

META全新發布開源LLAMA3
模型，媲美閉源模型

META

GOOGLE CLOUD NEXT
2024 揭露革命性 GEMINI 1.5
PRO 更新

GOOGLE

OPENAI 擴展至東京：領航亞
洲AI發展

OPENAI

人工智慧技術月報 AI TRENDS

Artificial Intelligence Technology Monthly Report

目錄

精選文章

-
- Google Cloud Next 2024 揭露革命性生成式 AI 與 Gemini 更新 9
 - OpenAI 擴展至東京：領航亞洲AI發展 11
 - Meta 發布開創性的 Llama 3 語言模型：向先進 AI 應用的重大進展 13

模型技術

-
- 利用Amazon的PyTorch 2.0 FSDP在EKS上革命性地改進AI模型訓練 16
 - 革新LLM基準測試：Gradient搭配AWS Inferentia的飛躍 17
 - 透過 Amazon SageMaker JumpStart 發掘太陽能模型的力量 18
 - 本周於Microsoft Research：重要創新 19
 - 連接無碼與碼先機器學習：Amazon SageMaker 的無縫整合 20
 - 以 Amazon 進階 AI 模型革新電子商務搜尋 21
 - AWS與Mistral AI：為所有人開創生成式AI的未來 23
 - 轉型虛擬世界：NVIDIA 革命性的 ACE 微服務 24
 - 利用 Cohere 的 Command R+ 在 Azure 上解鎖企業 AI 的未來 25
 - Nielsen Sports透過Amazon SageMaker革新視頻分析，削減成本75% 26

● OpenAI 提升 AI 模型自定義與細調功能	27
● 利用 Amazon Rekognition 的最新特性革新內容審核	29
● Microsoft AI 加強英國影響力，新倫敦中心開幕	30
● 透過 Amazon Bedrock 新的元數據篩選功能增強數據檢索	31
● 以 LlamaIndex 及 Llama 2-Chat 革命化對話式應用	32
● 用自然語言解鎖數據：Mixtral 8x7B 的飛躍進步	33
● 利用 Amazon SageMaker 最新更新讓 AI 模型推論革命化	35
● LLMOps：安全且準確聊天機器人的前沿	36
● 革命性的語言處理：Stability AI 的最新躍進	38
● 利用 Amazon Bedrock 最新更新實現回應準確度的新層次	39
● 加速 AI 發展：NVIDIA 與 Google Cloud 建立新聯盟	40
● 利用 AWS 機器學習革新汽車安全技術	41
● AI 在數據完整性革命中的關鍵角色	43
● 解鎖人工智慧的力量：基礎模型初探	45
● Microsoft Research 開創語言技術新視野，由 Kalika Bali 領銜	47
● 美國與日本推進革命性的 AI 與科技協定	49
● 轉型文件分類：Amazon Titan 多模態嵌入表示法的革命	51
● 在 NVIDIA GTC 2024 上的精彩合作與創新：AWS 與 NVIDIA 領導生成式 AI 革命	53
● OpenAI 在日本推出針對日語的尖端 GPT-4 定製模型	55
● NVIDIA 的願景：AI 作為社會進步的催化劑	56
● Microsoft 在 2024 年 NSDI 的開創性貢獻：塑造未來網絡系統	58
● 以 Spectrumize 革命化全球連接：物聯網 (IoT) 的未來	60
● 介紹 Idefics2：Hugging Face 的視覺語言模型未來	61

● SageMaker Studio 透過 Text-to-SQL 及直接資料庫連線革新 SQL 查詢	63
● Amazon SageMaker 透過新函式庫革新大型語言模型訓練	65
● 透過 NVIDIA 最新 GPU 釋放創造力與效率	67
● 革命性的影片編輯：DaVinci Resolve 19 與 AI 的力量	69
● 本週聚焦Microsoft Research的創新	71
● SAS 透過為所有技能水平打包模型革命化 AI 的可及性	73
● 三星以LPDDR5X DRAM設立新標準，提升AI性能	75
● 透過生成式AI轉型業務：來自Ikigai Labs的Kamal Ahluwalia洞察	77
● 創新分析：透過 Amazon Bedrock 和 Neptune 提升投資策略	79
● 利用 AWS 上的開源可觀察性解鎖機器學習的未來	81
● 解鎖隱形光譜：Living Optics 提升高光譜成像技術	83
● 與NVIDIA的Instant NeRF一起踏進未來：將2D圖像轉變為3D世界	85
● Microsoft Research 推出 SAMMO：透過提示優化轉變 AI	87
● 介紹 Mixtral 8x22B：開源 AI 模型的重大突破	89
● 探索Amazon SageMaker JumpStart中Meta Llama 3模型的力量	91
● NVIDIA 以 Meta Llama 3 加速人工智慧發展	93
● 革新機器學習部署：Amazon SageMaker 遇上 Kubernetes	95
● 利用 Amazon Bedrock 先進 AI 革新您的簡報方式	97
● 利用 AI 加速藥物發現：Iambic Therapeutics 的重大突破	99
● 利用 AWS 自動訓練實現個人化的革命	101

資訊安全

- 英美聯手在AI安全領域取得重大進展 104
- 深偽困境中的航向：保護企業於數位幻象之中 105
- 航向數位假訊息的新時代 106
- AI首爾峰會：AI安全與創新的全球對話 107
- 使用 AWS 的尖端 Nitro 系統確保生成式 AI 工作流的安全 109
- 在人工智慧領域，資料隱私和安全成為主要關切 110

應用

- 深入探索 AWS 與 Hugging Face 在北美的 AI 未來巡迴演講 113
- 簡化機器學習存取：Amazon SageMaker Canvas 與 AWS IAM Identity Center 114
- Google 利用 AI 為開發中國家鋪路 115
- 以創新照亮道路：Dotlumen 在輔助技術上的飛躍 116
- 在數位時代，資料品質對於GenAI在行銷領域的關鍵作用 117
- 以 Amazon Personalize 革新新聞消費方式 118
- Gramener利用Amazon SageMaker對抗都市熱島效應 119
- 🌟 深入機器學習的世界，參加 ML 奧林匹亞競賽！🌟 120
- 利用AI創新：Google Cloud對企業的變革性影響 121

● 探索AI的全球影響：Google AI Podcast系列洞見	123
● 解鎖未來：在 Google Cloud Next '24 上實現的實際 AI 轉型	124
● 用 Google 的生成式 AI 革新您的廣告視覺故事講述	126
● 透過Bing和GenAI革新競爭情報與銷售策略	128
● 用亞馬遜的即時會議助理革新您的會議體驗	129
● Slack 透過 AI 提高生產力：工作效率的巨大飛躍	130
● NVIDIA 在 EMEA 地區慶祝 AI 成就，頒發合作夥伴獎項	132

服務

● 立即體驗 ChatGPT，開啟 AI 世界的探索之旅	135
● 以NVIDIA GeForce NOW遊戲陣容邁入未來	136
● 利用 AWS CloudFormation 簡化 Amazon Lex 聊天機器人的部署	138
● 利用 Amazon Bedrock 的 AI 驅動 IaC 腳本革新雲端遷移	140

01 精選文章

Google Cloud Next 2024 揭露革命性 生成式 AI 與 Gemini 更新

Google Cloud Next | 生成式 AI | Gemini 更新 | AI 超級電腦 | 雲計算 | 網絡安全 | **Google**
Workspace | 行業創新

