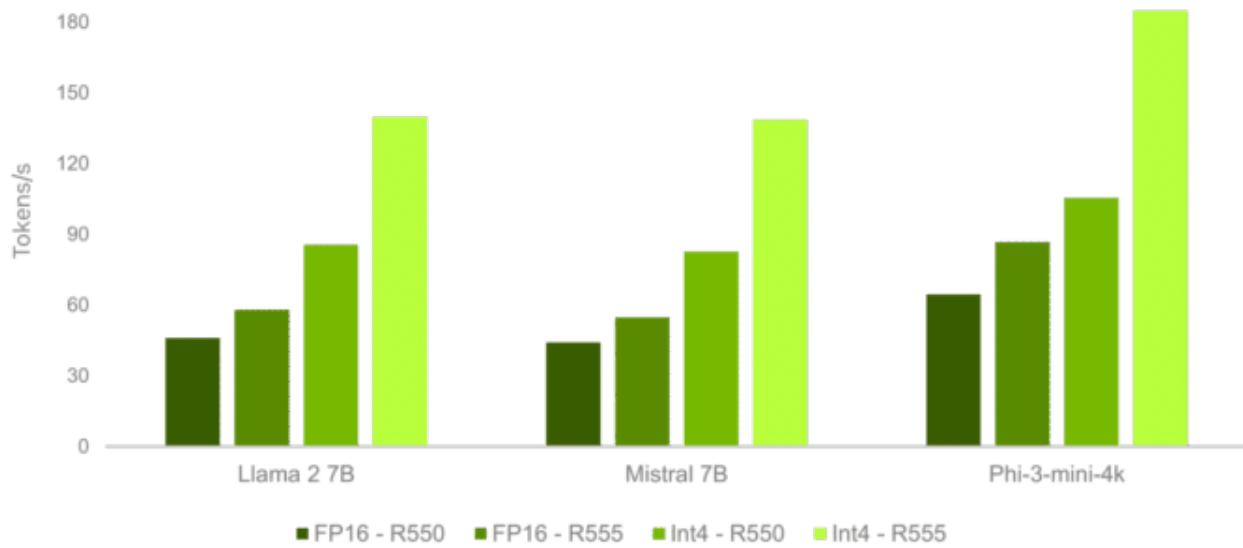
[閱讀更多](#)

NVIDIA 超級增強 RTX AI 電腦的新性能優化

NVIDIA | RTX | AI | 性能優化 | GeForce | ONNX Runtime | DirectML | WebNN | PyTorch | Adobe
 Premiere | Blender | DLSS | TensorRT

2024-05-21

LLM Inference Performance with ORT-DML
 (GeForce RTX 4090, INSEQ = 2000, BS = 1)



NVIDIA 剛剛宣布了針對其 GeForce RTX AI 電腦和工作站的突破性 AI 性能優化，這些優化在最近的 Microsoft Build 活動上揭曉。這些進步承諾為遊戲玩家、創作者和開發者帶來更快、更高效的 AI 應用。

增強的 AI 性能 使用新的 NVIDIA R555 遊戲準備驅動程式，大型語言模型（LLMs）現在可以使用 ONNX Runtime（ORT）和 DirectML 快速運行，性能提升高達三倍。這些工具允許 AI 模型直接在 Windows 電腦上操作，提高效率。

增強的網頁和應用開發 WebNN 是一個用於在網頁上部署 AI 模型的 API，現在透過 DirectML 獲得 RTX 的加速。這意味著網頁應用可以更快地整合先進的 AI 功能。此外，PyTorch 現在支持 DirectML 執行後端，允許開發者在 Windows 上本地訓練和推論複雜的 AI 模型。

針對內容創作者和遊戲玩家 RTX AI 電腦在 Adobe Premiere 和 Blender 等應用中提供 AI 輔助工作流程，簡化任務並釋放創意潛力。遊戲玩家受益於 NVIDIA DLSS，這可以將幀率提高高達 4 倍，以及 NVIDIA ACE，它通過 AI 驅動的功能增強遊戲角色的互動。

開發者工具 開發者可以利用 NVIDIA 的 AI 工具，包括 AI Workbench 和 TensorRT，直接在 RTX AI 電腦上構建和微調 AI 模型。這套工具確保了強大且流暢的開發過程。

NVIDIA 最新的性能優化不僅提升了用戶體驗，還擴展了 AI 在多個領域中的可能性。

[閱讀更多](#)

Azure AI Studio 中的新生成式 AI 評估和監控工具

Azure AI Studio | 生成式 AI | 評估工具 | 監控功能 | 自訂指標 | 測試數據 | 安全 | 可靠

2024-05-21



Microsoft 最近在 Azure AI Studio 中推出了用於評估和監控生成式 AI 應用程式的突破性工具，使這些先進技術更加安全、可靠且使用者友好。

Evaluation Tools

一個名為「Evaluators」的新功能現已預覽。Evaluators 使得開發者能夠設計 AI 輔助的評估資產，來評估使用指定的 Azure OpenAI 模型部署的各種指標。這些評估工具可以自訂並在組織內部分享，確保評估的一致性。它們可以在本地運行，或集成到 Azure AI Studio UI 中的自動化評估工作流程中。這種自動化程度允許對 AI 應用程式的進行全面的評估，以達到商業標準和目標。

Monitoring Capabilities

同樣在預覽中，Azure AI Studio 現在提供生成式 AI 應用程式的監控功能。這個工具追蹤關鍵指標，如 token 使用量、生成質量、請求數量、延遲和錯誤率。通過可視化趨勢並及時收到警報，組織可以對它們的應用程式進行持續改進。這對於在不斷演變的 AI 形勢中維持 AI 應用程式的性能和安全至關重要。

Custom Metrics and Test Data

開發者可以使用基於代碼或基於提示的方法創建特定任務的自訂評估指標。此外，Azure AI Studio 支援生成高質量的測試數據集，包括合成對抗數據集，以確保穩健的評估。

Real-World Applications

例如，一家公司如 Contoso Camping 可以使用這些工具來確保其所有的 AI 助手符合內部標準，達到品牌一致性和友好性。評估工具可以設置來一致性地衡量這些品質，跨不同的專案。

通過利用這些新功能，開發者可以確保其 AI 應用程式不僅質量高，而且符合安全和倫理標準，這使得 Azure AI Studio 成為管理生成式 AI 應用程式全生命週期的綜合平台。

推出 JAIS：Azure 上以阿拉伯語為中心的大型語言模型

JAIS Azure 阿拉伯語 語言模型 GPT-3 AI 雙語 開發者 企業 自動化 客戶服務 公共部門

2024-05-21



推出 JAIS：Azure 上以阿拉伯語為中心的大型語言模型

阿拉伯語，作為超過四億人使用的語言，擁有豐富的文化和語言多樣性。為了迎合這龐大的受眾，一項突破性的發展已經推出：JAIS，一個以阿拉伯語為中心的大型語言模型，現在在 Microsoft 的 Azure 上可用。

JAIS 30B Chat，由 Core42 開發，是一個基於 GPT-3 架構、專為阿拉伯語和英語設計的雙語模型，優化於包括摘要、翻譯和文本生成等多種應用。值得注意的是，它使用了一個自訂的 tokenizer，顯著減少了處理 tokens 的數量，因此能夠高效處理阿拉伯語文本。

通過 Microsoft 的 Azure AI 模型目錄提供，JAIS 30B Chat 為開發者和企業提供了最先進的多語言功能。該平台確保安全、可擴展且合規的 AI 部署，有助於醫療保健、金融和媒體等行業。

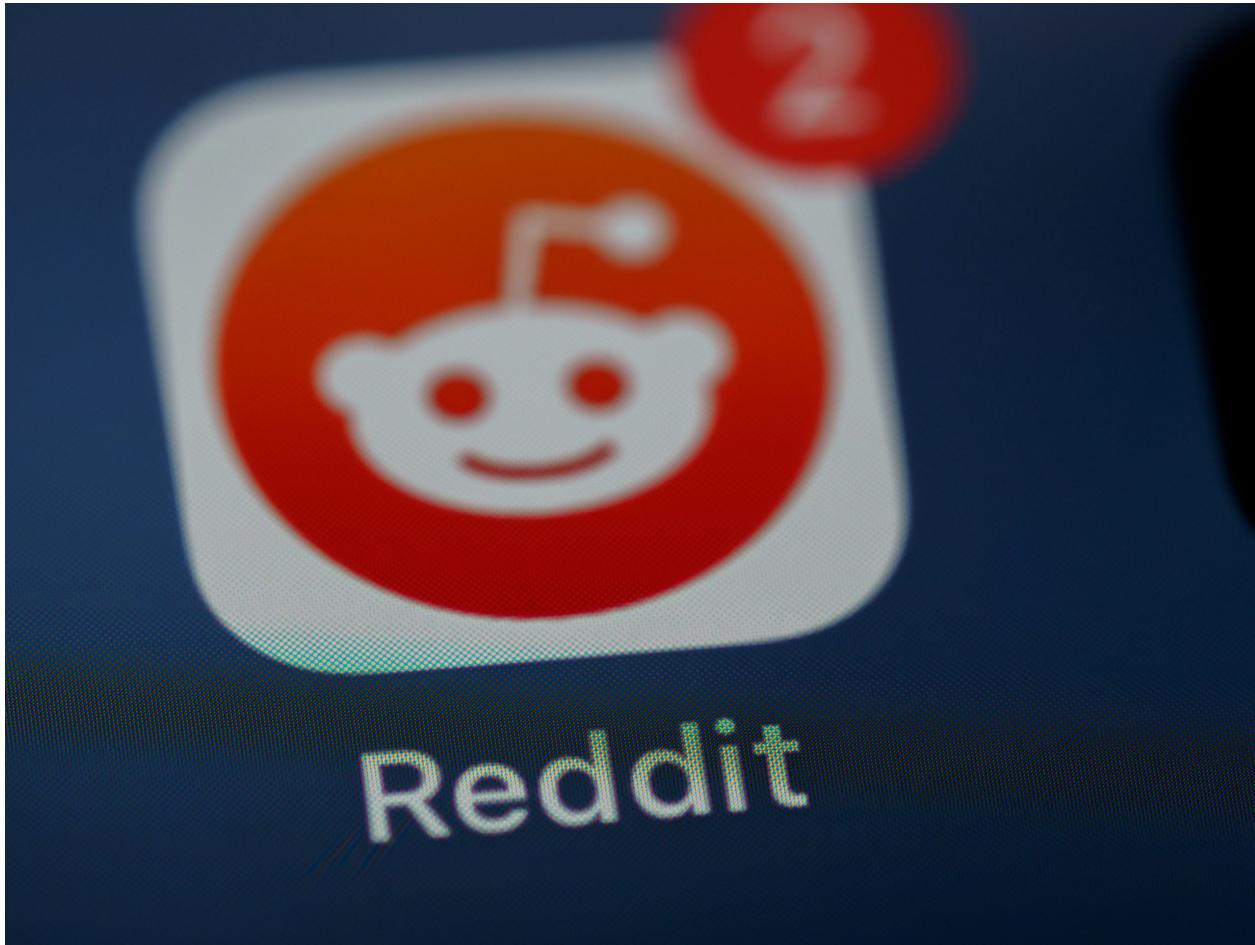
通過在 Azure 上利用 JAIS 30B Chat，企業可以增強多語言溝通、自動化客戶服務，並改進公共部門服務。這項創新標誌著向全球阿拉伯語社群提供先進 AI 的重要一步。

[閱讀更多](#)

OpenAI 與 Reddit 合作進行即時數據整合

OpenAI | Reddit | 數據整合 | ChatGPT | 自然語言處理 | AI互動 | 廣告合作 | 隱私 | 用戶信任 | 倫理標準

2024-05-22



OpenAI 最近與 Reddit 合作，通過 Reddit 的數據 API 訪問即時內容。這項合作使 OpenAI 能夠將 Reddit 討論納入其 AI 模型，包括 ChatGPT，以提高其精確度和上下文感知能力。

通過利用 Reddit 廣泛的公開討論，OpenAI 可以確保其模型隨時更新最新的趨勢和對話。這項整合旨在改進自然語言處理，使 AI 互動更加有效和細緻。

此外，Reddit 將利用 OpenAI 的先進語言模型為用戶和版主創造新的 AI 驅動工具。這些創新可能包括內容摘要和回應輔助等功能，旨在簡化用戶互動並提高管理效率。

此外，OpenAI 將成為 Reddit 的廣告合作夥伴，通過用戶行為分析幫助傳送更個性化和相關性的廣告。

儘管這次合作帶來了重大進展，但也引發了有關隱私和用戶自由的重要倫理問題。這次合作的成功將取決於維持用戶信任和遵守倫理標準。

[閱讀更多](#)

GigaPath：數位病理學的突破

GigaPath 數位病理學 Microsoft 視覺Transformer 癌症診斷 擴張自注意 Prov-GigaPath 精準健康 視覺-語言整合 PubMedBERT GPT-3.5

2024-05-22



數位病理學正經歷著革命性的變革，這要歸功於Microsoft的GigaPath。這項創新技術利用最先進的視覺Transformer來分析整片影像，這些影像是用於癌症診斷的顯微鏡載玻片的數位版本。

GigaPath的特別之處？

傳統的視覺Transformer在處理巨像素載玻片的巨大尺寸時遇到困難。GigaPath通過使用"擴張自注意"來解決這一問題，這種方法可以在不增加計算負擔的情況下，仍能捕捉整片載玻片的細節。這種方法使得GigaPath能夠考慮載玻片的整體情境，不像舊有模型那樣，只能通過分析較小的部分而錯過重要細節。

現實世界的影響

在與Providence Health System和University of Washington的合作中，Microsoft的團隊已經在超過十七萬張真實載玻片的一億多個小圖像塊上對Prov-GigaPath模型進行了預訓練。這種大規模的預訓練取得了令人印象深刻的成果，設立了癌症分類和病理任務的新基準。

例如，Prov-GigaPath在預測基因突變和癌症分型方面表現出色，在測試的26項任務中有25項達到了最先進的性能。這一能力對於精準健康至關重要，為癌症診斷和治療策略開辟了新途徑。

視覺-語言整合

GigaPath在視覺-語言任務中也展現了潛力。通過將病理載玻片影像與病理報告文本對齊，它可以執行如零樣本癌症分型和突變預測等任務。這是通過使用PubMedBERT和GPT-3.5等先進模型來進行文本編碼和處理實現的。

邁向精準健康的第一步

GigaPath代表了數位病理學的重大進步，為更準確和全面的癌症診斷鋪平了道路。隨著持續發展，它有望增進我們對腫瘤微環境的理解，並改善整體病患護理。

請繼續關注Microsoft在數位病理學領域的最新進展。

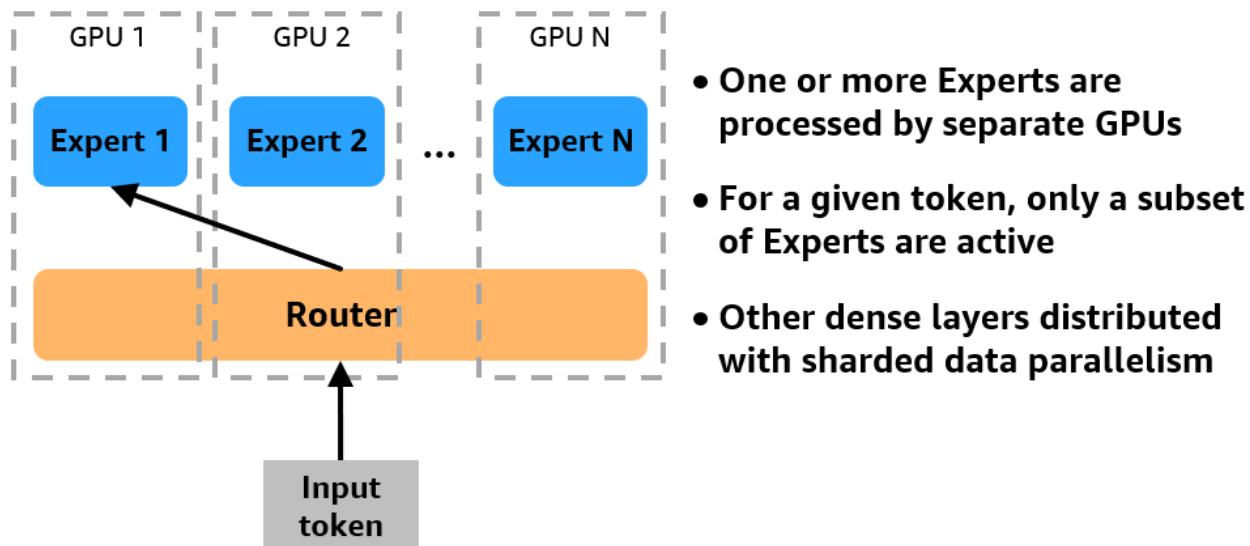
[閱讀更多](#)