2024-04-09



在 Google Cloud Next 2024 的激動人心的發展中，Google 已經用重大更新和創新推動了 AI 和雲計算未來，為技術和商業轉型的可能性設定了新的標準。擴展 AI 視野與 Gemini 1.5 Pro 其中一個引人注目的宣布是 Gemini 1.5 Pro 的公開預覽。Google 的AI 模型的這一最新版本展示了在性能上的巨大飛躍，特別是在理解長達 100 萬個 token 的長文本方面。想像一下，企業分析廣泛的數據的潛力，從全面的報告到詳細的客戶反饋，都具有無與倫比的深度和準確性。Gemini 1.5 Pro 的多模態功能意味著它可以處理各種數據類型——音頻、視頻、文字和代碼——將複雜的任務轉化為創新的解決方案。一家遊戲公司可以通過詳細的視頻分析革命化玩家反饋，而保險公司可能通過整合數據分析加快索賠處理。以 Gemini 1.5 Pro 在多模態功能方面的突出應用為例，它可以無縫分析、分類和總結龐大且複雜的文檔，例如上传的Apollo 11號登月任務的402頁PDF記錄，並能夠在不到一分鐘內精確回答特定查詢。此外，Gemini 1.5 Pro 在處理超長視頻方面也表現出色，例如分析一部完整的無聲電影，並迅速準確地提供情節摘要和關鍵事件的時間點。這證明了 Gemini 1.5 Pro 如何利用其多模態能力來提高信息處理的效率和精確性。AI 超級電腦：生成式 AI 的支柱 支持這些

進步，Google 的 AI 超級電腦將 TPUs、GPUs 和 AI 軟件的最佳功能合而為一，提供了一個強大的基礎設施來支持這些 AI 模型。作為這一基礎設施的一部分引入 TPU v5p，凸顯了 Google 提供先進 AI 開發資源的承諾，擁有前代四倍的計算能力。通過 Gemini 強化雲計算和網絡安全 更新擴展了 Gemini 在雲計算和網絡安全方面的應用，引入了AI 輔助功能，簡化編碼、提高生產力並加強安全。憑藉 Gemini Code Assist 和新的威脅情報能力，Google 正在設定高效解決問題和複雜威脅分析的新標準。生成式 AI 打造更智能的工作空間 Google Workspace 也正在獲得 AI 升級，整合生成式 AI 功能，簡化並提高合作和創造力。Gmail、Docs、Sheets 的新工具，以及 Google Vids 的首次亮相，一款全方位視

[閱讀更多](#)

OpenAI 擴展至東京：領航亞洲AI發展

OpenAI 東京 人工智慧 GPT-4 日本 技術合作 商業應用 社會可持續發展

2024-04-15



OpenAI 在連接各大洲的技術與創新方面，採取大膽的一步，策略性地在日本東京開設了其在亞洲的首個辦公室。這一開創性的舉措旨在將 OpenAI 的先進人工智慧技術與日本享譽世界的技術景觀和創新精神整合。

為什麼選擇東京？這個城市站在全球技術和服務文化的最前沿，使其成為 OpenAI 促進合作的完美樞紐。目標很明確：與日本政府、當地企業和研究機構密切合作，量身定制符合日本特定需求的 AI 工具。

OpenAI Japan 的掌舵人是長崎忠夫總裁，他擁有豐富的經驗，引領商業和市場參與努力。OpenAI 給日本市場的首份禮物之一是提前獲得定制的 GPT-4 模型早期存取權。這個版本不僅僅是任何更新；它針對日語進行了優化，提供了無與倫比的翻譯和摘要任務效率——同時成本效益高，且速度比其前身快三倍。

這一擴張的影響已經顯現。像 Daikin、Rakuten 和 TOYOTA Connected 這樣的公司正在利用 ChatGPT Enterprise 提升業務操作，從數據分析到內部報告。即使是地方政府也在軌道上，橫須賀市利用這項技術提高公共服務效率，這一舉措已使 80% 的市政府員工報告了顯著的生產力提升。

但 OpenAI 的願景遠遠超越商業成功。與日本在 AI 政策領導地位保持一致，OpenAI 致力於開發尊重人類尊嚴、多樣性、包容性並促進可持續社會的 AI。這一使命在解決地區性挑戰，如農村人口減少和勞動力短缺等問題上至關重要。

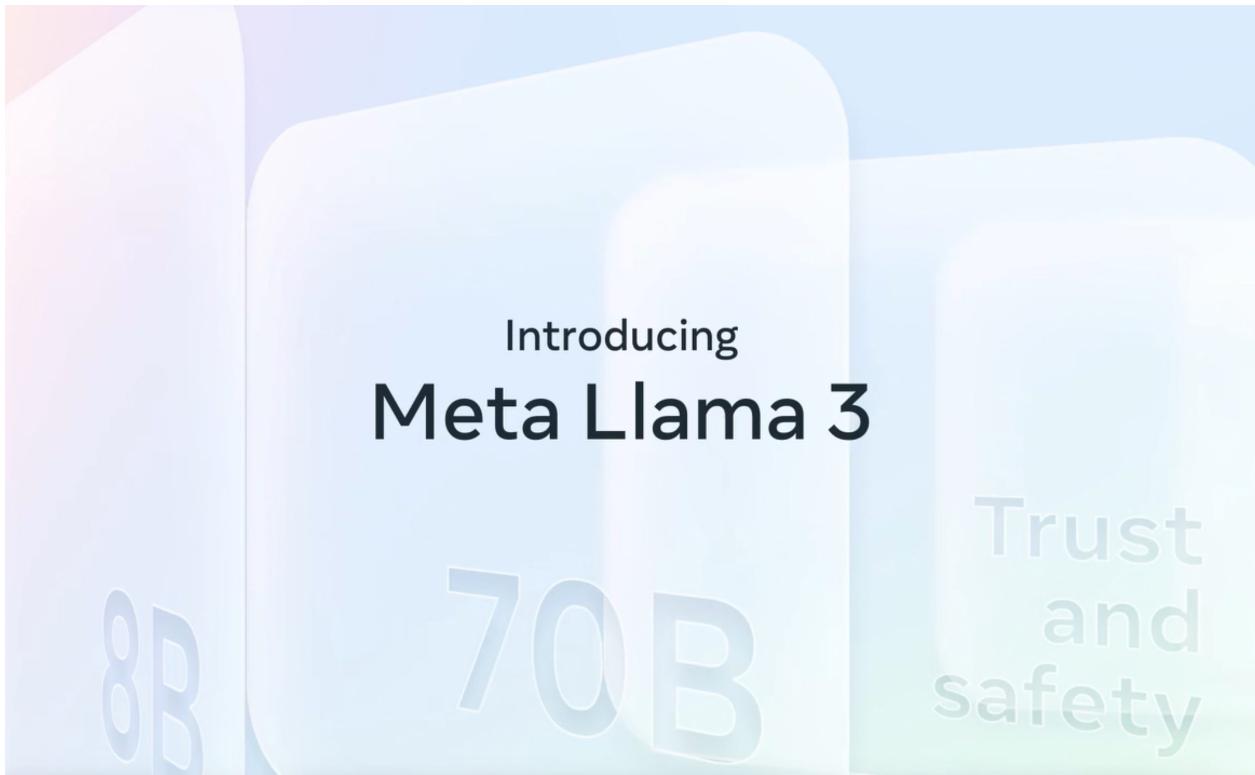
OpenAI 在東京的設立不僅僅是新辦公室的開幕；它是將多元視角納入 AI 發展的大門。這是一步確保人工普遍智能的好處能夠跨越每個行業和社區，為全球所有人帶來益處的方向邁進。