加速在 Amazon SageMaker 上使用專家並行進行 Mixtral 8x7B 的預訓練

Amazon SageMaker | Mixtral 8x7B | 專家並行 | MoE | 預訓練 | 大型語言模型 | SMP 庫 | NVIDIA
 Megatron | PyTorch | FSDP | GPU

2024-05-23

Simplified expert parallelism diagram



Amazon SageMaker 在其模型並行庫中引入了新功能，使得使用專家混合架構 (Mixture of Experts, MoE) 訓練大型語言模型 (LLMs) 變得更加容易。Mixtral 8x7B 模型受益於這些進步，利用專家並行來進行高效的預訓練和微調。

專家混合架構 (MoE) 模型：MoE 模型通過使用稀疏的專家子網來處理部分的 tokens，提高了計算效率，並且相較於密集模型減少了整體計算量。然而，訓練 MoE 模型面臨挑戰，如負載平衡和高內存需求。

專家並行：現在由 SageMaker 模型並行 (SMP) 庫支持的專家並行解決了這些問題。它將 MoE 專家拆分到不同的 GPU 或設備上。這種方法確保了資源的高效利用並減少了內存負擔，使得像 Mixtral 8x7B 這樣擁有 470 億參數的大型模型得以訓練。

主要特點：

1. SMP 庫增強： - 專家並行： 將每個專家子網拆分到多個設備上。 - 混合分片： 將專家並行與分片數據並行結合，以提高內存效率。
- NVIDIA Megatron 集成： 在 PyTorch 完全分片數據並行 (FSDP) API 上利用 Megatron 進行專家並行。

1. 訓練 Mixtral 8x7B :
2. Token 路由： 一個路由器決定哪些 tokens 會發送到哪些專家，優化計算效率。

-
- 3. SMP 配置：使用 `expert_parallel_degree` 和 `hybrid_shard_degree` 等參數簡化設置，以分配 GPU 的工作負載。

通過納入這些 SMP 功能，Mixtral 8x7B 模型可以在 Amazon SageMaker 上更高效地進行訓練，這使其成為處理大規模語言模型的強大解決方案。

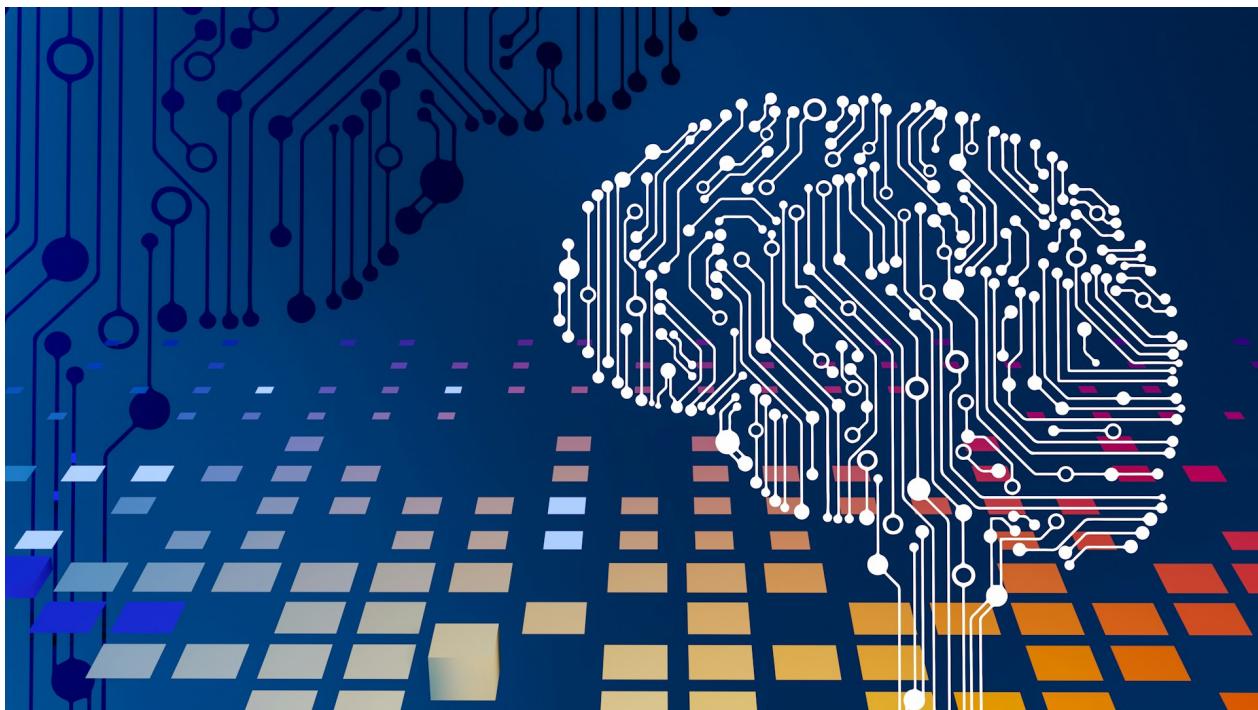
要獲取更詳細的說明和更新，您可以參考 SageMaker 模型並行庫文檔和 GitHub 儲存庫。

[閱讀更多](#)

Atera 的 AI 演進：革新 IT 管理

Atera **AI** **IT 管理** **Copilot** **Autopilot** **Action AI**

2024-05-28



Atera 的 AI 演進：革新 IT 管理

TechForge 最近與 Atera 的 CEO Gil Pekelman 討論了 AI 如何重塑 IT 管理。

Atera 於 2016 年推出的全方位 IT 管理平台，將從修補和安全到工單管理和分析等一切整合到一個直觀的界面中。這裡的關鍵創新在於 Atera 使用了統一的代碼庫和數據庫，簡化了傳統上需要多個不同工具的工作。

今年，Atera 正在開創“Action AI”，它不僅僅是生成內容，還能自動與硬體和軟體互動以解決 IT 問題。這是通過 Atera 的 Copilot 和 Autopilot 功能實現的。Autopilot 自動解決諸如重設密碼和打印機問題等常規 IT 問題，而 Copilot 則協助 IT 專業人員診斷問題並提出建議，而不會剝奪用戶的控制權。

這些 AI 工具顯著減少了 IT 專業人員的日常任務，提高了生產力和工作滿意度。隨著 AI 變得越來越智能和有同理心，Atera 的創新承諾帶來無縫、高效且類似人類的互動，改變了 IT 管理的格局。

[閱讀更多](#)

NVIDIA 在 COMPUTEX 最佳選擇獎大放異彩

NVIDIA | COMPUTEX | GH200 Grace Hopper | AI | 高效能運算 | Spectrum-X | 乙太網結構 | AI Enterprise | 雲原生軟體

2024-05-28



Explore What's Next in AI With the Best of GTC

Watch On Demand

NVIDIA 再次展示其領先的技術，並在 COMPUTEX 最佳選擇獎中贏得多項殊榮。讓我們深入探討這些創新技術的亮點。

NVIDIA GH200 Grace Hopper 超級晶片：革新 AI 和高效能運算

GH200 Grace Hopper 超級晶片在電腦與系統類別中獲得了獎項。這個突破性的平臺將高效能的 NVIDIA Grace CPU 與基於 NVIDIA Hopper 架構的 GPU 結合在一起，並通過高帶寬 900GB/s 的 NVLink 晶片對晶片互連相連接。這種協同作用實現了加速運算，大幅提升 AI 推理、大型語言模

型、高效能運算應用等。該超級晶片已經在全球超過 40 台 AI 超級電腦中使用，推動了前所未有的科學創新。

NVIDIA Spectrum-X：重新定義 AI 網路

NVIDIA Spectrum-X 平臺在網路與通訊類別中獲得獎項，這是全球首個專為 AI 打造的乙太網結構。NVIDIA Spectrum SN5600 交換機和 BlueField-3 SuperNICs 使 Spectrum-X 的生成性 AI 網路性能比傳統乙太網結構提高了 1.6 倍。這種先進的網路解決方案有望成為 AI 雲端和大型企業部署的骨幹。

NVIDIA AI Enterprise：簡化 AI 開發

NVIDIA AI Enterprise 獲得了享譽盛名的金獎。這個雲原生軟體平臺簡化了生成性 AI 應用的創建和部署。像 NVIDIA NIM 和 CUDA-X 這樣的關鍵微服務增強了模型性能，同時確保企業級的安全性和穩定性。這個平臺還優化了計算資源的利用，幫助企業降低能源成本和數據中心佔地。

持續關注 NVIDIA 繼續推動 AI、計算和網路技術的邊界。

[閱讀更多](#)

03 資訊安全

HiddenLayer Model Scanner：提升 AI 模型安全性

HiddenLayer Model Scanner | **AI 模型安全** | **惡意軟體** | **漏洞** | **後門檢測** | **Microsoft** | **Azure AI**

2024-05-21



Microsoft 最近在 Microsoft Build 活動上宣佈了一款創新安全工具——HiddenLayer Model Scanner。此工具旨在幫助開發者在將第三方和開放模型加入 Azure AI 的模型目錄之前，評估其安全性。這項計畫目標是解決新興的網路安全威脅，如惡意軟體和漏洞，確保 AI 部署的安全性。

HiddenLayer Model Scanner 採用先進技術來分析 AI 模型。它會掃描各種模型格式和框架，檢查層、張量、函數和模塊，以檢測任何惡意代碼或漏洞。這種徹底的檢查幫助開發者以更大的信心部署模型，確保其安全性。

Model Scanner 進行了多項檢查，包括：
- 惡意軟體分析：識別嵌入的惡意代碼。
- 漏洞評估：檢測常見漏洞和零日威脅。
- 後門檢測：尋找供應鏈攻擊和後門。
- 模型完整性：確保模型的結構完整性。

此掃描器代表著保護開放合作和創新空間的重要一步，使組織能夠更安全地微調和部署 AI 模型。

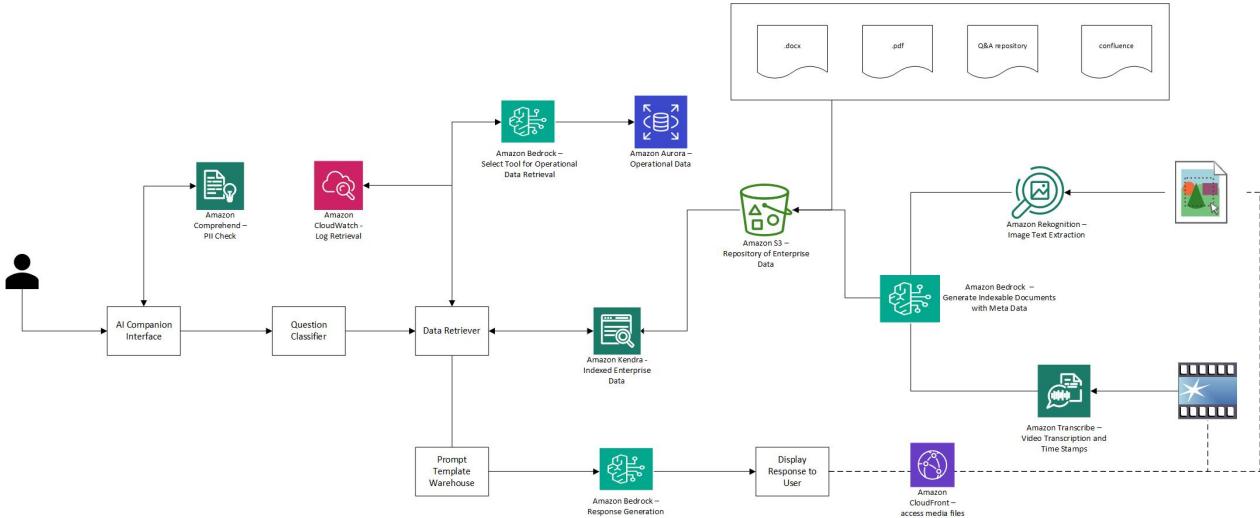
[閱讀更多](#)

04 應用

釋放生成式 AI 的能量：Verisk 的即時洞察引擎提升客戶支持

AI | 生成式AI | 即時洞察引擎 | 客戶支持 | Verisk | FAST平台 | Amazon Comprehend | Amazon Kendra | Amazon Bedrock | Amazon Rekognition | Amazon Transcribe | 保險業務 | 檢索增強生成

2024-05-09



Verisk 最推出了一個令人振奮的客戶支持創新，即其 AI 驅動的「即時洞察引擎」。這項新技術運用了生成式人工智慧 (AI)，顯著提升了保險業務中的運營效率和客戶服務。

FAST 平台與即時洞察引擎

這項創新背後的核心是 Verisk 的 FAST 平台，這個平台以其在壽險和退休領域的領導地位而聞名。現在該平台整合了一個 AI 夥伴，旨在通過回答業務流程和配置問題來提供 24/7 支持。這個 AI 夥伴使用了檢索增強生成 (RAG) 方法，通過從多個資料來源存取最新資訊來生成準確且富有上下文的回應。

主要技術和組件

1. Amazon Comprehend：識別並屏蔽用戶查詢中的個人識別資訊 (PII)。
2. Amazon Kendra：利用強大的語義搜索能力來實現上下文檢索。
3. Amazon Bedrock：包含 Anthropic Claude 模型，擅長生成自然語言回應和 SQL 查詢。
4. Amazon Rekognition 和 Amazon Transcribe：增強影像和視頻資料處理的自然語言索引。

影響和未來前景

通過即時洞察引擎，Verisk 旨在減輕支持團隊的負擔，使他們能夠專注於解決複雜問題並進一步創新。這個 AI 夥伴將通過提供即時且準確的資訊來改變客戶體驗，使產品更快上市並提高客戶服務效率。

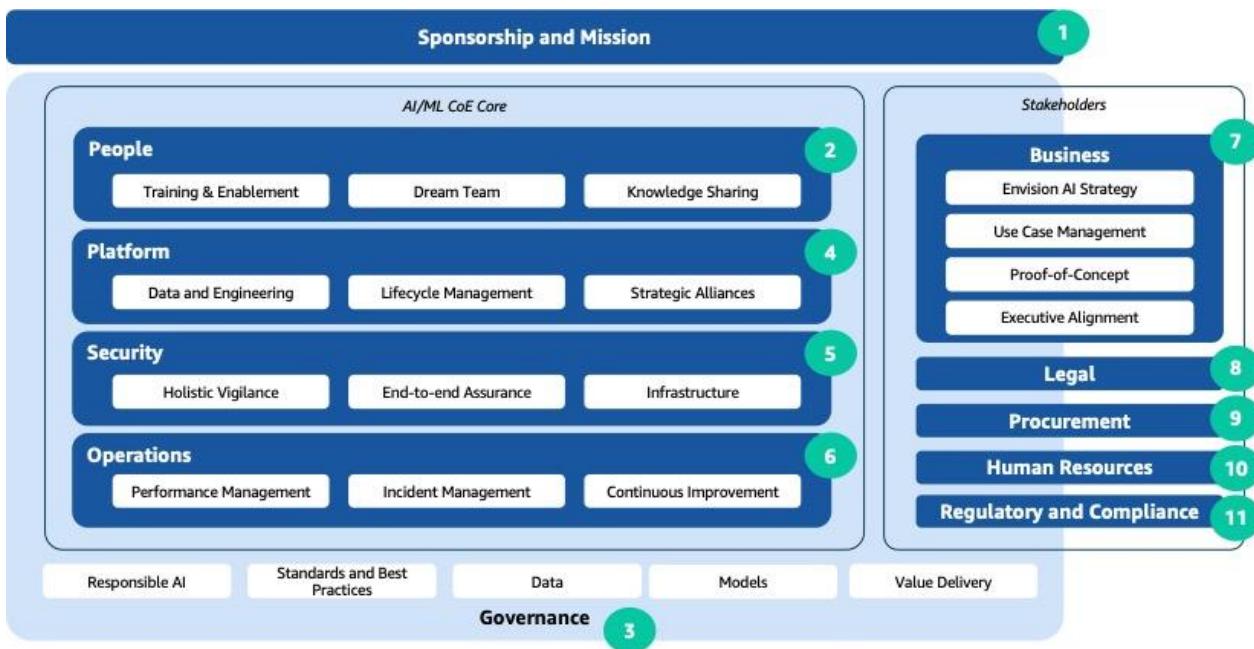
隨著生成式 AI 的不斷發展，Verisk 計劃擴展其 AI 夥伴的能力，有可能徹底改變保險業中客戶支持和業務流程的管理方式。

[閱讀更多](#)

建立 AI/ML 卓越中心：企業的遊戲規則改變者

AI | ML | 卓越中心 | 企業 | 生產力 | 風險管理 | 治理 | 投資回報

2024-05-09



人工智能和機器學習 (AI/ML) 的顯著進展正在革新各種產業。根據 McKinsey 的一項最新研究，僅生成式 AI 預計就能為金融服務業帶來超過 4000 億美元的生產力效益。隨著超過 80% 的企業預計在 2026 年之前部署 AI，建立一個 AI/ML 卓越中心 (CoE) 已成為各組織的關鍵。

AI/ML CoE 是一個專門的單位，負責協調和監督組織內所有 AI/ML 項目。其使命是將 AI/ML 的應用案例與商業策略對齊，確保解決方案的可擴展性和互操作性。這個中央樞紐促進創新，減少風險，並維持治理標準，從而加快上市時間，優化風險管理，並最大化投資回報。

通過採用像 Cloud Adoption Framework for AI/ML 這樣的框架中的最佳實踐，組織可以建立明確的領導角色，提供有針對性的培訓，並實施強有力的治理。CoE 還確保道德的 AI 部署，維護數據的完整性，並促進持續改進，這使其成為任何前瞻性企業不可或缺的資產。

請繼續關注我們即將發布的有關如何創建成功的 AI/ML CoE 的更多見解。

[閱讀更多](#)

AWS DeepRacer：人人都能接觸的機器學習入門

AWS DeepRacer 機器學習 人工智慧 自動駕駛 模型訓練 全球賽車聯盟 AI 教育 內容生成

影像分類

2024-05-10



人工智能 (AI) 和機器學習 (ML) 正在重塑我們的世界，而 AWS DeepRacer 讓這些技術對所有人都變得可及，無論技能水平如何。AWS DeepRacer 於 2019 年推出，是一個有趣的平台，透過自動駕駛賽車的動態方式教授 ML 的基本知識。

透過 AWS DeepRacer，使用者可以在模擬環境中定義獎勵函數、設置訓練參數和配置工作，從而深入了解模型訓練和評估。這種實際操作經驗對於理解和應用 ML 概念於各種現實世界應用，如內容生成和影像分類，具有無價的價值。

此外，AWS DeepRacer 還提供一個全球賽車聯盟，參賽者可以在其中競爭、獲取獎勵並展示他們的 ML 技能。像是由 JPMorgan Chase 主辦的“世界最大規模的全球女性 AWS DeepRacer 聯盟”這樣的活動，突顯了該平台在普及 ML 教育和促進創新方面的角色。

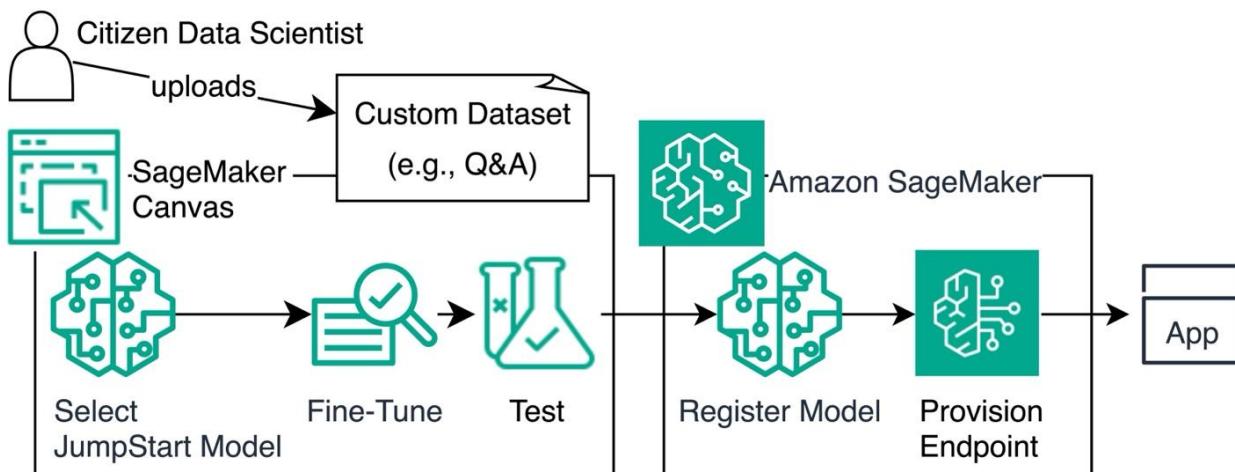
無論您是經驗豐富的開發者還是好奇的新手，AWS DeepRacer 都提供了一個有趣且實用的方式來開始您的 AI 和 ML 之旅。

[閱讀更多](#)

使用 Amazon SageMaker Canvas 和 JumpStart 進行無程式碼 LLM 微調，提升客戶互動

Amazon SageMaker Canvas | **SageMaker JumpStart** | **LLM微調** | **客戶互動** | **無程式碼** | **AI技術** | **個性化體驗** | **AWS安全環境**

2024-05-10



Amazon 推出了一種創新方式，透過其無程式碼解決方案，使用 Amazon SageMaker Canvas 和 SageMaker JumpStart 來微調大型語言模型（LLMs），以提升客戶互動。這種新方法使企業能夠根據品牌聲音精確地自訂 LLMs，而無需深入的技術專業知識。

SageMaker Canvas 提供一個使用者友好、點擊操作介面，用於微調 LLMs。這意味著商業用戶只需通過與介面互動即可調整模型，以滿足其需求，而無需撰寫複雜的程式碼。該工具能與 SageMaker JumpStart 和 Amazon Bedrock 模型無縫協作，提供選擇基礎模型（FM）的靈活性。

透過在公司特定數據上微調 LLMs，企業能確保在所有客戶接觸點上的一致訊息傳遞，從而創造個性化的客戶體驗。此外，使用於微調的數據保持在公司安全的 AWS 環境內，確保數據隱私。

這個工具對於已經投資於 Amazon SageMaker 生態系統的公司特別有利，因為它確保了運行的連續性，並提供了對部署選項的細膩掌控。

總結來說，Amazon SageMaker Canvas 和 SageMaker JumpStart 讓企業比以往更容易利用先進的 AI 技術來改善客戶互動和推動增長。

[閱讀更多](#)

AWS 引入生成式 AI 用於臨床報告摘要

AWS | 生成式 AI | 臨床報告摘要 | 大型語言模型 | Amazon Bedrock | 提示工程 | 檢索增強生成 | 放射學 | Claude 3 Sonnet | 醫療文書

2024-05-13

----- Zero-shot -----

Human: Generate radiology report impressions based on the following findings and output it within <impression> tags.

Findings: Heart size is normal. The mediastinal and hilar contours are unremarkable. The pulmonary vasculature is not engorged. Subsegmental atelectasis is seen in the lung bases without focal consolidation. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.

Assistant:

<impression>

The chest x-ray demonstrates no acute cardiopulmonary abnormalities. Specifically, the heart size and mediastinal contours are within normal limits. The lungs are clear and expanded without evidence of focal consolidation, pulmonary edema or pleural effusion. Minor atelectasis is present in the lung bases. The visualized bony structures appear intact.
</impression>

Amazon Web Services (AWS) 近期在使用生成式 AI 來提升臨床報告摘要方面取得了重大進展。這包括對大型語言模型 (LLMs) 進行微調，並採用新技術來改進醫療文書，特別是在放射學領域。

AWS 推出了 Amazon Bedrock，一種完全託管的服務，提供來自領先 AI 公司的一系列高效能基礎模型。Amazon Bedrock 允許用戶構建具有增強安全性和隱私性的生成式 AI 應用程序，並將這些功能無縫整合到現有的 AWS 服務中。

討論的一項關鍵技術是提示工程，這涉及設計提示以引導 LLMs 產生更好的結果。AWS 比較了兩種方法：零範例提示和少量範例提示。零範例提示不使用任何範例，而少量範例提示則包含一些範例以提升性能。

此外，AWS 還探討了檢索增強生成 (RAG)。這種技術在相關數據集中進行搜索，以為 AI 提供上下文，大大提高了生成摘要的質量。實驗顯示，RAG 在使用 ROUGE 指標評估時，表現優於零範例和少量範例提示技術。

在他們的試驗中，AWS 通過 Amazon Bedrock 使用了 Anthropic 的 Claude 3 Sonnet 模型。數據集包括 2,000 份去識別的放射學報告，展示了 AI 如何從臨床筆記生成簡明而準確的摘要。

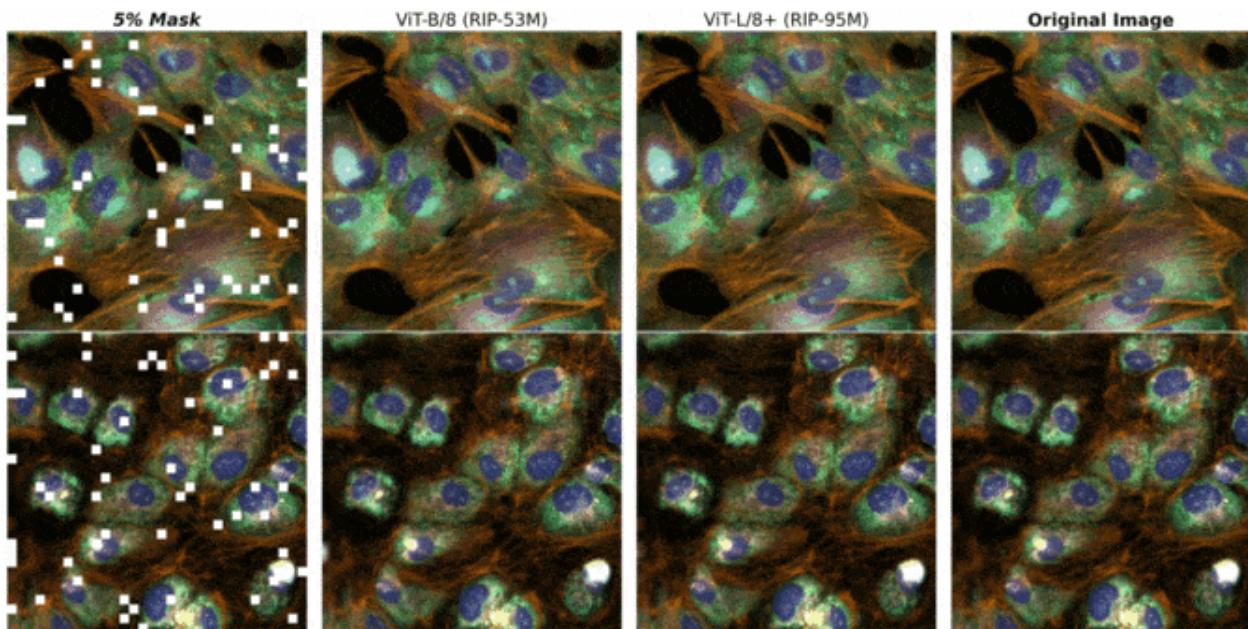
這一發展標誌著 AI 在醫療領域應用的重大一步，承諾更好的文書處理和潛在的患者護理成果改進。

[閱讀更多](#)

NVIDIA 和 Recursion 加速藥物發現與 AI 超級電腦

**NVIDIA | Recursion | AI 超級電腦 | BioHive-2 | 藥物發現 | NVIDIA H100 Tensor Core GPU | NVIDIA
Quantum-2 InfiniBand | AI 性能 | 生物數據 | Phenom-Beta | NVIDIA BioNeMo**

2024-05-13



NVIDIA 和 Recursion 透過他們的 AI 驅動超級電腦 BioHive-2，在藥物研究領域取得了重大進展。最近，位於 Recursion 鹽湖城總部的 BioHive-2，在全球最快超級電腦 TOP500 名單中攀升了超過 100 名，現在排名第 35 位。

這個尖端系統整合了 504 個 NVIDIA H100 Tensor Core GPU，通過 NVIDIA Quantum-2 InfiniBand 網絡連接，實現了 2 exaflops 的 AI 性能。計算能力的提升使得 BioHive-2 比其前身 BioHive-1 快了近五倍。

憑藉這一性能，BioHive-2 加速了藥物發現過程。通過使用 AI 模型來分析大量生物數據，Recursion 可以減少一半的傳統濕實驗次數，同時仍然能獲得 80% 的所需結果。考慮到生物系統的複雜性以及發現新藥物候選者所需的數百萬次實驗，這種效率至關重要。

BioHive-2 的能力不僅限於速度，還使得開發先進的 AI 模型如 Phenom-Beta 成為可能，該模型現已作為雲端 API 在 NVIDIA BioNeMo 上提供。這一創新有望通過優化各種疾病的分子發現來改變醫療保健行業。

隨著這些進展，NVIDIA 和 Recursion 站在藥物設計革命的最前沿，標誌著醫療技術的一個轉折點。

AWS 引入混合人機工作流程於生成式 AI 應用

AWS | 生成式 AI | 人機工作流程 | RLHF | Amazon SageMaker | Amazon Bedrock

2024-05-14

Amazon SageMaker Q&A Bot

The screenshot shows a chatbot interface titled "Amazon SageMaker Q&A Bot". The user asks "Can we train models on Sagemaker?" and the bot responds "yes". It then provides a link to Amazon SageMaker FAQs for more details. The bot continues to explain that reinforcement learning models can be trained in SageMaker, while other models like Neo can be optimized for SageMaker ML instances or AWS IoT Greengrass supported devices. It also mentions SageMaker Edge Manager for edge optimization.

Can we train models on Sagemaker?

yes

You can also refer to the following documents for details

1. Amazon SageMaker FAQs - page 24
- environment by trial and error using feedback from its own actions and experiences.Q: Can I train reinforcement learning models in SageMaker?Yes, you can train reinforcement learning models in SageMak.....
2. Amazon SageMaker FAQs - page 33
- model?No. You can train models elsewhere and use SageMaker Neo to optimize them for SageMaker MLinstances or AWS IoT Greengrass supported devices.Q: Which models does SageMaker Neo support?Currently,
3. Amazon SageMaker FAQs - page 30
- you can request predictions for a large number of data records and transform the data quickly andeasily.Q: What is Amazon SageMaker Edge Manager?SageMaker Edge Manager makes it easier to optimize, sec.....