[閱讀更多](#)

Meta 發布開創性的 Llama 3 語言模型：向先進 AI 應用的重大進展

Meta | Llama 3 | 語言模型 | 開源 | AI 安全工具包 | 道德發展 | 語言處理 | 指令細調技術

2024-04-19



Meta 推出開創性的 Llama 3 語言模型：邁向先進 AI 應用的一大步

在 AI 愛好者和開發者們熱切期待中，Meta 推出了最新的開源大型語言模型（LLM）Llama 3。這款新模型不僅僅是一次升級；它旨在超越其前輩，包括眾所周知的 GPT-3.5，尤其是在實際應用中。

Llama 3 包含了從 80 億到 700 億參數的模型範圍，且 Meta 不僅止步於此。計畫揭露更大的模型，超過 4000 億參數，已經在進行中。這一雄心勃勃的擴展標誌著 AI 在理解和處理人類語言方面的能力有了一次飛躍。

Llama 3 的開發歷時兩年，涉及收集高品質訓練數據、精煉模型架構，以及開創性地進行指令細調技術的努力。結果如何？一個 700 億參數的模型在編碼、推理和創意寫作等多項任務中表現優異，通過人類評估擊敗了其他同等規模的模型。

Meta 致力於推動 AI 透明度和道德發展，在其「默認開放」策略下推出 Llama 3 就是一個明證。這個模型將在所有主要雲平台上可用，為科技社區提供了一個獨特的機會去探索、創新，並為負責任的 AI 發展做出貢獻。

與 Llama 3 同時，Meta 也更新了其 AI 安全工具包，引入新功能設計來分類風險、評估潛在濫用情況，和過濾不安全的代碼建議。這些增強功能強調了以道德原則和用戶安全為基礎發展 AI 的重要性。

隨著我們期待 Llama 3 融入 Meta 的服務陣容，包括充滿希望的 Meta AI 助手，轉變生產力、學習和創造力的潛力是巨大的。這項開源倡議不僅凸顯了 Meta 在 AI 領域的領導地位，也為開發道德和強大的 AI 工具設定了新的標準。

[閱讀更多](#)

02 模型技術

利用Amazon的PyTorch 2.0 FSDP在EKS上革命性地改進AI模型訓練

Amazon | PyTorch 2.0 | FSDP | EKS | AI模型訓練 | GPU | Llama2模型 | 全分片數據並行 | 大型語言模型 | NVIDIA A100 | H100 Tensor Core GPU

2024-04-01

利用Amazon的PyTorch 2.0 FSDP在EKS上革命性地改進AI模型訓練

在機器學習領域，Amazon Web Services (AWS) 與Meta的PyTorch團隊合作，引入了一種使用PyTorch 2.0的全分片數據並行 (Fully Sharded Data Parallel , FSDP) 在Amazon彈性Kubernetes服務 (EKS) 上擴展大型語言模型 (LLMs) 的先進方法。這一突破使得在多個GPU上訓練巨大的AI模型，例如引人入勝的Llama2模型，效率達到了前所未有的水平。

這項創新的本質在於其能夠對AI模型進行分片，或分割，到不同的數據並行工作器上，從而顯著減少每個GPU上的內存需求。這種新穎的方法不僅克服了單個GPU的內存容量限制，還為訓練更複雜的AI模型鋪平了道路，這些模型以前由於硬件限制而難以實現。

AWS與Meta的合作展示了FSDP的實際應用，展示了在AWS強大的計算實例上訓練擁有數十億參數的Llama2模型，這些實例配備了NVIDIA A100和H100 Tensor Core GPU。結果呢？吞吐量幾乎線性增長，訓練時間大幅縮短，為企業和開發人員開啟了新的可能性。

這一技術進步將機器學習帶給了更廣泛的觀眾，使創建更準確、更強大的AI工具成為可能，這些工具的應用範圍從虛擬助手到內容創建和計算機視覺。通過簡化訓練大規模AI模型的複雜性，AWS和Meta正在為行業設定新的標準，使企業更容易利用AI的全部潛能來推動創新和效率。

隨著AWS繼續推動機器學習的界限，承諾即將推出的功能，如每個參數的分片和對FP8的支持，以進一步提高在尖端H100 GPU上的性能和效率。這僅僅是由AWS和Meta的PyTorch團隊合作帶來的AI模型訓練新時代的開始。

[閱讀更多](#)

革新LLM基準測試：Gradient搭配AWS Inferentia的飛躍

LLM AWS Inferentia 基準測試 Gradient AI開發 模型評估 成本降低

2024-04-02

革新LLM基準測試：Gradient搭配AWS Inferentia的飛躍

在最近的一次合作中，專注於定製大型語言模型(LLM)開發的先驅Gradient，引入了一種既經濟又簡便的LLM基準測試新方法，這要歸功於AWS Inferentia。這項進步對於那些希望在LLM上線之前，更有效地進行微調和預訓練的企業來說，尤為關鍵。

Gradient的AI開發實驗室一直處於打造私有、定製LLM和AI共駕系統的前沿。在其開發過程中，他們認識到了需要一種更靈活、更經濟的模型評估工具。傳統工具不但昂貴，而且受到硬件資源的限制。引入AWS Inferentia2及其底層庫AWS Neuron後，Gradient巧妙地將其整合到了lm-evaluation-harness中——一個用於對生成式語言模型進行評分的開源框架。

這一整合使得Gradient可以在訓練過程中及之後，將其模型，包括早期的Albatross模型版本v-alpha-tross，與公開基準進行對比。通過利用AWS Inferentia2實例，Gradient現在能夠存取多達384 GB的共享加速器內存，非常適合他們全面的公開架構。此外，採用AWS Spot Instances使成本降低了多達90%，使得更頻繁和多樣化的測試場景成為可能。

AWS Inferentia2提供的靈活性極大，使Gradient能夠生成與Hugging Face上多種模型在Open LLM Leaderboard上的得分相當的成績，同時保持應用私有基準的能力。這種方法不僅標準化了基準測試過程，還為廣泛的模型測試鋪平了道路，而不需要付出通常的高昂代價。

對於那些有興趣探索這種創新基準測試方法的人來說，Gradient使他們的過程變得可訪問且可複製，承諾為技術社區內的LLM開發和評估提供一條更流暢的途徑。

這一突破象徵著在民主化LLM基準測試方面向前邁出了重要一步，使其更易於各種規模的組織使用，並強化了像Gradient和AWS這樣的技術先驅之間合作創新的潛力。

[閱讀更多](#)

透過 Amazon SageMaker JumpStart 發掘太陽能模型的力量

Amazon SageMaker JumpStart | 太陽能模型 | 機器學習 | 大型語言模型 | 對話應用 | **Hugging Face**
Llama 2 | **Mistral 7B** | **Solar Mini Chat**

2024-04-02

透過 Amazon SageMaker JumpStart 探索太陽能模型的力量

對於機器學習愛好者和開發人員來說，一個激動人心的進展是，由 Upstage 開發的開創性太陽能基礎模型（Solar foundation model），現已可透過 Amazon SageMaker JumpStart 存取。這個創新的大型語言模型（LLM）經過精心預訓練，擅長於各種應用，擁有跨多種語言、領域和任務的多功能性。