How can we train models on Sagemaker?

Ask me anything about Amazon SageMaker...

Submit

在一個重大的進展中，AWS 推出了一個能夠無縫整合線上和線下人機工作流程的解決方案，專為生成式 AI 應用設計。此創新旨在提升生成式 AI 的性能和可靠性，可創造出類似人類的內容，如圖像、文字和音頻。

核心技術

1. 從人類反饋中學習的強化學習 (RLHF)： RLHF 結合了人類反饋和機器學習模型，微調 AI 的性能。通過評估和批評輸出，模型不斷改進，使 AI 更加準確和負責任。
2. 近實時人類工作流程：此解決方案允許人類專家在必要時介入正在進行的 AI 用戶互動。例如，如果用戶在使用聊天機器人時有負面體驗，可以實時通知人類接管對話。
3. Amazon SageMaker Ground Truth： SageMaker Ground Truth 提供全面的人在迴圈中的能力，支持數據生成、模型審查和評估等任務。這個反饋迴路對於完善 AI 模型至關重要。
4. Amazon Bedrock：作為一個完全託管的服務，Amazon Bedrock 提供來自領先 AI 公司高性能基礎模型的訪問。它允許輕鬆的實驗、定制和部署生成式 AI 應用。

實際應用

想像一個使用生成式 AI 驅動的問答機器人的客戶，想了解 Amazon SageMaker 的信息。如果機器人檢測到客戶查詢中的負面情緒，它可以觸發一個工作流程，通知人類專家介入。這種實時的人類參與確保能及時有效地解決客戶問題。

此外，離線的 RLHF 工作流程涉及存儲聊天記錄並讓人類專家進行評估。這些評估有助於微調機器學習模型，確保 AI 回應的持續改進。

這些工作流程共同創造了一個堅實的系統，使 AI 和人類智能相輔相成，導致更準確和值得信賴的生成式 AI 應用。通過這些進步，AWS 正在為更具責任心和更有效的 AI 解決方案鋪平道路。

[閱讀更多](#)

Cerence 使用生成式 AI 轉變車內體驗

Cerence 生成式AI **NVIDIA** 車內體驗 語音助手 自然語言處理 **Microsoft Azure** **NVIDIA DGX Cloud**

2024-05-14



Explore What's Next in AI With the Best of GTC

Watch On Demand

人工智能 (AI) 的整合正在戲劇性地改變我們的駕駛體驗，將車輛變成智慧、互聯的實體。

Cerence，作為AI驅動的移動解決方案的先驅，與NVIDIA合作，創造下一代車內體驗，承諾將革新我們與車輛互動的方式。

Cerence創新核心是Cerence Automotive Large Language Model (CaLLM)。這個模型是該公司下一代車內計算平台的基礎，該平台運行於NVIDIA DRIVE。這個平台於12月揭幕，介紹了一個專為汽車設計的助手，旨在提供無縫、整合的車內體驗。

生成式AI使車輛能夠用驚人的準確性理解和回應人類語言。無論是啟動導航的語音指令、管理娛樂系統，還是進行自然語言對話，這項技術為使駕駛更加方便和愉快開啟了無限可能。

Cerence 借助 Microsoft Azure 上的 NVIDIA DGX Cloud 進行可擴展的AI工作負載訓練，確保實時性能和無縫的用戶體驗。結果是，駕駛將演變成每個人都更加個人化、互聯和安全的體驗。

[閱讀更多](#)

NVIDIA 與日本合作推動生成式 AI 基礎建設

NVIDIA | 日本 | 生成式 AI | 基礎建設 | 雲端基礎設施 | AI 工廠 | 超級計算機 | 數據中心 | 自然災害 | 氣候變化

2024-05-14



Explore What's Next in AI With the Best of GTC

Watch On Demand

在技術進步的雄心壯志下，日本將投資超過 7 億 4 千萬美元發展其生成式 AI 基礎建設。與 NVIDIA 以及本地公司如 GMO Internet Group、KDDI Corporation 和 SoftBank Corp. 合作，這項計畫旨在顯著增強該國的 AI 能力。

NVIDIA 的合作重點在於建立堅固的雲端基礎設施，這對 AI 應用至關重要。這項計畫符合日本建立'AI 工廠'的願景——這些最先進的數據中心旨在處理複雜的 AI 任務。這些 AI 工廠將在將大量數據集轉化為可操作的見解方面發揮關鍵作用，從而成為現代經濟的支柱。

此外，日本政府計劃補貼 AI 超級計算機的創建。這不僅會加速 AI 的採用，還會增強勞動力技能，支持日語模型的開發，並提高應對自然災害和氣候變化的韌性。

通過此舉，日本加入了全球增強國內計算能力的浪潮，旨在以尖端 AI 技術支持初創企業、企業和研究。

[閱讀更多](#)

釋放 Google AI 在 Android 上的威力

[Google AI](#) [Android](#) [Gemini](#) [圈選搜尋](#) [多模態功能](#) [智慧型手機](#) [詐騙警報](#)

2024-05-14



在一項突破性的更新中，Google 正在將先進的 AI 直接整合到 Android 作業系統中，革新用戶與設備互動的方式。以下是一些主要亮點：

圈選搜尋 作業幫助 圈選搜尋變得更加智能，使學生能夠直接在手機和平板電腦上解決物理和數學問題。通過圈出問題，學生可以獲得分步說明，讓他們更容易掌握複雜的概念，而無需切換應用程式。

Gemini 於 Android 上的更新 Gemini 是 Android 上的新 AI 助手，提升了語境理解能力。用戶可以拖放生成的圖像，對視頻和 PDF 提出詳細問題，並根據屏幕上的內容接收動態建議。此功能將很快推送到數百萬台設備上。

具多模態功能的 Gemini Nano 從 Pixel 設備開始，Gemini Nano 將帶來多模態功能，使手機能夠處理文字、圖像、聲音和語音，確保更豐富的互動和在通話期間的實時詐騙警報。

這些更新標誌著 AI 如何讓您的智慧型手機體驗更加直觀和安全的重大飛躍。請繼續關注更多創新，因為 Google 將繼續深入整合 AI 到 Android 生態系統中。

[閱讀更多](#)

3 種利用 Gemini for Google Workspace 提升生產力的新方法

[Gemini for Google Workspace](#) | 生產力 | [Gmail](#) | [Docs](#) | [Drive](#) | [Slides](#) | [Sheets](#) | [幫我寫](#) | [多語言支援](#)

2024-05-14



Google 為 Gemini for Google Workspace 推出了令人興奮的更新，帶來了旨在提升個人和專業用戶生產力的功能。

1. 在 Workspace 側邊欄中增強的 Gemini 1.5 Pro 從今天開始，Gmail、Docs、Drive、Slides 和 Sheets 側邊欄中的 Gemini 將使用 Gemini 1.5 Pro。這次升級允許 Gemini 參考更大範圍的資料集，提供更有見地的回應和摘要。無論你是需要總結大量的電子郵件，還是快速提取重要細節，Gemini 較長的上下文窗口都能讓這些變得輕而易舉。
2. Gmail 手機應用程式中的新功能 Gmail 手機應用程式現在包含一些工具，幫助你隨時隨地保持高效：
 - 總結電子郵件：快速獲取長電子郵件線索的重點。
 - 語境智慧回覆：接收詳細且語境感知的回覆建議。
 - Gmail 問答：請 Gemini 總結電子郵件、列出下一步行動或在收件箱中尋找特定資訊。
3. 幫我寫功能的多語言支援 Google 正在擴展 Gmail 和 Docs 中“幫我寫”功能的語言支援。現已在桌面版中支援西班牙語和葡萄牙語，將來還會支援更多語言。

這些更新旨在簡化你的工作流程，讓日常任務變得更輕鬆。查看 Gemini for Workspace 插件以獲取更多詳細資訊。

[閱讀更多](#)

介紹Google的LearnLM：教育中的生成式AI革命

[Google](#) [LearnLM](#) [教育](#) [生成式AI](#) [DeepMind](#) [Google Research](#) [Google Search](#) [YouTube](#)
[Android](#) [Google Classroom](#) [Illuminate](#) [Learn About](#)

2024-05-14



Google已經發佈了LearnLM，一個為教育用途而精心調整的革命性生成式AI模型家族。LearnLM是透過Google DeepMind、Google Research以及各產品團隊的合作開發而成，旨在使教學和學習體驗變得更有動力、更個性化和更具吸引力。

LearnLM根植於教育研究，並結合學習科學的原則，旨在激發主動學習、管理認知負荷、適應學習者的需求、激發好奇心並加深元認知。這些AI模型已經在提升Google Search、YouTube和Android等熟悉的產品。例如，在Google Search中，用戶現在可以調整AI概覽，更好地理解複雜的主題，而YouTube則使用對話式AI來幫助學習視頻播放過程中的內容。

此外，LearnLM正在Google Classroom中試點，旨在簡化課程規劃，幫助教師發現新活動並根據學生的個別需求量身定制課程。此外，Google還在嘗試新的工具，如Illuminate，將研究論文轉化為易於理解的音頻對話，還有Learn About，一個基於聊天的體驗，引導用戶以自己的步調學習各種主題。

總之，Google的LearnLM將通過整合AI來改變教育，使學習變得更加可及和個性化。

[閱讀更多](#)

四位藝術家與 Imagen 2：通過 AI 重新想像奇幻世界

AI | Imagen 2 | Google Creative Lab | 無限仙境 | 愛麗絲夢遊仙境 | DeepMind | StyleDrop | Eric Hu
Shawna X | Erik Carter | Haruko Hayakawa

2024-05-14



四位富有遠見的藝術家——Eric Hu、Shawna X、Erik Carter 和 Haruko Hayakawa——與 Google Creative Lab 合作，利用最先進的 AI 技術為《愛麗絲夢遊仙境》注入新生命。

利用 Google DeepMind 的 Imagen 2，這些藝術家創造了名為「無限仙境」的創新 AI 驅動體驗。每位藝術家都使用他們獨特的風格對 Imagen 2 模型進行微調，從而能夠為這部經典小說中的每個句子生成無限且個性化的圖像。這不僅為讀者提供了一次獨特的視覺奇幻之旅，還重新定義了創意的邊界。

傳統上，要實現風格轉換需要數千張圖像。然而，隨著 Imagen 2 的 StyleDrop，這些藝術家只需約十多張圖像就能對模型進行微調，實現了新的創作表現規模。這個迭代過程使藝術家能夠精確地操作 AI 的輸出，以符合他們的美學願景。

「無限仙境」展示了 AI 如何增強人類創意，提供新的敘事維度，並使藝術家能夠以前所未有的方式與觀眾共同創作。

[閱讀更多](#)

使用 Amazon Bedrock 自動化考試創建

自動化考試生成 | **Amazon Bedrock** | 人工智慧 | 無伺服器 | 教育技術 | **Anthropic Claude**

考試問題

2024-05-15



自動化考試創建與 Amazon Bedrock

為教育工作者創建考試和測驗是一項艱鉅的任務，涉及到細緻的計劃和關注細節。引入 Amazon Bedrock 這一突破性的解決方案，利用人工智慧來簡化這一過程。讓我們深入探討這款無伺服器應用程序如何改變考試生成。

創新：Amazon Bedrock

Amazon Bedrock 是一項完全管理的服務，提供從領先的人工智慧公司如 Anthropic、AI21 Labs、Cohere、Meta 等提供的各種高性能基礎模型 (FMs)。通過使用單一 API，教育工作者可以利用這些模型來構建以安全、隱私和負責任的 AI 實踐為重點的生成式 AI 應用程序。

考試生成過程

該解決方案的核心是 Amazon Bedrock 上的 Anthropic Claude v2.1 模型，該模型擅長處理相當於多達 500 頁的廣泛文件。此模型使用先進的提示工程來將講座內容分解為清晰、可管理的步驟，創建高質量的選擇題和判斷題。

以下是簡化的工作流程：

1. 教育者路徑：
 - 上傳講座內容作為 PDF。
 - 系統處理此文件並使用 Anthropic Claude 模型生成考試問題。
 - 考試準備好後，教育者會收到通知。

1. 學習者路徑：

-
2. 學生登錄參加考試，從生成的測試中進行選擇。
 3. 系統為學習者和教育者提供即時反饋和分數。

主要優點

- 節省時間：自動化生成考試問題的繁瑣任務，讓教育者能更多地專注於教學。
- 即時反饋：學生能夠立即獲得結果，有助於學習和記憶。
- 可擴展性：AWS 無伺服器技術確保應用程序能夠在沒有人工干預的情況下處理不同的負載。

擴展可能性

此解決方案提供了多種擴展機會。例如，支持批量上傳多個文件，或者整合先進的分析功能來追蹤和提高學生的表現。

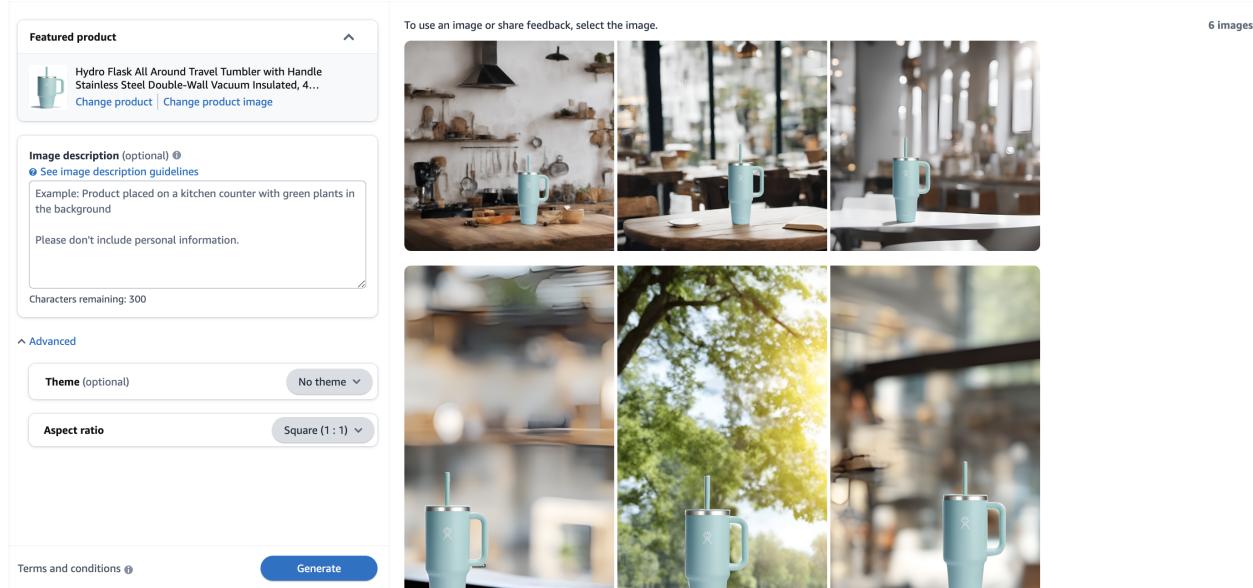
總之，Amazon Bedrock 的無伺服器考試生成器代表了教育技術的一大飛躍，使教育工作者的生活更輕鬆，同時提升了學生的學習體驗。

[閱讀更多](#)

Amazon Ads 推出由 Amazon SageMaker 支持的 AI 生成圖像功能

Amazon Ads | **Amazon SageMaker** | **AI生成圖像** | **廣告** | **生成式AI** | **Amazon Comprehend** | **Amazon Rekognition**

2024-05-15



Amazon Ads 推出了一項突破性的生成式 AI 圖像生成功能，使用 Amazon SageMaker 技術。這項創新旨在幫助廣告主輕鬆創建引人入勝且具情境相關性的圖像，即使沒有技術或設計專業知識。