Solar 的特點是它緊湊而強大的設計，專為特定用途訓練，從而顯著提高性能。值得注意的是，Solar 在多輪對話應用中表現出色，為英語和韓語提供精準而細膩的語言處理能力。它基於 Solar 10.7B 模型，結合了 32 層 Llama 2 結構和預訓練的 Mistral 7B 權重，非常適合處理詳細對話。

這個模型在 Hugging Face 的開放 LLM 排行榜上取得了顯著的成功，提供比其主要競爭對手更快、更準確的回應。通過 SageMaker JumpStart 部署 Solar 模型來進行對話應用，從未如此簡單，承諾在對話環境中帶來新的交互性和效率水平。

在 SageMaker Studio 中開始使用 Solar 模型很簡單——Amazon 的全面管理機器學習中心。它提供了將預建 ML 模型部署到生產中所需的一切，提供從開發到部署的無縫過渡。現已準備好部署的 Solar Mini Chat 及其量化對應物，均針對對話應用進行了優化，承諾能夠把握語言細微之處並提供優越的用戶互動。

這一進展不僅為開發人員簡化了部署過程，還為更多創新和定制的生成式 AI 應用鋪平了道路。深入探索使用 SageMaker JumpStart 上的 Solar 模型增強的機器學習世界，並立即轉變您的應用。

[閱讀更多](#)

本周於Microsoft Research：重要創新

大型語言模型 模擬試錯法 跨語言評估 音頻描述 對比學習

2024-04-03

透過想像力試錯法解鎖大型語言模型(LLMs)的潛力

Microsoft Research正在開創革新方法，以增強大型語言模型(LLMs)的能力。最新研究引入了一種名為模擬試錯法(STE)的創新技術，靈感來自於生物學習過程。STE旨在橋接工具增強型LLMs的缺口，這些模型在工具使用上往往只有30%到60%的正確率。透過整合試錯、想像和記憶，STE顯著提升了LLM透過與工具互動學習的能力，不論是在情境學習還是細微調整的情境中都顯示出有希望的改進。這項進展可能會革新LLMs適應和與新工具互動的方式，使它們更加可靠和高效。

跨語言、跨任務對LLMs的基準評測

在全球數位社群中，理解並利用各種語言的LLMs至關重要。Microsoft的研究團隊介紹了一項全面的基準評測努力，名為MEGAVERSE，旨在評估81種語言（包括數種低代表性語言）中最先進LLMs的性能。這項研究不僅測試了模型在多語言背景下的能力，也探索了它們在多模態任務中的效率。初步發現顯示不同模型間有明顯的表現模式，特別是在低資源語言處理方面的顯著成就。這項工作強調了為未來LLM發展設定廣泛且包容性評估標準的重要性。

用純文字訓練革新音頻描述

Microsoft Research與卡內基梅隆大學(Carnegie Mellon University)合作，正在重新定義自動音頻描述(AAC)，方法是無需音頻即可訓練模型。這一突破性方法利用了一種對比學習模型CLAP，創建音頻和文字之間的共享向量表示，允許僅使用文字描述生成準確的描述。這種方法不僅能與頂級AAC模型競爭，還為描述風格和詞彙多樣性開辟了新途徑。它標誌著向更加無障礙和多樣化的AAC系統邁進的一大步，預示著對各種應用的重大影響。

請繼續關注來自技術研究最前沿的更多更新，Microsoft將繼續突破可能的界限。

[閱讀更多](#)

連接無碼與碼先機器學習：Amazon SageMaker 的無縫整合

Amazon SageMaker 機器學習 無碼 碼先 整合 協作

2024-04-03

跨越無碼和碼先機器學習的鴻溝：Amazon SageMaker 的無縫整合

Amazon Web Services 為資深機器學習（ML）專業人員和新手提供了一種創新方式，以便更有效地協作。通過 Amazon SageMaker Canvas 和 Amazon SageMaker Studio 的無縫整合，用戶現在可以輕鬆地在無碼和碼先機器學習之間過渡，使更廣泛範圍的專業人員能夠貢獻於 ML 項目。

Amazon SageMaker Studio 提供一個全面的、基於網頁的集成開發環境（IDE），專為開發、訓練和部署 ML 模型而設計。它旨在提高 ML 生命週期每個階段的生產力，從數據準備到生產。

另一方面，Amazon SageMaker Canvas 設計給商業分析師和數據團隊，提供一個不妥協準確性的無碼 ML 解決方案。其直觀的視覺界面簡化了一度複雜的數據加載、清理和 ML 模型建構過程，使非專家也能輕鬆接觸 ML。

這一整合之美在於其適應性。用戶可以從 SageMaker Canvas 的直觀無碼界面開始，隨著他們的項目或技能演進，利用 SageMaker Studio 的進階能力進行更自定義和控制的 ML 開發。

Amazon 概述了兩種主要的協作方法：1. SageMaker 模型註冊 選項允許 Canvas 用戶註冊他們的 ML 模型，促進 ML 專家的審查和治理過程。2. 筆記本導出 選項使得項目的筆記本完全共享，以便進行徹底評審，允許 ML 專家在 SageMaker Studio 中進行更深入的自定義和增強。

這種創新的整合不僅促進了 ML 專家和非專家之間的協作，而且還簡化了從數據準備到部署的工作流程，使從無碼到碼先方法的轉換隨著項目需求的演進變得更加容易。

隨著機器學習領域的持續增長，像 SageMaker Canvas 和 Studio 這樣的工具對於使 ML 技術更加易於存取和協作至關重要，加速了從想法到部署的途徑。

對於那些有興趣探索這種整合 ML 開發方式的人，Amazon 提供了從設立 AWS 帳戶到部署您的第一個模型的清晰途徑，確保各種技能水平的用戶都能自信地開始他們的 ML 之旅。

[閱讀更多](#)

以 Amazon 進階 AI 模型革新電子商務搜尋

Amazon | AI | 電子商務 | 搜尋引擎 | 多模態嵌入模型 | 生成式AI | Amazon Titan | Amazon OpenSearch Serverless | Amazon Bedrock | Amazon SageMaker Studio | 語義搜尋 | 產品推薦

2024-04-03

以 Amazon 進階 AI 模型革新電子商務搜尋

在繁忙的電子商務世界中，找到完美的產品有時候就像在大海撈針。但如果我們告訴你，那個大海能夠智慧地引導你找到針呢？多虧了 Amazon Web Services (AWS) 的創新團隊，這現在已經接近現實，他們在搜尋引擎技術上的最新突破讓這一切成為可能。

Amazon 透過一種改變遊戲規則的方式介紹產品推薦，透過一個結合文本和圖像的情境搜尋引擎，利用 Amazon Bedrock 和 Amazon OpenSearch Serverless。這個先進的系統依靠生成式 AI 來革命性地改變消費者在線上找產品的方式，創造一個無縫且直覺的購物體驗。

這項技術的核心是 Amazon Titan 多模態嵌入模型。這個模型是 AI 工程的奇蹟，設計來理解和處理各種形式的數據 - 不論是文本、圖像，甚至是聲音 - 以一種統一的方式。透過分析大規模的圖像與其描述的數據集，該模型學習發現隱藏的連結，將這些多樣的信息嵌入到一個共享的數位空間中。

使這個系統脫穎而出的是其使用餘弦相似度概念進行語義搜尋的能力。這個指標衡量兩個數據片段之間的相似度，例如購物者的查詢和可用的產品，使得搜尋引擎能夠返回真正符合客戶需求的結果。