這個過程從 Amazon SageMaker JumpStart 開始，允許使用各種基礎模型進行快速原型設計和迭代。廣告主可以從基本的產品照片中生成豐富的生活方式圖像，同時保留原始產品特徵。這意味著即使產品照片起初只是一張白色背景的簡單圖像，該 AI 也能將其轉變為引人注目且具情境的廣告圖像。

解決方案架構涉及幾個關鍵步驟：模型開發、人機互動評論以確保圖像質量，並使用 Amazon SageMaker 工具套件進行部署。生成的圖像通過 API 處理，確保它們符合內容指導方針，這過程使用了 Amazon Comprehend 和 Amazon Rekognition。

最終，這個生成式 AI 解決方案使廣告主能夠生成多個引人入勝的產品圖像，簡化創作過程並提升消費者體驗。通過這項創新，Amazon Ads 繼續推動數字廣告的可能性邊界。

[閱讀更多](#)

機器人的未來：AI 正在改變產業

機器人 | AI | 產業 | 自動化 | 數字雙胞胎 | 生成式 AI | 合作 | NVIDIA | Teradyne | Siemens

2024-05-15



Explore What's Next in AI With the Best of GTC

Watch On Demand

最近，來自 NVIDIA、Siemens 和 Teradyne Robotics 的業界領袖齊聚丹麥的歐登塞市——一個以機器人創新著稱的城市——討論自動化機器和 AI 的突破性進展。這次聚會標誌著 Teradyne Robotics 新總部的啟動，該總部將成為促進機器人協作與創新的一个樞紐。

一個重要的亮點是生成式 AI、模擬和數字雙胞胎在機器人革命中的角色。這些技術使機器人能夠理解、學習和做出決策，增強其認知能力。例如，NVIDIA 的 Isaac 平台在模擬和優化邊緣計算方面迅速取得了進展。

合作是中心主題。NVIDIA 和 Teradyne 展示了一種基於 AI 的內部物流解決方案，強調了合作夥伴關係的力量。這種協同效應正在推動機器人市場，特別是在歐洲，協作機器人市場預計將從 2022 年的 2.86 億美元激增至 2032 年的 67 億美元。

隨著產業越來越多地採用先進的機器人技術，AI 將機器人從僅僅是工具轉變為智能夥伴，推動各個領域的效率、精確性和生產力。

[閱讀更多](#)

Mozilla Firefox 增加 AI 驅動的 RTX Video 以提升串流品質

Mozilla Firefox | **AI** | **RTX Video** | 視頻串流 | **NVIDIA** | **GeForce RTX** | 超解析度 | 高動態範圍 | **HDR**
線上視頻 | **瀏覽器**

2024-05-15



**Explore What's Next in AI
With the Best of GTC**

Watch On Demand

Mozilla Firefox，這款深受喜愛的開源瀏覽器，通過整合 NVIDIA 的 AI 驅動 RTX Video 技術，在視頻品質上取得了重大突破。這一創新步驟利用本地 NVIDIA RTX GPU 來提升 Windows PC 和工作站上的視頻串流體驗。

什麼是 RTX Video？

首次在 CES 2023 上亮相，RTX Video 是一套 AI 驅動的視頻增強技術，旨在提升在線視頻的視覺效果。它包括兩個主要功能：

1. RTX Video 超解析度：這個工具可以升級低解析度視頻，使其看起來更清晰、更詳細。通過深度學習，它預測並生成更高解析度的版本，減少壓縮失真並增強清晰度。
2. RTX Video 高動態範圍：此功能通過添加高動態範圍（HDR）信息來增強標準動態範圍（SDR）內容。結果是視頻更加鮮豔、詳細且視覺上更具吸引力。

鑑於90%的在線視頻是1080p或更低解析度且為SDR，啟用RTX Video 可以顯著改善日常觀看體驗。

如何在 Firefox 中啟用 RTX Video

要享受這些增強功能，用戶需要最新的 GeForce RTX Game Ready Driver、NVIDIA Studio 或 NVIDIA RTX Enterprise Driver。然後，啟用 Windows HDR 功能並在 NVIDIA 控制面板中調整設置以開啟「超解析度」和「高動態範圍」。

更廣泛的影響

通過支持 RTX Video，Firefox 加入了其他流行瀏覽器如 Google Chrome 和 Microsoft Edge，以及視頻播放器如 VLC 的行列。這一舉措有望顯著提升廣大用戶的視頻串流品質，使在線內容更加身臨其境和愉悅。

敬請期待更多有關 AI 如何持續改變我們數位體驗的更新！

[閱讀更多](#)

Google Cloud AI 如何幫助發掘英格蘭的下一個足球明星

Google Cloud AI | 英格蘭足球協會 | FA | 人才發掘 | BigQuery | Vertex AI | 機器學習 | 人工智慧 | 球探報告

2024-05-15



英格蘭足球協會 (FA) 正在利用尖端技術來更有效地發掘未來的人才。通過使用 Google Cloud 的 AI 能力，FA 將大量的球探報告轉化為可行的見解。

傳統上，球探會對年輕球員進行詳細觀察，並編寫大量報告。這些報告的龐大數量讓人難以找出突出的人才。為了解決這個問題，FA 與 Google Cloud 合作，使用 BigQuery 這個數據倉庫解決方案，將數據集中並簡化處理。

現在，通過 Google 的機器學習平台 Vertex AI，這些球探報告被轉變為簡明的摘要。這種生成式 AI 方法提供了多個優點： - 一目了然的評估：快速摘要提供了對球員發展的即時見解。 - 更深入的背景：平衡球探的主觀評估和客觀數據。 - 一致性：消除個別球探的偏見，呈現球員表現的全面觀點。

這種創新的 AI 應用有可能徹底改變人才發掘的方式，確保有潛力的球員能夠被有效地認識和培養。

[閱讀更多](#)

Microsoft 在 CHI 2024 大放異彩：引領以人為本的設計

以人為本設計 | 自然語言界面 | 資訊多樣性 | 動態圖形 | 神經輻射場 | 拖延行為 | LLM | 虛擬協作

2024-05-15



在 2024 年 ACM CHI 計算系統人因學會議上，Microsoft 展示了其對以人為本設計的承諾，並推出了一系列創新研究項目。這些項目旨在加強我們與技術的互動，使其更加易於使用、個性化和具吸引力。

DynaVis 引入了動態合成的 UI 小工具來編輯視覺化，結合自然語言界面以簡化用戶交互。

Generative Echo Chamber? 探索了基於 LLM 的搜尋系統可能如何影響資訊多樣性，強調了平衡資訊查詢的必要性。

Piet 在動態圖形中促進顏色創作，通過互動式調色板和主題變更簡化設計過程。

SharedNeRF 利用神經輻射場技術實現即時、照片般真實的遠程協作，提升虛擬協作體驗。

Understanding Academic Procrastination 研究了 LLM 如何個性化干預拖延行為，提供以截止日期為導向的建議和適應性問題。

這些創新強調了 Microsoft 對改善數位體驗的承諾，專注於人類需求，為更直覺和沉浸式的技術互動鋪平了道路。

[閱讀更多](#)

GAAD 2024：Google 發表 AI 驅動的無障礙功能

GAAD 2024 | Google | 無障礙功能 | Android | Google Maps | Lookout | Look to Speak | Project Gameface | Project Relate | 聲音通知

2024-05-16



為了慶祝全球無障礙意識日（GAAD），Google 在多個平台上推出了八項新的無障礙更新，包括 Android 和 Google Maps。這些進步是為了與有殘疾的人一起設計的，旨在增強全球超過 13 億人的數位使用和包容性。

Lookout 增強功能

Android 上的 Lookout 現在具有測試版的「查找模式」，允許用戶使用手機相機找到特定物體，如座位或廁所。它還為圖像提供了 AI 生成的說明，為視覺障礙者提供更多上下文。

Look to Speak

Look to Speak 應用程式引入了一種無文字模式，允許用戶使用可自定義的表情符號、符號和照片進行交流，這使得有認知差異或識字困難的人更容易表達自己。

Project Gameface

開發人員現在可以訪問 Project Gameface，這是一款開源的、免手操作的 Android 設備遊戲滑鼠。這個工具允許面部表情自定義，旨在增強教育和工作環境中的無障礙性。

Google Maps 更新

Google Maps 在全球範圍內擴展了詳細的語音導航和螢幕閱讀功能，幫助視障用戶更有效地導航。此外，現在提供了超過 5000 萬個地方的無障礙資訊，包括無障礙入口和 Auracast 廣播音訊等屬性，以增強聽力支援。

Project Relate 和聲音通知

Project Relate 現在允許用戶自定義語音識別模型，從其他應用程式中使用短語。基於用戶反饋，聲音通知系統已重新設計，改進了門鈴或煙霧警報等家庭聲音的提示系統。

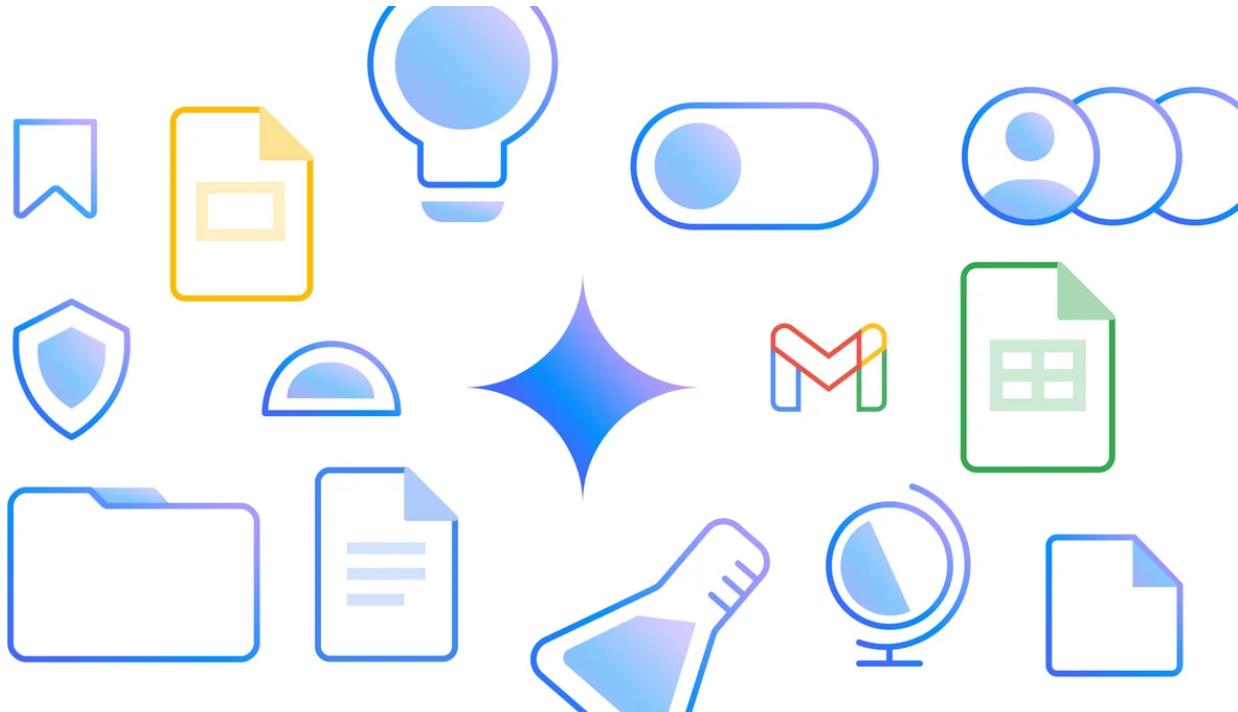
這些更新強調了 Google 對使科技更加包容和無障礙的承諾。

[閱讀更多](#)

Google AI : Gemini 透過 Workspace 整合提升教育

[Google AI](#) [Gemini](#) [Workspace](#) [教育](#) [生成式 AI](#) [Docs](#) [Gmail](#) [Slides](#) [課程計畫](#) [Email 總結](#)
[簡報](#) [OpenStax](#) [Data Commons](#) [練習測驗](#) [數據保護](#) [教育機構](#)

2024-05-16



Google AI : Gemini 透過 Workspace 整合提升教育

Google 正在透過為 18 歲及以上的用戶推出的 Gemini for Google Workspace 改變教育領域。這個最先進的生成式 AI 工具旨在透過與受歡迎的 Workspace 應用程式如 Docs、Gmail 和 Slides 的無縫整合來豐富教學和學習。

Gemini 幫助教育者設計課程計畫、總結電子郵件並為簡報生成獨特的圖像，提高創造力和生產力。值得注意的是，它包括 OpenStax 和 Data Commons 擴充功能等高級功能，用於可信的教育內容，並提供引導式練習測驗以增強學習。

此外，Google 確保企業級數據保護，保護用戶資訊，並免費為教育機構提供這些保護措施。

Gemini 通過兩個新的付費附加元件提供：經濟實惠的 Gemini Education，以及包括 AI 驅動的 Meet 筆記功能在內的全方位功能 Gemini Education Premium。

從 5 月 23 日起，教育機構可以整合 Gemini，開啟 AI 強化教育的新時代。

智能自動化崛起為戰略差異化因素

智能自動化 | 機器人流程自動化 | 光學字符識別 | 生產力 | 合規 | 健康保險 | 財務

2024-05-17



智能自動化崛起為戰略差異化因素

智能自動化 (IA) 正在從操作工具演變為各行各業的戰略強大力量。SS&C Blue Prism委託 Forrester Consulting進行的最新研究強調了這一轉變，顯示出每位客戶從生產力和合規成本避免等收益中獲得的淨現值達到驚人的5340萬美元。

強調實際應用，SS&C Blue Prism的技術已為醫療保健客戶節省了數百萬美元。其中一個例子是一個幫助護士追回加班費的機器人，在疫情期間提供了4300萬美元的效益。這些系統基於機器人流程自動化 (RPA) 和光學字符識別 (OCR) 等技術，簡化了數據提取和流程自動化。

SS&C Blue Prism的Chorus，一種流程協作工具，被認為是該領域的領導者。該工具每天促進數百萬次交易，提高了遠超傳統金融和保險行業的效率。

本質上，智能自動化為公司提供了一種變革性的方式來重新構想其流程，將操作任務轉化為戰略優勢。

[閱讀更多](#)

IBM 和 Tech Mahindra 發布可信賴的 AI 與 Watsonx

IBM | Tech Mahindra | AI | Watsonx | 數據平台 | 合作 | 可信賴 AI | 自動化 | 虛擬卓越中心 | 數位轉型
生成 AI | 混合雲 | 本地環境

2024-05-17



在人工智慧 (AI) 領域取得重大進展，IBM 和 Tech Mahindra 合作推出了一個旨在通過 IBM 的 Watsonx AI 和數據平台推動可信賴 AI 的合作計劃。

這個合作將 IBM 的 Watsonx 能力與 Tech Mahindra 的 AI 諮詢和工程專業知識結合在一起，創造出大量的新生成 AI 服務和解決方案框架。目標是幫助組織利用其可信數據來自動化流程，構建可靠的 AI 模型，並確保在混合雲和本地環境中的可擴展性。

Tech Mahindra 還推出了一個虛擬的 Watsonx 卓越中心 (CoE)。這個 CoE 將作為創新中心，允許專門的團隊開發結合兩家公司優勢的獨特 AI 解決方案。

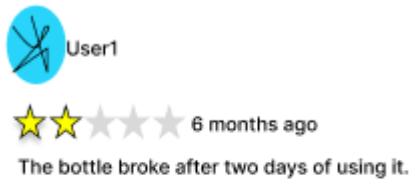
這些合作努力預期將推動數位轉型的進步，提升生產力，並通過負責任的 AI 實踐促進業務增長。這項計劃為生成 AI 的開發和部署設定了新的標準，強調透明性、可解釋性和信任。

[閱讀更多](#)

利用 AWS 建立高效能 AI 工作流程： 提示鏈結與人機互動指南

AWS 高效能 AI 工作流程 提示鏈結 人機互動 Foundation Models LLMs 事件驅動架構
Amazon Comprehend DetectToxicContent Amazon EventBridge

2024-05-17



Response from the product team

We are sorry to hear about your experience. We are happy to send a replacement. Please give us call to our customer care number

利用 AWS 建立高效能 AI 工作流程：提示鏈結與人機互動指南

生成式 AI 正在革新內容創作，從對話、故事到圖片和影片。這種創新背後的核心是 Foundation Models (FMs) 和 Large Language Models (LLMs)，它們根據輸入的提示生成回應。然而，當任務變得更加複雜時，這些模型可能會出現失誤。此時，「提示鏈結」便派上用場。

提示鏈結的概念是將複雜的任務分解為較小的、可管理的子任務。每個子任務按特定順序提供給 LLM，從而提高準確性和效率。例如，一間零售公司可能使用這種技術來自動回應顧客評論。系統利用 LLM 生成回應，但將不確定的回應標記給人工審核，以確保品質。

AWS Step Functions 編排這個工作流程，結合事件驅動架構 (EDA) 與人機互動過程。例如，當一則評論被發布時，Amazon EventBridge 會觸發一個工作流程。系統使用 Amazon Comprehend 的 DetectToxicContent API 評估評論內容。如果內容的毒性評分不明確，則將其發送給人工進行評估。人類判斷的整合提高了 AI 生成內容的可靠性。

這種方法不僅通過平行處理任務來加速過程，還通過聚焦提示來提高輸出品質。對於希望利用 AI 同時保持高客戶互動標準的公司來說，這是一種實用的做法。

對於那些希望實施這種工作流程的公司，AWS 在其 GitHub 儲存庫中提供了詳細的指導和示例。提示鏈結與人機互動機制的結合確保生成式 AI 應用既準確又安全。

[閱讀更多](#)

Google 支持中東歐AI發展的新舉措

Google | INSAIT | AI | 投資 | 雲端運算 | TPU | 博士獎學金 | 中東歐

2024-05-17



Google 宣布將投資超過 200 萬美元，支持位於保加利亞索非亞的計算機科學、人工智慧與技術研究所 (INSAIT)。這項計畫再次證明了 Google 對於培育中東歐地區 AI 研究與人才的承諾。

以下是這筆投資的內容： - 雲端運算資源：超過 100 萬美元將用於提供 INSAIT 使用 Google Cloud 的 Tensor Processing Units (TPUs)。這些專門的處理器設計用於機器學習，使研究人員能夠建立先進的 AI 模型並進行開創性的研究。 - 博士獎學金：另外 100 萬美元將用於資助八個博士獎學金。這旨在吸引並支持該地區最優秀的人才，培養下一代 AI 專家。

自 2022 年成立以來，INSAIT 已迅速成為領先的研究中心，致力於打造一個吸引並留住頂尖人才的世界級環境。這筆新投資建立在 Google 2022 年的初始 300 萬美元捐款基礎上，這筆捐款同樣支持了計算資源和獎學金。

Google 對 INSAIT 的持續支持突顯了中東歐地區在 AI 領域的巨大潛力。通過投資像 INSAIT 這樣的機構，Google 旨在使 AI 研究更加包容和創新，最終為該地區的高科技經濟繁榮做出貢獻。

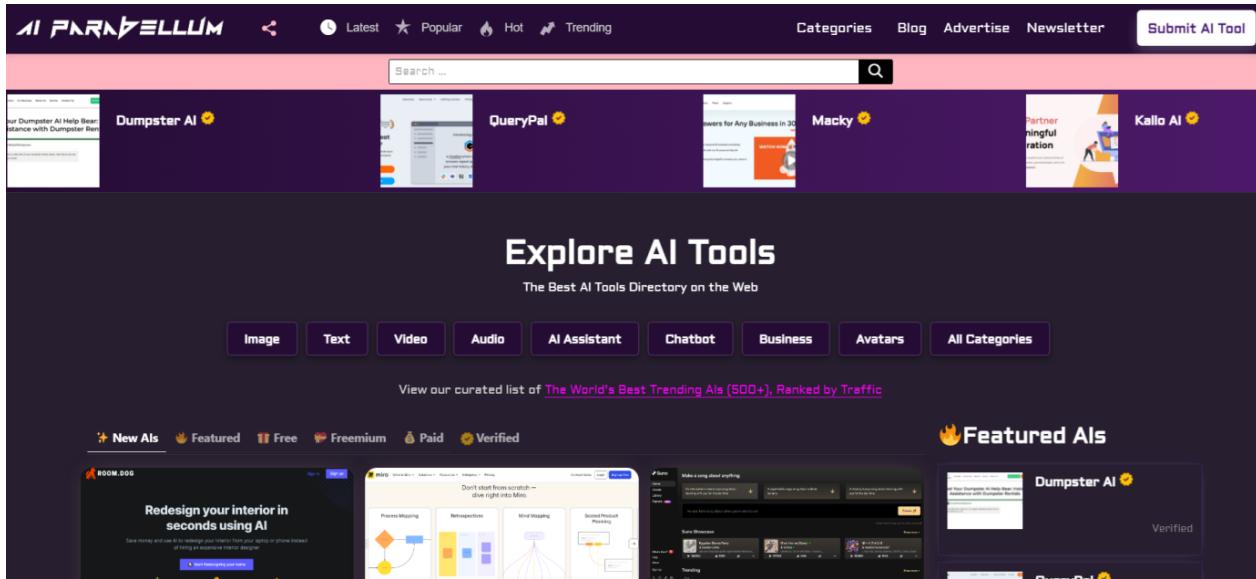
Google 和 INSAIT 正攜手努力，邁向一個 AI 發展惠及所有人的未來。

[閱讀更多](#)

探索 AI 的未來：介紹 AI Parabellum

AI | 人工智慧 | 工具目錄 | 開發者 | 研究人員 | 創新者

2024-05-20



探索 AI 的未來：介紹 AI Parabellum

在不斷發展的人工智慧領域，保持對最新工具的了解可能會讓人感到不知所措。這時候，AI Parabellum 登場了，這是一個先進的 AI 工具目錄，旨在簡化這項任務。AI Parabellum 作為一個綜合資源，將開發者、研究人員和愛好者聚集在一起，共同探索和評估最尖端的 AI 工具。

瀏覽 AI Parabellum 是一種無縫的體驗。用戶可以瀏覽範圍廣泛的類別，例如圖像、文本、視頻和音頻工具。這個目錄甚至按流量對超過 500 種 AI 工具進行排名，提供了一目了然的市場上最受歡迎和最有效解決方案的概覽。

但 AI Parabellum 不僅僅是用於發現；它也是創作者的平台。通過提交 AI 工具頁面，創新者可以展示他們的作品，接觸潛在的用戶和投資者。

對於任何希望站在 AI 最前沿的人來說，AI Parabellum 是一個不可或缺的資源，幫助你揭示這個革命性領域的未來。

[閱讀更多](#)

AI是否正在改變賭博業？

人工智慧 | 賭博 | 線上賭場 | 玩家行為分析 | 虛擬現實 | 個性化優惠 | 欺詐檢測 | 遊戲圖形

2024-05-20



人工智能 (AI) 正在迅速改變我們生活的許多方面，包括賭博產業。雖然AI的未來影響仍不確定，但其當前的應用已經備受矚目。

當前應用：AI廣泛應用於線上賭場的後端管理，確保遊戲運行順暢高效。它還在創造具挑戰性的對手方面發揮了重要作用，提升了整體遊戲體驗。

此外，AI還分析玩家行為以確保安全和個性化。它能檢測可能暗示欺詐的不尋常活動，並根據個人偏好定制優惠，從而提升玩家的體驗並促進負責任的賭博。

未來前景：展望未來，AI有望顯著改善遊戲圖形和虛擬現實體驗。增強的玩家分析將導致對玩家偏好的更準確預測，增加個性化優惠和忠誠度獎勵。

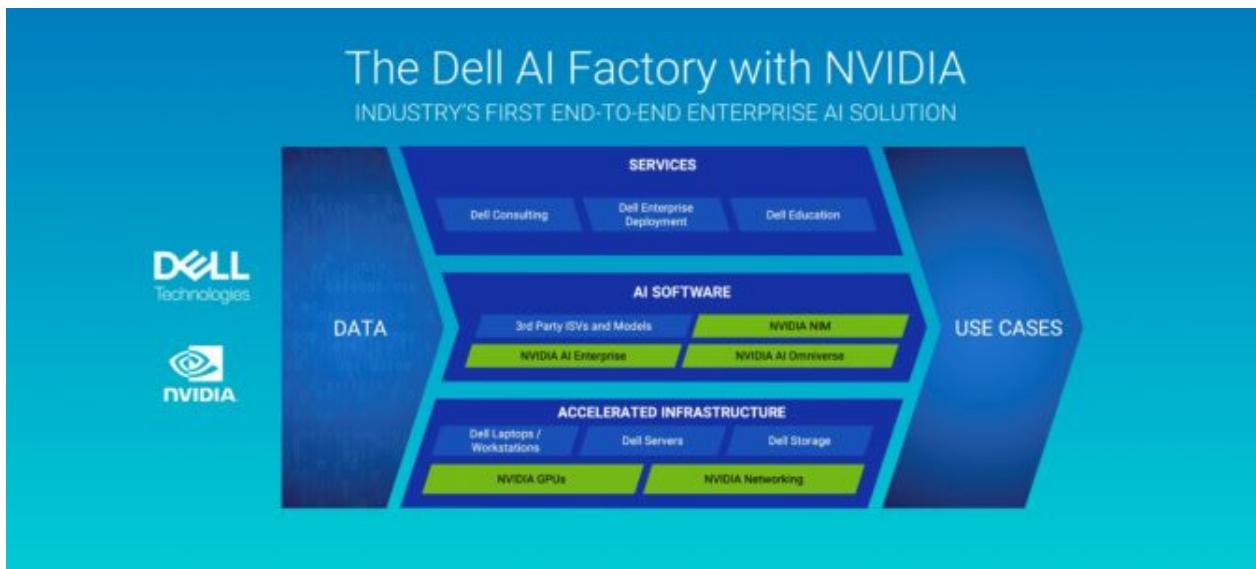
隨著AI技術的持續進步，賭博產業無疑會看到更多創新的變化，為我們展現更具沉浸感和安全的遊戲未來。

[閱讀更多](#)

NVIDIA 和 Dell 發布突破性 AI 創新

NVIDIA | Dell | AI | 生成式 AI | 數據中心 | Edge AI | PowerEdge XE9680L | 液冷系統 | Tensor Core GPU | AI Enterprise | NIM 推理微服務 | AI 工廠 | AI 解決方案

2024-05-20



在 Dell Technologies World 大會上，NVIDIA 執行長 Jensen Huang 強調了一個變革性的願景：每家公司都將成為“智慧製造商”。Huang 與 Dell 執行長 Michael Dell 一同強調生成式 AI 對全球產業的革命性影響，將現代數據中心比作將數據轉化為有價值智慧的 AI 工廠。

宣布的主要更新包括 Dell PowerEdge XE9680L 伺服器，配備八個 NVIDIA Blackwell Tensor Core GPU 和液冷系統。這種配置為大規模 AI 部署提供了最密集、最節能的解決方案。

此外，Dell NativeEdge 平台現在自動化 NVIDIA AI Enterprise 軟體交付，幫助開發者和 IT 操作人員快速部署邊緣 AI 應用。該平台經過 NVIDIA NIM 推理微服務的增強，簡化了 AI 應用開發，實現更快的價值轉化。

透過與 NVIDIA 的合作，Dell AI Factory 承諾提供從數據中心到邊緣環境的一整套 AI 解決方案。這一合作旨在為各行業提供先進的 AI 能力，確保數據安全並促進定制化 AI 服務的開發。

[閱讀更多](#)

Microsoft 推出具備先進 AI 功能的 Copilot+ 電腦

**Microsoft | Copilot+ | AI 功能 | Surface Laptop | Surface Pro | Qualcomm Snapdragon | AI 加速器
Recall 功能**

2024-05-21



Microsoft 推出了其在個人計算領域的最新創新：Copilot+ 電腦。這些新設備配備了先進的 AI 功能，旨在將更多的計算能力帶到使用者本地，減少對雲端資料中心的依賴。

Copilot+ 系列包括 Surface Laptop 和 Surface Pro，這兩款設備均搭載了 Qualcomm Snapdragon X 系列晶片。這些晶片內建 AI 加速器，每秒可達 45 兆次運算 (TOPS)，大幅提升性能。Surface Laptop 有 13.5 吋和 15 吋型號可選，最高配備 64GB RAM 和 1TB 閃存。Surface Pro 是一款多功能二合一電腦，透過可拆卸的 Flex 鍵盤轉變為筆記型電腦，而這款鍵盤同時也充當無線連接器和觸控筆充電器。

Copilot+ 系列的一個亮點功能是“Recall”，讓使用者能夠記錄並儲存所有活動，創建一個可搜索的歷史記錄，數月後仍可訪問。這種創新方法旨在通過在設備上直接增強 AI 助理的功能，重新定義使用者體驗。

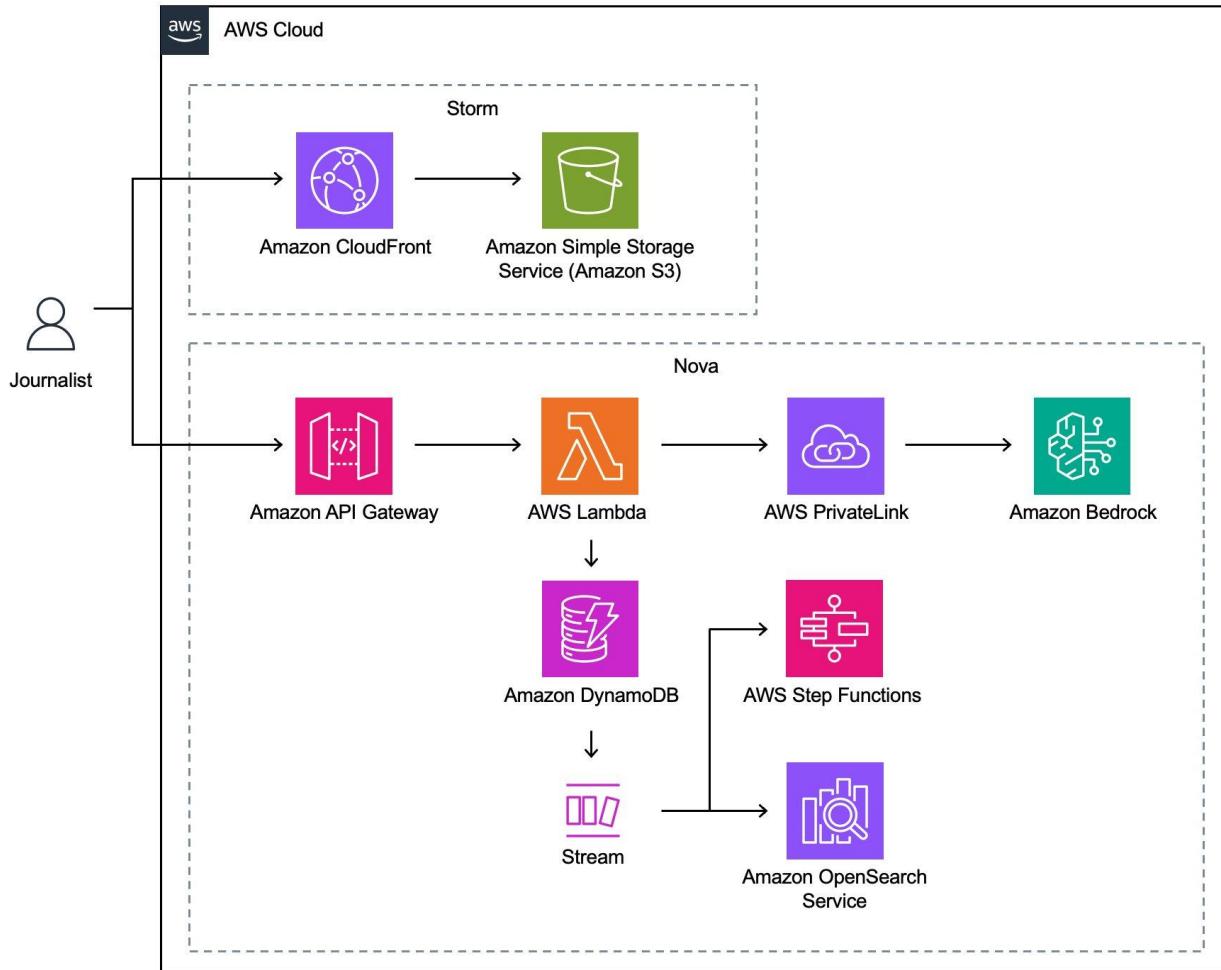
這些 AI 驅動的電腦將在 6 月 18 日開始出貨，定價從 \$1,000 起，標誌著 Microsoft 在將最先進技術整合到日常計算中的重大飛躍。