想像一下，輸入一個查詢或上傳一個想要的商品圖片；Amazon Titan 模型立即行動起來，處理你的輸入以提取相關的產品推薦。這不僅簡化了搜尋過程，還提高了準確度，使在線購物變得更加高效和愉快。

這項技術的實施涉及幾個關鍵組件： - Amazon Titan 多模態嵌入模型：為產品圖像和文本生成嵌入，允許對產品目錄有深入的理解。 - Amazon OpenSearch Serverless：作為一個可擴展的向量數據庫，儲存生成的嵌入以便高效地檢索。 - Amazon SageMaker Studio：提供一個全面的機器學習開發環境，從數據準備到模型部署。

這個解決方案不僅讓電子商務平台在提供更好的搜尋功能方面受益，還開啟了個性化購物體驗的新途徑。它標誌著 AI 理解和與世界互動方式上的一個重大進步，使數位市場更加貼近消費者的需要。

總之，當我們繼續探索網際網路的廣闊資源時，像 Amazon 的情境搜尋引擎這樣的工具為我們照亮了前方，確保我們的搜尋準確無誤地引領我們到達我們想去的地方，以最少的麻煩和最大的相關性。這不僅僅是電子商務的一大步，更是我們與數位環境互動方式未來的一瞥，讓技術更加反應靈敏、理解並最終更加人性化。

[閱讀更多](#)

AWS與Mistral AI：為所有人開創生成式AI的未來

生成式AI AWS Mistral AI Amazon Bedrock 大型語言模型 隱私保護 客戶服務

2024-04-03

AWS與Mistral AI：為所有人開創生成式AI的未來

在生成式人工智慧（AI）領域中，AWS與Mistral AI的聯手，標誌著AI可及性新時代的到來。這次合作預計將革命性地改變各種規模或行業的企業利用AI轉型其運營方式，使得複雜決策更迅速，工作流程比以往任何時候都更直觀。

利用Amazon Bedrock，像是西門子之類的公司已經能夠增強其平台的AI能力，顯著提升生產力和創新力。同樣地，這次合作承諾將生成式AI的轉型潛力擴展到更廣泛的應用範圍，從醫療保健和公共部門運作到Amazon Pharmacy的客戶服務改善，確保隱私和效率是至關重要的。

這次合作中最引人注目的是在Amazon Bedrock上引入的Mistral Large模型。這個模型是語言理解和生成進步的一大光點，特別是在需要深層推理或處理專門內容的任務上。它的專業範圍橫跨多種語言，包括法語、德語、西班牙語和意大利語，以及英語，展示了向全球AI可及性邁進的顯著一躍。

這次合作不僅擴大了在Amazon Bedrock上可用的大型語言模型（LLMs）和基礎模型（FMs）的庫存，也體現了AWS致力於提供多樣化AI工具的承諾。通過提供適用於廣泛用途的優化模型，AWS使客戶能夠找到滿足其獨特需求的完美解決方案，同時確保頂級的安全性和隱私保護。

隨著AWS與Mistral AI持續前進，他們的聯合努力預計將使複雜的AI技術更加普及，使全球的組織能夠創新、提升客戶體驗並像從未有過地優化運營。

[閱讀更多](#)

轉型虛擬世界：NVIDIA 革命性的 ACE 微服務

NVIDIA | ACE 微服務 | 虛擬世界 | NPC | AI 技術 | Riva | Audio2Face | 生成式 AI | 數字人護士 | 客服頭像

2024-04-03

轉型虛擬世界：NVIDIA 革命性的 ACE 微服務

在視頻遊戲變得日益沉浸式的時代，NVIDIA 正在用其尖端的 ACE 微服務推動真實感的極限，重新定義非玩家角色（NPC）在數字宇宙中的互動方式。這些先進的工具賦予開發者能力，讓 NPC 擁有栩栩如生的個性和對話，使遊戲中的每次遭遇都更加引人入勝和動態。

歷史上，NPC 的角色有限，經常重複相同的台詞和動作，這可能會破壞廣闊遊戲世界的沉浸式體驗。然而，NVIDIA 的 ACE 微服務將改變這一點，通過啟用 NPC 能夠與玩家進行實時、與情境相關的對話，這得益於尖端生成式 AI 模型的整合。

這種開創性的方式利用了一系列 AI 技術，從 NVIDIA Riva 開始，將玩家的口述詞轉換為文本。然後，這些文本由強大的語言模型處理以生成自然反應，這些反應被轉換回語音並與 NVIDIA Audio2Face 技術提供的生動、逼真的面部動畫同步。結果呢？NPC 能夠聽見、理解、回應和即時表達情緒，創造出真正互動的遊戲體驗。

ACE 微服務的潛在應用不僅限於遊戲。從改善醫療互動的數字人護士到提供更個人化體驗的客服頭像，這項技術將通過賦予數字人以生命來革命化各個行業。

當我們站在這個數字互動新時代的邊緣時，NVIDIA 的 ACE 微服務不僅增強了遊戲角色的深度和真實感，還為虛擬和現實世界應用中的 AI 驅動對話和互動開啟了無限可能。現在，開發者和創作者擁有了構建更複雜和更引人入勝的數字生命的工具，現實與虛擬現實之間的界限繼續模糊，展示了數字通信和故事講述未來的一瞥。

[閱讀更多](#)

利用 Cohere 的 Command R+ 在 Azure 上解鎖企業 AI 的未來

Cohere **Command R+** **Azure** **企業 AI** **生成型 AI 模型** **知識助理** **客戶支持聊天機器人**

2024-04-04

利用 Cohere 的 Command R+ 在 Azure 上解鎖企業 AI 的未來

在人工智慧領域中的激動人心的新聞：Cohere 與 Microsoft 合作，將一款先進的生成型 AI 模型 Command R+ 帶到 Azure 的 AI 模型目錄中。這一舉措將革新企業使用 AI 的方式，結合了 Cohere 的尖端技術與 Azure 的堅固基礎設施。

Command R+ 不僅僅是一款 AI 模型；它是為企業設計的強大動力，經過精調，專為檢索增強生成（RAG）以確保效率與準確性相結合。這意味著企業現在可以利用其龐大的內部資料和文件庫，創建既高度定制又極為準確的 AI 應用程序。該模型在為其輸出提供清晰引用的能力方面表現優異，減少了不準確性的風險，為它處理的信息帶來更多上下文。

但它不僅止於此。Command R+ 字面上講的是商業語言——實際上，流利地使用 10 種關鍵商業語言，包括主要的亞洲語言。這一能力是構建像知識助理和客戶支持聊天機器人這樣的 AI 應用程序的遊戲規則改變者，可以為全球觀眾服務。

與 Azure AI Studio 工具（如 Azure AI Content Safety 和 Azure AI Search）的整合進一步增強了該模型的產品供應，確保企業能夠開發出不僅強大而且負責任且安全的 AI 解決方案。具有像大型語言模型（LLM）流程的簡化評估和部署簡化等功能，Command R+ 在 Azure AI Studio 上代表著使企業級 AI 變得既可取又實際的一大飛躍。

總而言之，Cohere 與 Microsoft 通過在 Azure AI Studio 上的 Command R+ 的合作，不僅僅是一項技術進步；它是創建能夠在大規模理解、溝通和創新的智能系統的途徑。這證明了 AI 轉變業務運營，使其更加高效、包容且為明天的挑戰做好準備的潛力。

[閱讀更多](#)