[閱讀更多](#)

使用 Amazon Bedrock 提升新聞業的 生成式 AI

Amazon Bedrock | 生成式 AI | 20 Minutes | 新聞業 | 數位出版 | SEO | 內容自動化 | 品牌安全 | AWS PrivateLink

2024-05-21



20 分鐘革命性提升新聞室效率和參與度

法國媒體 20 Minutes 正在利用 Amazon Bedrock 的生成式 AI 技術來改變其數位出版流程。擁有每月 1900 萬讀者的 20 Minutes，旨在透過整合尖端 AI 技術來提升新聞品質和讀者參與度。

簡化重複性工作

其中一項關鍵創新是自動化重複性工作，如生成文章摘要、分類內容和標籤。透過部署 Amazon Bedrock 的強大基礎模型，20 Minutes 的記者可以專注於核心報導工作，而 AI 則處理對搜尋引擎優化（SEO）至關重要的元數據生成。

高效內容重發

另一個重要應用是重發新聞社的稿件。使用 Amazon Bedrock，20 Minutes 開發了一個 AI 工具來重寫新聞社內容，保持編輯標準並優化 SEO。這個 AI 驅動的過程大大減少了手動工作量，使記者能夠花更多時間創作原創內容。

確保品牌安全

為了為廣告商提供一個品牌安全的環境，20 Minutes 使用生成式 AI 來評估單篇文章的品牌安全性。透過對內容進行行業指導原則分析，AI 提供了即時的、精細的安全評估，確保廣告放置在合適的上下文中。

技術整合

20 Minutes 通過 AWS PrivateLink 使用 Amazon Bedrock，無縫整合到他們現有的無伺服器架構中。這種設置確保了安全、私密和高效的 AI 模型訪問，增強了他們的數位編輯工具。

生成式 AI 正在成為像 20 Minutes 這樣的數位出版商的改變遊戲規則，為更創新和高效的新聞室運作鋪平道路。20 Minutes 與 Amazon Bedrock 的合作展示了 AI 如何在數位時代提升新聞業。

[閱讀更多](#)

NVIDIA 與 Microsoft 聯手加速 AI 發展

NVIDIA Microsoft AI GPU TensorRT-LLM Azure Phi-3 cuOpt DirectX DirectML

2024-05-21



Explore What's Next in AI With the Best of GTC

Watch On Demand

NVIDIA 擴展了與 Microsoft 的合作夥伴關係，以簡化 AI 應用程式的部署，使開發者更容易利用人工智慧的力量。在最近的 Microsoft Build 活動中，NVIDIA 展示了多種與 Microsoft Azure 和 Windows 電腦整合的解決方案，這些解決方案可以簡化 AI 模型的部署並提升效能。

其中一個亮點是為 NVIDIA GPU 優化 Microsoft 的 Phi-3 系列小型語言模型。這些模型，包括新的 Phi-3-vision 多模態模型，利用 NVIDIA TensorRT-LLM，可以作為 NVIDIA NIM 推理微服務使用。這使得開發者可以通過標準 API 高效地部署高階 AI 模型。

在一個重要的舉措中，NVIDIA 用於路徑優化的 GPU 加速 cuOpt 服務現在已在 Azure Marketplace 上線。這個微服務使用平行算法實時管理物流，在優化車輛路徑和減少閒置時間方面，為物流和供應鏈行業提供了潛在的節省。

此外，NVIDIA 也在優化由 NVIDIA GeForce RTX 和 NVIDIA RTX GPU 提供動力的電腦上的 AI 效能。NVIDIA DirectX 驅動程式和 DirectML 的增強功能為大型語言模型提供更快的推理效能，讓致力於下一代 AI 應用程式的開發者受益。

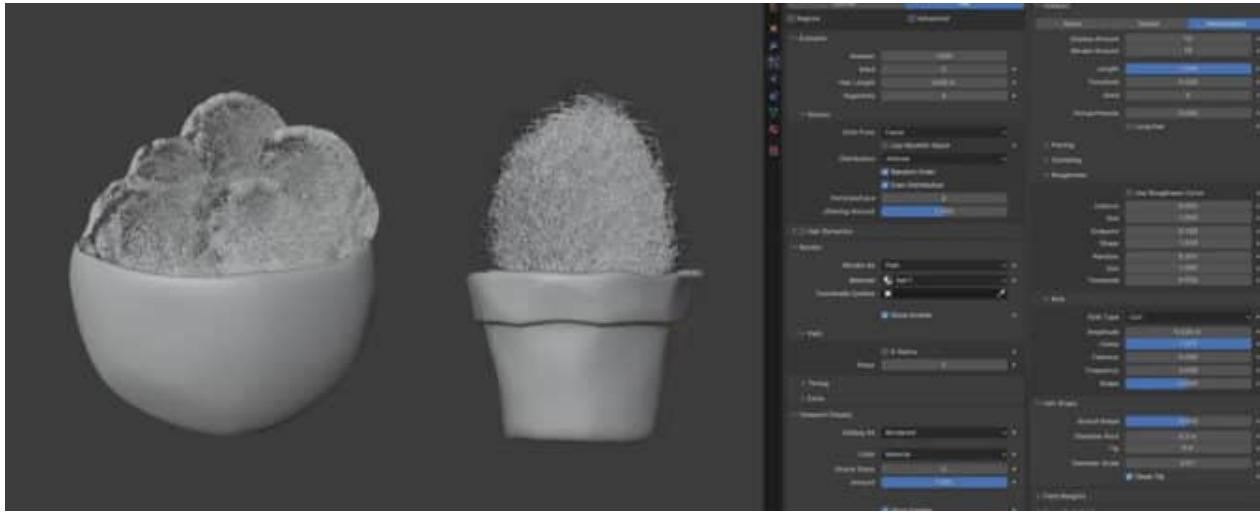
這些進步展示了 NVIDIA 與 Microsoft 持續合作以賦能開發者，確保他們擁有必要的工具來高效地創建高效能的 AI 應用程式。

[閱讀更多](#)

NVIDIA五月Studio驅動程式革新內容創作

NVIDIA | Studio驅動程式 | GPU加速 | 人工智慧 | RTX Video | Mozilla Firefox | NVIDIA RTX Remix
Maxon's Cinema 4D | NVIDIA Omniverse Audio2Face | 創意應用

2024-05-21



NVIDIA最新的五月Studio驅動程式釋出在創意社群中掀起了波瀾，推出了一系列創新、GPU加速及人工智慧驅動的功能來提升內容創作。

創意應用的增強功能：

- RTX Video：
 - 此功能提升了YouTube和Disney+等平台的視頻質量。RTX Video超解析度 (VSR) 可為視頻進行升級，提供更清晰的畫面，而RTX Video HDR則將標準視頻轉換為高動態範圍 (HDR10)，提高可見度和色彩鮮豔度。
- Mozilla Firefox：
 - 現在支援RTX VSR和HDR，為大多數串流視頻提供人工智慧增強的升級和HDR效果。
- NVIDIA RTX Remix：
 - 讓模組製作者能夠捕捉遊戲資產並使用人工智慧工具提升素材。新增的DLSS 3.5支援，搭配光線重建功能，在光線追蹤遊戲中提供更高質量的圖像。
- Maxon's Cinema 4D：
 -

- 最新版本整合了模擬系統，允許對復雜場景中的發射場和互動進行精確控制。這些模擬受益於RTX GPU加速。
- NVIDIA Omniverse Audio2Face：
 - 此應用程式可從音頻輸入生成具有表情的面部動畫。最新版本支援多語言唇同步和唱歌動畫，並增強了細緻的面部表情。

在一次令人欣喜的展示中，3D藝術家Yao Chan使用她的NVIDIA RTX GPU創建了一個以春天為靈感的3D場景。她的作品突顯了人工智慧驅動的模擬和RTX加速渲染的無縫整合，展示了這些新工具的潛力。

這些更新現已提供給NVIDIA和GeForce RTX GPU用戶，保證能簡化和提升創作流程。

[閱讀更多](#)

加速 GenAI 應用程式開發與增強追蹤和除錯

GenAI | prompt flow SDK | 追蹤 | 除錯 | RAG | AI Agent | Azure AI Studio | 性能 | 監控 | 優化

2024-05-21



Generative AI (GenAI) 應用程式的開發正在快速演進，從基本的提示工程進化到更複雜的系統。這些先進的系統包括一系列複雜的步驟，如檢索、前處理、後處理和記憶體管理，這些步驟以稱為 Retrieval-Augmented Generation (RAG) 和 AI Agent 情景的混合方式進行協作。

雖然這種複雜性增強了能力，但也帶來了挑戰。開發人員必須掌握工作流程中的每一步，以獲得洞察力、排除問題並改進應用程式。隨著越來越多的 GenAI 應用程式從實驗階段轉向全面生產，部署後的可觀測性變得至關重要。這涉及監控應用程式性能、回應用戶反饋和調試挑戰性案例。

這時，prompt flow SDK 出現了，顯著簡化了這個過程。SDK 允許開發人員從輸入到輸出追蹤和監控他們的 GenAI 應用程式，提供對中間結果和執行時間的洞察。這種可見性有助於調試和優化應用程式。

追蹤功能支持各種端點，包括本地環境和 Azure AI Studio。這種靈活性促進了在不同框架中的測試、評估和部署。基於雲的 Azure AI Studio 追蹤進一步增強了協作和測試歷史管理。

增強的追蹤功能為每個功能提供詳細的日誌和指標，幫助調試和優化。開發人員可以監控多代理情景中的對話流程，改進整體應用程式性能，並確保無縫部署到 Azure AI Studio，以實現安全和可擴展的操作。

總結來說，通過 prompt flow SDK 增強的追蹤和除錯功能使開發人員能夠有效地管理 GenAI 應用程式的複雜性，確保強大的性能和優化的資源利用。

[閱讀更多](#)

釋放 AI 潛力：Azure AI MaaS 與 ClearML、Arize AI 和 Dataloop 的整合提升 LLM 模型的可及性

Azure AI MaaS | ClearML | Arize AI | Dataloop | LLM | GenAI | API | 整合 | 模型 | 數據管理

2024-05-21



Microsoft 已經宣布了一個重大進展，即將先進的 AI 模型的可及性提升至新高度，通過將 Azure AI Models as a Service (MaaS) 整合到 ClearML、Arize AI 和 Dataloop 平台，這一舉措旨在簡化生成式 AI (GenAI) 應用的開發，提供企業級且安全的按需付費 API。

Arize AI 整合：Arize AI 提供了評估、追蹤和觀察 AI 模型的工具，確保開發者能夠自信地部署複雜的 GenAI 系統。通過 Azure AI MaaS，用戶可以訪問到像 Llama 3 和 Cohere Command R+ 等先進模型，簡化開發過程，並提供增強的可見性和調試能力。

Dataloop 整合：Dataloop 加速了 AI 項目中非結構化數據管道的處理。通過這次整合，開發者能夠更好地管理和標註數據、精煉模型並快速部署 AI 解決方案。這一合作關係提高了數據管理和模型精準度，使 AI 構建過程更加高效。

ClearML 整合：ClearML 專注於自動化和擴展 AI 工作流程。通過整合 Azure AI MaaS，用戶可以優化其 AI 開發流程，同時保持數據隱私和安全性。ClearML 的低代碼工具和企業級安全功能確保了流暢和安全的模型開發與部署。

這些整合標誌著在使先進 AI 工具變得更為可及和高效方面邁出了重要一步，賦予更廣泛的開發者和企業在競爭激烈的科技領域中創新和蓬勃發展的能力。

[閱讀更多](#)

Microsoft 與 Hugging Face 提升生成式 AI 合作

Microsoft | Hugging Face | 生成式 AI | Azure AI | AMD | Visual Studio Code

2024-05-21



來自人工智慧領域的振奮消息！Microsoft 與 Hugging Face 宣布了一項深化合作，旨在讓生成式 AI 更加易於接觸且更高效。此次合作重點在於增強 Azure AI 模型目錄並整合 Hugging Face 的尖端模型和工具。

主要亮點：

1. Azure AI 目錄中的新模型：
2. Microsoft 已經將 20 種受歡迎的 Hugging Face 模型，包括 Rhea-72B-v0.5 和 Multiverse-70B，添加到 Azure AI 模型目錄中。這項擴展為開發者提供了更多 AI 專案的選擇。
3. 增強的 Azure 基礎架構：
4. 與 AMD 合作，Microsoft 已經優化了 Azure AI 基礎架構，以支援在 ND MI300X v5 上的 Hugging Face 模型，利用 AMD GPU 提升效能和效率。
5. HuggingChat 上的 Phi-3-mini：
6. Microsoft 的小型模型家族 Phi-3 現在可在 HuggingChat 上使用，使開發者能夠輕鬆實驗和利用這些模型。
7. Visual Studio Code 整合：
8. Hugging Face Spaces 中的新“開發模式”功能允許與 Visual Studio Code 無縫連接。開發者現在可以更高效地編輯和部署 AI 驅動的演示。

這次合作旨在為 AI 開發者提供強大的工具和基礎設施，簡化可擴展且負責任的 AI 解決方案的開發和部署。

[閱讀更多](#)

Azure AI Studio 與變革性的AI開發

Azure AI Studio | 生成式人工智慧 | 負責任AI | 安全性 | 數據檢索 | 超過1600個模型 | GPT | AI
Copilot | **SuperTOBI** | **AI Tax Assist**

2024-05-21



生成式人工智慧 (Generative AI) 一直是熱門話題，重塑了企業開發軟體和利用技術的方式。然而，生成式人工智慧的不可預測性可能導致錯誤，如創建不存在的網址或編造數據。認識到這一點，企業專注於通過嚴格的測試和持續的監控來提高人工智慧的質量。

進入 Azure AI Studio，這是Microsoft的平台，專為負責任且安全地開發和部署生成式人工智慧應用程式而設計。Azure AI Studio 加速了生成式人工智慧的開發週期，提供自訂和配置功能，並有 Azure等級的安全性、隱私和合規性作為後盾。

Azure AI Studio的主要功能：

1. Pro-Code 平台：使用視覺和代碼優先工具自訂和配置生成式人工智慧應用程式。與Azure AI 服務整合，簡化了助手的創建過程。
2. 模型和API整合：支持來自Meta、Microsoft和OpenAI等提供商的超過1600個模型。模型目錄提供簡便的部署選項和安全驗證。
3. 開發工具鏈：一個資源共享、基於團隊的開發和專案管理的中心。像提示流SDK等功能支持高效調試和追蹤。
4. 數據和安全：Azure AI Search支持數據檢索，而內容安全措施則防止有害輸出。
5. 可擴展性：促進大規模部署和持續監控，確保靈活的資源管理和治理。