Nielsen Sports透過Amazon SageMaker革新視頻分析，削減成本75%

Amazon SageMaker | 影片分析 | 機器學習 | 成本削減 | **Nielsen Sports**

2024-04-04

Nielsen Sports透過Amazon SageMaker革新影片分析，削減成本75%

在影片分析技術方面取得顯著進展，Nielsen Sports通過採用Amazon SageMaker多模型端點，大幅減少了其操作和財務開支達75%。這個全球領先的觀眾洞察和分析公司，在追蹤和評估各種媒體渠道中品牌曝光度方面重新定義了效率。

Nielsen Sports的創新方法包括部署成千上萬的機器學習模型來識別影片中的品牌存在，這一任務對於衡量體育贊助廣告活動的影響至關重要。傳統上，管理如此巨大數量的模型不僅成本高昂，而且速度緩慢且容易出錯。Nielsen Sports戰略性轉向Amazon SageMaker的突破性之舉，使其能夠擁有更加流暢、靈活和成本效益的系統。

這一系統的全面改革意味著，過去涉及大量模型以處理成千上萬小時內容的繁重影片分析任務，現在顯著更加高效。借助SageMaker的多模型端點，Nielsen Sports可以在單一端點上同時運行多個模型，優化資源使用並大幅減少對龐大基礎設施的需求。

結果說明了一切：該公司不僅將操作成本削減了驚人的75%，而且整體管線運行時間也減少了33%。原本需要超過一個月才能部署的模型，現在可以在不到一周的時間內啟動，部署速度提升了75%。此外，使用更新的AWS實例立即帶來了性能提升，進一步說明了SageMaker對Nielsen Sports操作能力的深遠影響。

這一轉型充分展示了創新雲計算解決方案在提高生產力、降低成本和簡化影片分析領域複雜操作中的力量。

[閱讀更多](#)

OpenAI 提升 AI 模型自定義與細調功能

OpenAI | AI 模型 | 細調 API | 自定義模型 | 模型訓練 | GPT-3.5 | Weights and Biases | Indeed | SK Telecom | Harvey | 法律工具

2024-04-04

OpenAI 增強 AI 模型自定義和細調能力

為了賦予開發者和組織更大的能力，OpenAI 推出了對其細調 API 的新更新，並擴大了自定義模型計劃。這些增強功能旨在提供更多的靈活性和對 AI 模型訓練的控制，使得將 AI 工具定制於特定需求和任務變得更容易。

細調 API 獲得加強

為 GPT-3.5 引入的細調 API 現在包括諸如基於 epoch 的檢查點創建的功能，這減少了重新訓練的需求並緩解了過度擬合的擔憂。還添加了一個比較遊樂場，允許進行並排模型評估。此外，OpenAI 已經整合了對第三方平台的支持，從 Weights and Biases 開始，以簡化細調數據的共享。這些更新旨在提高模型準確性，同時降低成本和延遲。

在自定義中的成功故事

這些增強的實際應用已經顯示出希望。例如，Indeed 利用細調 API 精煉了其工作推薦系統，導致信息個性和效率的顯著增加。同樣，SK Telecom 與 OpenAI 合作開發了一款自定調模型，大大增強了他們在電信行業的客戶服務能力。

擴展自定義模型計劃

OpenAI 現在在其自定義模型計劃中提供協助細調，這是一項旨在實現獨特組織需求的最佳模型性能的合作努力。這種方法對於需要支持設置高效數據管道和自定義訓練方法的實體有益。

此外，該計劃支持開發完全自定義訓練的模型，Harvey 與 OpenAI 合作開發的法律工具就是一個例證，該工具由在大量法律信息數據集上訓練的 AI 模型驅動。

展望未來

OpenAI 展望一個個性化 AI 模型成為各行各業常態的未來，提供量身定制的解決方案，產生重大影響。這些對細調 API 的最新更新和自定義模型計劃的擴展標誌著邁向實現這一未來的一步，為組織提供利用定制 AI 的途徑。

如需開始您的細調之旅，或探索您的組織的自定義模型解決方案，請訪問 OpenAI 的文檔並尋求支持。

[閱讀更多](#)

利用 Amazon Rekognition 的最新特性 革新內容審核

Amazon Rekognition | 內容審查 | 深度學習 | 影像分析 | 視頻分析 | 自訂審查 | 批量分析

2024-04-05

Amazon Rekognition 透過其尖端的影像與視頻分析能力，正在改變各行各業管理用戶生成內容的方式。這項革命性技術基於深度學習，讓無機器學習專長的用戶能夠輕鬆地將先進的內容審查整合到他們的應用程式中。

最新更新，內容審查版本 7.0，帶來了一系列旨在精煉審查過程的增強功能。透過 26 個新的審查標籤和擴展的分類體系，用戶現在可以更準確地檢測細微概念。這個版本還引入了識別動畫和插圖內容的能力，使內容過濾更加精確。

Amazon Rekognition 透過批量分析進一步簡化了審查工作流程。此功能允許大量影像集合進行異步分析，有效識別不當內容。通過消除對定製審查解決方案的需求，批量分析簡化了流程，提供了大規模影像審查類別的洞見。

自訂審查是另一項創新功能，提高了審查預測的準確性。用戶可以用自己的影像訓練一個自訂審查適配器，細化基礎模型以滿足特定的審查需求。這種適應性確保內容根據個別內容政策進行審查，促進在廣告、遊戲、媒體和電子商務等各種平台上的用戶社群更安全。

Amazon Rekognition 在批量分析和自訂審查方面的進步，為內容審查設定了新的標準。通過利用這些功能，用戶可以保護他們的品牌聲譽並為他們的社群創造安全的環境，同時輕鬆、精確地應對審查大量用戶生成內容的挑戰。

[閱讀更多](#)

Microsoft AI加強英國影響力，新倫敦中心開幕

Microsoft AI | 倫敦 | 人工智能 | 語言模型 | GPU | 負責任AI

2024-04-08

在一個大膽的舉措中，以利用英國豐富的人才庫，Microsoft AI宣布在倫敦開設一個全新的據點。這項戰略行動旨在推動人工智能的界限，專注於發展先進的語言模型以及支持這些基礎模型的基礎架構。

這項開創性企業的帶領者是AI科學家及工程師Jordan Hoffmann，他因在AI產業的開創性工作而聞名。Microsoft AI擴展到倫敦不僅僅是關於創新；這是對英國AI專家和工程師的長期投資，表明公司致力於在該地區培養人才的決心。

Mustafa Suleyman，Microsoft AI的執行副總裁兼CEO，也是一位土生土長的倫敦人，他表達了對於為英國的AI景觀做出貢獻的自豪感。倫敦據點是Microsoft更廣泛願景的一部分，其中包括投資25億英鎊以提升英國勞動力技能和增強AI基礎設施，承諾到2026年將引進20,000個先進的GPU。

這次擴展不僅僅代表著技術上的進步；這是對負責任AI發展的承諾，強調安全和倫理考量。倫敦據點不僅代表著AI研究的一大步，也強調了英國在負責任人工智能演進中作為一個領導者的角色。

您是否被塑造AI未來的挑戰所激勵？Microsoft AI倫敦正在尋找準備好面對AI最大問題的卓越人才。這可能是您站在AI創新前沿的機會。

敬請期待更多技術和人工智能世界中的洞察和機會。

[閱讀更多](#)