實際應用案例：

- Sweco's GPT：增強了項目信息檢索和文件分析。
- Parloa's AI Copilot：簡化了多語言客戶服務。
- Vodafone's SuperTOBI：為帳單和網絡問題提供個性化客戶支持。
- H&R Block's AI Tax Assist：通過對話式人工智慧簡化報稅過程。

Azure AI Studio 讓開發人員能夠構建創新且負責任的人工智慧解決方案，提高各行業的效率和轉型。立即探索其功能，引領下一代人工智慧開發！

[閱讀更多](#)

Microsoft 公布 AI 應用程式範本以簡化 AI 開發

Microsoft | AI應用程式範本 | AZD | Azure | Prompty | 提示工程 | LLM開發 | GitHub Actions

2024-05-21



Microsoft 推出了 AI 應用程式範本套件，旨在簡化建立 AI 驅動應用程式的過程。這些新工具旨在簡化開發過程，使企業更容易用 AI 功能來現代化其系統。

AI 專案入門範本 這些範本作為各種熱門應用場景的基礎 GitHub 資料庫。它們內建了對 Azure Developer CLI (AZD) 的支持，加速了從本地開發到 Azure 部署的過程。主要特點包括 GitHub Actions 用於 CI/CD 部署、模型性能評估以及行業最佳實踐。

新工具與 Azure Developer CLI AZD 現在包括一個針對生成式 AI 模型和工作流程的 `ai.endpoint` 主機類型。這個統一的工具管理從初始化到部署整個 Azure 應用的生命周期，並增強了對 AI 特定資源的支持。

提示工程與 Prompty Prompty 是一個新的語言無關的提示資產，整合到了像 Prompt flow、LangChain 和 Semantic Kernel 這樣的流行編排系統中。它簡化了創建和管理提示的工作流程，使 LLM 開發變得更加容易。

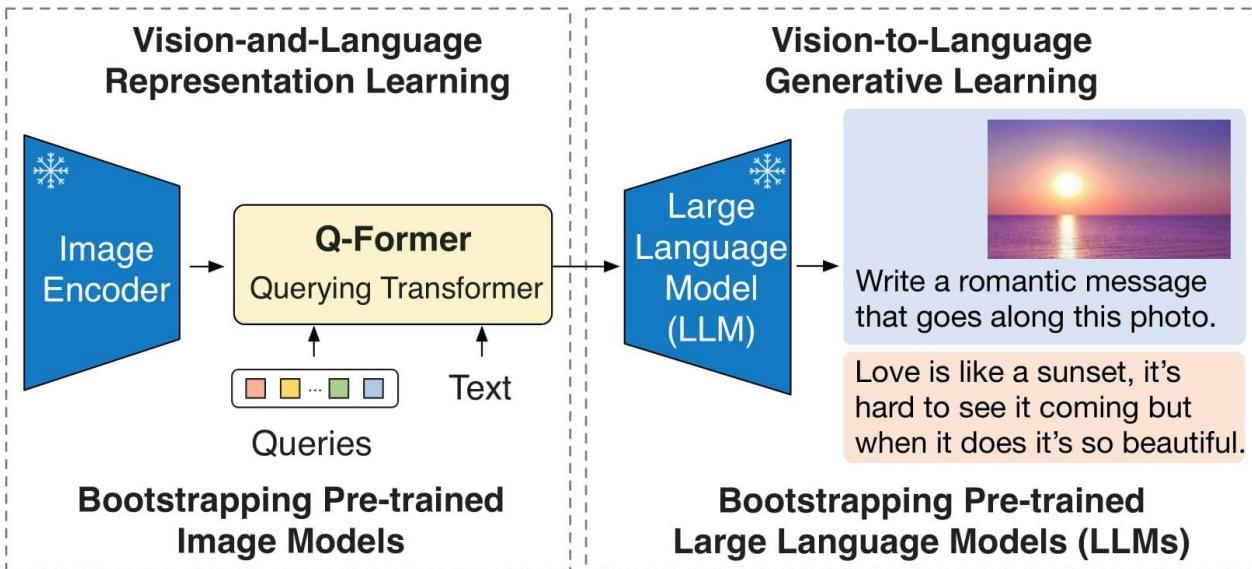
這些創新提供了一個強大的工具包，讓企業從初始開發到部署無縫整合 AI，提高生產力和創新。

[閱讀更多](#)

使用視覺語言模型和 Amazon Bedrock 自動化時尚產品描述

視覺語言模型 | **Amazon Bedrock** | 產品描述 | 電子商務 | 機器學習 | 自然語言處理 | 生成式 AI
 BLIP-2 | CLIP | Amazon SageMaker | 低秩適應 | Anthropic | Claude

2024-05-22



為眾多在線零售商品創建高質量的產品描述是一項艱巨的任務。近日，機器學習 (ML) 和自然語言處理 (NLP) 的進步提供了一個有前途的解決方案。這項技術不僅可以自動創建這些描述，還可以提高電子商務平台的可搜索性和個性化。

視覺語言模型 (VLMs) 和 Amazon Bedrock

使用像 BLIP-2 這樣的視覺語言模型 (VLMs)，我們可以直接從圖片中預測產品屬性。BLIP-2 模型結合了類似 CLIP 的圖像編碼器、Querying Transformer (Q-Former) 和大型語言模型 (LLM) 來生成時尚圖片的詳細屬性。

為了進一步完善這些預測，使用了時尚圖片的自定義數據集。這些數據幫助模型理解具體的細微差別，例如面料類型、合身度、圖案和袖長。使用 Amazon SageMaker，可以對 BLIP-2 模型進行微調，利用低秩適應 (LoRA) 等技術來保持效率。

使用 Amazon Bedrock 生成描述

一旦預測出產品屬性，Amazon Bedrock 就會介入。這個託管服務提供了來自領先 AI 供應商的高性能基礎模型的訪問。通過將預測的屬性輸入到像 Anthropic 的 Claude 這樣的生成模型中，Bedrock 會創建引人入勝且精確的產品描述。

例如，如果模型識別出「長袖棉襯衫」，Bedrock 會生成不僅包含這些細節還增添靈感的描述，提升購物體驗。

對電子商務的影響

使用這些技術自動化產品描述可以顯著改善搜索功能，使客戶更容易找到他們正在尋找的產品。此外，個性化的描述可以提升客戶滿意度並增加購買可能性。

隨著生成式 AI 領域的擴展，VLMs 和 LLMs 的整合為在線零售的內容生成帶來了強大的工具。對於那些渴望探索該技術的人，代碼和詳細步驟可在 GitHub 上找到，邀請您測試和改進該過程以適應自己的使用案例。

[閱讀更多](#)

SoftServe 和 Continental 透過 OpenUSD 和生成式 AI 推動數位化

SoftServe | **Continental** | **OpenUSD** | 生成式 AI | 數位化 | 虛擬代理 | 工業副駕駛 | **NVIDIA**
Omniverse | 汽車製造 | 維護工作流程

2024-05-23



汽車技術的進步日益增長，而 SoftServe 與 Continental 攜手合作，領導這一趨勢。他們共同推出了一款虛擬代理——工業副駕駛，這款代理由生成式 AI 驅動，旨在簡化汽車製造中的維護工作流程。

這項創新技術的核心是將通用場景描述 (OpenUSD) 框架整合到 NVIDIA Omniverse 平臺中。這個框架允許靈活和可擴展的數位轉型，從而使操作更加高效。工業副駕駛利用 OpenUSD 和生成式 AI 提供身臨其境的 3D 視覺化，幫助工程師監控和管理生產線。

通過提供逐步工作指示和根本原因分析，工業副駕駛減少了文檔流程並改善了維護程序，預計能降低 10% 的維護工作量和停機時間。這一合作不僅促進了多平台合作，還推動了汽車製造工作流程中的創新和效率。

通過探索 SoftServe 和 Continental 在 NVIDIA GTC 演講中了解更多這些變革性技術。

[閱讀更多](#)

三項改變職場的 Google AI 功能在 Android 上

Google AI | Android | 生產力 | 應用程式開發 | 商業安全 | Gmail | Gemini | Pixel | 錄音 | 圈選搜尋
Gemini Nano | Vertex AI Studio | AI 驅動保護 | Gemini Business & Enterprise

2024-05-23



Google 已在 Android 上推出了新的 AI 驅動工具，旨在提升生產力、應用程式開發和商業安全。以下是三項突破性的功能：

1. 提升員工生產力:
2. Gmail 上的 Gemini: 員工現在可以總結冗長的電子郵件線索，使用情境智能回覆獲得建議回覆，甚至可以在應用程式內提出問題——這一切都能在移動中完成。
3. Pixel 的錄音應用程式: 自動轉錄面對面的會議，使團隊更容易跟蹤討論和決策。
4. 圈選搜尋: 在您的 Android 設備上通過圈選圖片或突出顯示文字快速找到資訊。
5. 賦能開發者:
6. Gemini Nano: 這個工具允許開發者在應用程式中整合生成式 AI 功能，而不需要廣泛的 AI 專業知識。
7. Vertex AI Studio: 提供強大的計算資源和超過 130 種基礎模型，支持開發先進的 AI 應用程式。

8. 增強商業安全:

9. AI 驅動保護: 像防盜鎖定這樣的功能可以在設備被偷竊時幫助保護敏感信息。
10. Gemini Business & Enterprise: 提供強大的安全控制和管理設置，用於管理 AI 功能和用戶權限。

探索這些工具如何通過 Google AI 在 Android 上改變工作的未來。

[閱讀更多](#)

為人們設計 AI：與 Abigail Sellen 的見解

AI | 以人為本計算 | 人類認知 | 工作場所 | 經濟 | Microsoft Research | AICE | 協作 | 創造力 | 生產力

2024-05-23



在最新一集的 Microsoft Research Podcast 系列《Ideas》中，主持人 Gretchen Huizinga 與專注於以人為本計算的傑出科學家 Abigail Sellen 進行對話。Sellen 的旅程始於電腦還遠未用戶友好的時代，那時為人們設計技術是一個革命性的想法。

今天，Sellen 領導 Microsoft Research 的 AI、Cognition 和經濟 (AICE) 項目。這個跨學科項目旨在探索生成式 AI 如何影響人類認知、工作場所動態以及更廣泛的經濟。目標不僅僅是了解 AI 的潛在影響，而是幫助塑造一個 AI 增強人類技能並支持有意義工作的未來。

Sellen 強調設計符合人類需求、願望和價值觀的 AI 系統的重要性。通過利用多樣化的科學家、設計師和社會科學家團隊，AICE 項目旨在確保 AI 的開發建立在人類行為深刻理解的基礎上。

這一集強調了 AI 支持而不是取代人類能力的潛力。Sellen 的願景是未來 AI 作為一個協作的副駕駛，放大人類的創造力和生產力，同時保護基本的認知技能。

欲了解更多有關 AICE 項目的詳細信息，請訪問 Microsoft Research 頁面。

[閱讀更多](#)

革命化醫療保健：雲端運算和人工智慧的影響

雲端運算 | 人工智慧 | 醫療保健 | 預測分析 | 診斷工具 | 個性化醫療 | 遠程醫療 | 電子健康記錄 | 精準醫療

2024-05-23



近年來，由於技術的推動，特別是透過雲端運算和人工智慧（AI），醫療保健發生了變革性的變化。這些進步正在重塑病患護理、診斷以及整體醫療操作。

雲端運算在醫療保健中的應用 雲端運算涉及透過互聯網提供計算資源，使醫療提供者能夠遠程訪問、存儲和管理數據。主要優勢包括：

- 數據存儲和管理：為大量醫療數據提供安全且可擴展的解決方案。
- 互操作性：不同醫療系統之間無縫的數據共享與協作。
- 遠程訪問：醫療專業人員可隨時隨地訪問關鍵數據，提升護理質量。

人工智慧在醫療保健中的應用 AI 開發能夠執行需要人類智慧的任務的系統，例如學習和解決問題。醫療保健中的主要應用包括：

- 預測分析：分析數據集以識別模式並預測結果，從而實現主動護理。
- 診斷工具：高準確度的 AI 驅動工具，用於解釋醫學影像和識別疾病。
- 個性化醫療：根據基因和病患數據制定量身定做的治療方案。

雲端運算和人工智慧的結合 這些技術的整合提升了病患護理並簡化了操作流程：

- AI 驅動的遠程醫療：支持視頻諮詢、虛擬健康助手和遠程監控。
- 增強診斷：基於雲端的 AI 工具用於處理醫學影像和病理樣本。
- 簡化操作：優化醫院管理和電子健康記錄（EHR）系統。
- 個性化治療計劃：利用基因數據進行精準醫療和慢性病管理。

隨著雲端運算和 AI 技術的不斷發展，它們有望解決重大醫療挑戰，創造出更加高效、有效且可及的醫療保健未來。

[閱讀更多](#)

Azure AI Studio：協作 AI 開發的新時代

[Azure AI Studio](#) [hubs](#) [AI 開發](#) [協作](#) [安全性](#) [原型設計](#) [客製化開發環境](#) [雲資源](#) [專案工作區](#)

Microsoft

2024-05-28



上週在 Microsoft Build '24 大會上，Microsoft 公佈了 Azure AI Studio 的全面上市，這是一個旨在大幅提升 AI 應用開發的平台。此次發佈的其中一個亮點功能是 "hubs" 的引入，它使開發者能夠獨立創建和管理專案工作區，無需持續的 IT 監督。

什麼是 Hubs？

Hubs 是集中管理的空間，團隊可以在這裡管理多個專案的安全性、連接性和計算資源。每個在 hub 中創建的專案工作區都會繼承 hub 的安全設置和資源訪問權限，促進無縫的開發環境。這種設置允許團隊有效地組織工作、管理數據並控制訪問權限，同時保持強大的安全性和合規標準。

加速 AI 原型設計

AI 開發中的常見挑戰之一是從原型過渡到全面專案時遇到的瓶頸。通常，需要一個專門的平台團隊來配置雲資源，這導致了延遲。Azure AI Studio 解決了這個問題，允許 IT 團隊設置可重複使用的環境 (hubs)，開發者可以用這些環境來創建自己的專案工作區。這種方法顯著減少了設置時間，加速了創新。

客製化開發環境

開發者可以從 Azure AI Studio 創建 hub，或使用 Azure Portal 進行進階配置，如網路和加密。在 hubs 中的預配置連接確保了對公司資源（如模型端點和數據庫）的無縫訪問，消除了通常由 IT 依賴引起的延遲。

原型設計與構建 AI 應用

設置好 hubs 後，開發者可以在受控的遊樂場環境中立即探索和部署來自 Azure AI 模型目錄的模型。這允許徹底的測試和可行性評估。準備就緒後，開發者可以將模型整合到專案工作區中，利用 Azure AI 服務進行無縫整合。

協作開發的專案

Azure AI Studio 中的專案作為容器來組織和管理 AI 應用開發。它們配有客製化工具，用於即時運行、內容安全、微調等。這些工具使開發團隊能夠連接和上傳數據、應用內容過濾器，並在部署前評估應用的質量和安全指標。

簡化協作

對於同時使用 AI Studio 和 Azure Machine Learning 的團隊，專案工作區提供了共享環境，促進開發者和數據科學家之間的協作。這個共享空間有助於構建和運行 GenAI 應用和自訂機器學習模型。

入門指南

總結來說，Azure AI Studio 的 hubs 和專案工作區提供了一個可擴展的解決方案來管理 AI 開發，提升了敏捷性和治理。Microsoft 提供了詳細的文檔和資源，幫助團隊入門。

通過簡化對共享資源的訪問和簡化專案設置，Azure AI Studio 將徹底改變 AI 應用的開發和部署方式。

[閱讀更多](#)