透過 Amazon Bedrock 新的元數據篩選功能增強數據檢索

Amazon Bedrock | 元數據篩選 | 數據檢索 | **Retrieval Augmented Generation** | **AWS**

2024-04-08

透過 Amazon Bedrock 新的元數據篩選增強數據檢索

在數據世界中，快速且準確地找到正確的信息就像在大海撈針——尤其是當「海」非常龐大時。Amazon Web Services (AWS) 在 Amazon Bedrock 的知識庫中引入了一項突破性功能來正面迎戰這一挑戰。這個新功能，元數據篩選，旨在顯著改善檢索信息的過程，使之更加精確且針對特定需求。

Amazon Bedrock 的知識庫現在允許用戶透過一個完全管理的 Retrieval Augmented Generation (RAG) 模型，更安全地將他們的數據與基礎模型連接。此次更新的核心在於能夠根據文件的元數據篩選搜索結果。當需要從特定時期檢索文件或將文件標記在某些類別下時，這特別有用，增強了基礎模型產生的回應的準確性。

想象一下，必須篩選具有相似術語但不同上下文的法律文件，或者情節相似但發行年份不同的電影。有了元數據篩選，您可以指定創作年份或特定標籤等條件，將搜索結果縮小到最相關的文件。這不僅節省時間，還確保生成的回應更符合給定的上下文。

這個過程涉及為每個文件提供一個自定義的元數據文件，其中可以包括項目標籤、年份和團隊名稱等屬性。通過指示向量存儲在進行搜索之前基於這些元數據進行預篩選，用戶可以控制檢索到的文件，特別是在處理含糊查詢時。這項功能不僅僅提高準確性；它還通過減少查詢向量存儲時所需的 CPU 周期和成本來提高性能。

元數據篩選為許多應用打開了大門，從為軟件公司創建更高效的文件聊天機器人到促進組織資產內的會話式搜索。它甚至幫助軟件開發者確定與特定發布版本或文件類型相關的具體信息。

對 Amazon Bedrock 的知識庫進行的這次更新代表了在使數據檢索不僅更快，而且更智能方面的重大進步。通過使用戶能夠使用元數據過濾器來精緻化他們的搜索，AWS 為跨越廣泛行業和應用的更有效和準確的信息檢索鋪平了道路。

[閱讀更多](#)

以LlamaIndex及Llama 2-Chat革命化對話式應用

LlamaIndex | Llama 2-70B-Chat | LangChain | 對話式應用程式 | 大型語言模型 | 檢索增強生成 | 聊天機器人 | 企業搜索助手

2024-04-08

以LlamaIndex和Llama 2-Chat革新對話式應用程式

在提升對話式應用方面邁出重大的一步，AWS推出了LlamaIndex和Llama 2-70B-Chat，並與LangChain框架一同標誌著建立強大問答系統的新時代。這些創新技術旨在消化和分析大量文本語料庫，使得創建的應用程式不僅能理解複雜的查詢，還能以精確和清晰的方式回應。

Llama 2-70B-Chat：一個先進的大型語言模型（LLM），預先訓練了兩兆的文本token。它被設計來提供可與市場上領先模型匹敵的聊天輔助，使其成為開發需要細緻對話能力應用程式的首選方案。

LlamaIndex：一個為LLM應用量身定做的複雜數據框架，提供了一個廣泛的工具套件，用於從多樣化來源和格式吸收數據。其無縫整合能力意味著，無論您的數據存放在資料庫還是文件中—LlamaIndex都能輕鬆將其納入您的應用程式，提高您的LLM效率和響應速度。

這些技術共同為創建基於檢索增強生成（RAG）的應用程式鋪平了道路。RAG是一種將信息檢索與自然語言生成結合的方法，結果不僅洞察力深刻，而且在事實準確性上也有深厚的根基。這種方法特別適合於旨在提供類人對話的應用程式，如聊天機器人和企業搜索助手，將用戶體驗提升到前所未有的水平。

通過利用LlamaIndex、Llama 2-70B-Chat和LangChain的結合動力，開發者現在可以建立位於技術進步前沿的對話式應用程式。這些應用程式能夠遍尋廣泛的知識庫，以提供既與上下文相關又顯著類人的回應。無論是提供客戶支持、回答技術查詢，還是協助研究，其潛在應用範圍廣泛且變革性。

本質上，AWS的最新產品不僅是在機器學習和人工智能方面向前邁進的一步；它是向創建更直觀、智能和可訪問的對話式應用程式邁進的一大步，這些應用程式將重新定義我們與科技的互動方式。

[閱讀更多](#)

用自然語言解鎖數據：Mixtral 8x7B的飛躍進步

自然語言處理 | 數據分析 | **Amazon SageMaker JumpStart** | **Sparse Mixture of Experts** | 商業智能

2024-04-08

用自然語言解鎖數據：Mixtral 8x7B的飛躍進步

在一個數據為王的世界裡，獲取和理解數據不應該感覺像學習一門新語言。引進Mixtral 8x7B，一項人工智慧（AI）的尖端發展，正改變著從資深分析師到可能不知道他們的SQL與他們的HTML的業務使用者的遊戲規則。

想像一下，如果想知道上週特定產品的銷售數字，但不是與複雜的數據庫查詢抗爭，你只需像問一位同事一樣問你的電腦。這不是遠景的一瞥，而是由於今天在Amazon SageMaker JumpStart上的Mixtral 8x7B成為可能的現實。

由Mistral AI開發，Mixtral 8x7B是一個理解你的自然語言請求的模型，潛入數據庫中抓取答案，無需你編寫一行代碼。就好像你已經給你的電腦上了一門人類對話速成課程，使它能夠抓住你正在尋找的內容並迅速交付。

它是如何工作的？

在其核心，Mixtral 8x7B使用所謂的Sparse Mixture of Experts (MoE)架構。不深陷技術迷宮，想像它擁有一個專家團隊（即「專家」）在模型內，每個專家都準備好處理不同類型的數據請求。這種方法使Mixtral 8x7B能夠高效地處理信息，提供快速而準確的回應。

對於業務使用者來說，這意味著不再需要等待IT部門運行複雜報告，或掙扎於學習數據庫語言的學習曲線。無論你是Michelle，一位尋求簡化她的每週銷售報告準備工作的分析師，或是任何有關你數據問題的團隊成員，Mixtral 8x7B都能為你提供幫助。

部署變得容易

Amazon SageMaker JumpStart是Mixtral 8x7B的發射台，確保將這強大的AI整合到你的工作流程中像點擊一個按鈕一樣簡單。這項服務簡化了部署AI模型的原本令人望而卻步的任務，讓你可以訪問Mixtral 8x7B以及其他大量預構建解決方案的廣泛庫。

更大的圖景

Mixtral 8x7B的引入不僅僅標誌著一項技術進步；它代表了我們與數據互動方式的轉變。通過打破複雜數據查詢與終端用戶之間的障礙，Mixtral 8x7B民主化了數據訪問。對於商業智能、運營效率和決策制定的影響是深遠的。

在一個靈活性和知情決策至關重要的世界裡，像Mixtral 8x7B這樣的工具不僅僅是便利，而是必需品。隨著我們在AI旅程中繼續前進，自然語言處理和數據分析的融合，如Mixtral 8x7B所提供的，標誌著邁向一個我們與技術互動更直觀、高效和賦權的未來的關鍵一步。

[閱讀更多](#)

利用Amazon SageMaker最新更新讓AI模型推論革命化

Amazon SageMaker | AI模型推論 | 深度學習容器 | 生成式AI模型 | 大型語言模型 | 推論優化

硬件選擇 | 張量並行性 | 高效批處理 | 量化 | 令牌流傳輸

2024-04-08

用Amazon SageMaker最新更新革新AI模型推論

在提升人工智慧（AI）應用的最近一次努力中，Amazon SageMaker已經推出了其大型模型推論（LMI）深度學習容器（DLCs），版本0.26.0的更新。這次更新對於旨在以效率、速度和可擴展性部署生成式AI模型的開發者和企業來說，是一次遊戲規則的改變。

Amazon SageMaker的LMI DLCs現在支援一系列新模型，包括創新的Mixtral和Llama 2模型，提供了一個低代碼接口，簡化了先進推論優化技術和硬件選擇的應用。這意味著，僅憑模型ID和可選參數，用戶就可以利用張量並行性、高效批處理、量化、令牌流傳輸等等，為首選硬件優化大型語言模型（LLMs），實現無與倫比的價格性能指標。

主要增強功能包括對專家混合模型的支援、增強的推論後端和新一代詳細資料，以獲得對預測過程的更深入洞察和控制。值得注意的是，該更新引入了如旋轉位置嵌入（RoPE）縮放的能力，以在不重新訓練的情況下虛擬擴展上下文窗口，以及如生成完成原因和令牌級別的對數概率等細緻的生成詳細資料，為用戶提供了模型輸出的增加可預測性和解釋性。

在幕後，LMI DLCs由一套適用於不同需求和硬件配置的精密後端支持。從LMI-Distributed Library為延遲和精確度優化，到用於AWS Inferentia2和Trainium基礎實例部署的LMI NeuronX，每個後端都提供了獨特的功能以提高性能和效率。

對於那些部署AI模型的人來說，新功能以及對像Mistral-7B、基於MoE的Mixtral和Llama2-70B等流行模型在各種後端的支援，承諾了一個無縫且高效的推論過程。無論是通過RoPE縮放擴展上下文窗口還是通過CUDA圖形和連續批處理實現性能提升，SageMaker上的LMI DLCs使開發者能夠卸下基礎設施管理的重擔，同時專注於提供一流的AI服務。

這次更新不僅象徵著Amazon對推進AI技術的承諾，也賦予用戶充分發揮其模型潛力的能力，確保AI應用更快的價值實現時間和更流暢、更經濟高效的部署過程。

[閱讀更多](#)

LLMOps：安全且準確聊天機器人的前沿

LLMOps 聊天機器人 安全性 準確性 生成式AI Microsoft Azure AI Content Safety

2024-04-09

LLMOps：安全且準確的聊天機器人前沿技術

在迅速演進的生成式AI領域中，常被比喻為「狂野西部」，Microsoft在確保語言模型（LLM）聊天機器人在投入生產前的安全性和準確性方面，採取了開創性的步驟。這項被稱為LLMOps的倡議，為開發人員在部署既創新又可靠的生成式AI模型時導航的明燈。

為什麼關注LLMOps？

實施LLM聊天機器人的道路充滿了陷阱，從誤導信息到意外的版權侵犯。LLMOps作為緩解這些風險的關鍵框架出現，確保聊天機器人不僅功能性強，而且對公眾互動安全。

LLMOps的三大支柱

1. 評估：這個關鍵的第一步涉及對聊天機器人進行嚴格測試，以確保其在向用戶介紹前滿足高標準的準確性和安全性。這一階段為之後的順暢運營奠定了基礎。
2. 監控：一旦部署，即實時監控聊天機器人，迅速識別並糾正任何新出現的問題，確保持續符合安全標準。
3. 反饋：以用戶反饋閉環，這是幫助完善並增強聊天機器人的性能和用戶體驗的關鍵組成部分。

安全部署之路

Microsoft的LLMOps框架利用了從自動LLM評估到手動評分的一系列評估方法，確保對聊天機器人的能力有一個全面的了解。通過將LLMOps納入持續集成/持續部署（CI/CD）流程，開發人員可以快速迭代並改進其模型，維持高標準的質量和安全性。

確保安全互動

為了保障聊天機器人及其用戶的安全，Microsoft使用先進工具如Azure AI Content Safety，該工具篩選有害語言並檢測潛在的安全威脅。對模型輸出和用戶輸入進行雙重審查，對維護一個安全的互動環境至關重要。

反饋：改進的基石

通過積極徵求和分析用戶反饋，開發人員獲得有關聊天機器人性能的寶貴見解，指導進一步的改進，確保聊天機器人繼續符合用戶需求和期望。

隨著LLM聊天機器人的使用變得更加廣泛，對LLMOps的強調無疑將增長。Microsoft的前瞻性方法不僅解決了部署生成式AI模型的當前挑戰，而且為負責任開發AI技術設定了新標準。

[閱讀更多](#)

革命性的語言處理：Stability AI的最新躍進

Stability AI | 語言模型 | Stable LM 2 | AI工具 | 多語言處理

2024-04-09

革命性的語言處理：Stability AI的最新進展

在最近的一次突破中，Stability AI推出了其語言模型系列的激動人心的更新，引入了擁有高達120億參數的Stable LM 2，以及一個經過改良的16億參數變體。這項創新在AI世界中是一場遊戲改變者，特別是對於希望利用先進語言處理能力而不需承擔沉重計算成本的開發者和企業來說。

Stable LM 2系列現在擁有能夠理解和處理包括英語、西班牙語和法語在內的七種語言的版本，代表著向創造更包容、更高效的AI工具邁出的重要一步。120億參數模型實現了令人印象深刻的平衡，提供了高性能和效率，同時比您期望的該等級別的記憶體和速度要求低得多。

但這對日常用戶或企業意味著什麼？想像一下擁有一個能夠理解和互動多種語言的虛擬助理，不僅僅是基本指令，還能夠進行更深入、更有意義的對話。或者想象一個能夠迅速篩選各種語言大量文本的AI工具，總結信息、回答問題，甚至以接近人類理解的流暢度和準確性生成內容。

Stability AI並不僅僅是推出了12B模型來突破界限。更新的1.6B版本經過了微調，以提高各方面的對話能力，確保互動比以往任何時候都更加流暢自然。

Stable LM 2模型的特點在於其開放和透明的設計，允許開發者在保持對其數據控制的同時自由創新。這種方法在AI領域中是一股清新之風，其中可訪問性和透明度往往是關鍵憂慮。

隨著這些最新成員的加入，Stability AI正在為更可訪問、更高效、更強大的AI工具鋪平道路，這些工具勢必將轉變從客戶服務和內容創建到多語言交流等廣泛應用的範疇。AI語言處理的未來比以往任何時候看起來都更加光明，對於準備探索這些新可能性的開發者和企業來說，這是一個令人興奮的時刻。

[閱讀更多](#)

利用 Amazon Bedrock 最新更新實現回應準確度的新層次

Amazon Bedrock | **RetrieveAndGenerate API** | 基礎模型 | 自定義提示模板 | **AWS**

2024-04-09

使用 Amazon Bedrock 最新更新解鎖新的回應準確度層級

Amazon Bedrock 為其知識庫推出了一些激動人心的更新，特別是針對其 RetrieveAndGenerate API。這些更新旨在賦予用戶更多控制權和自定義能力，確保基礎模型 (FMs) 生成的回應盡可能相關且準確。

首先，用戶現在可以設定可檢索並傳遞給 FMs 的搜尋結果的最大數量。這意味著您可以微調模型在生成回應時考慮的資訊量——對於複雜查詢，資訊量多一些，或者當問題簡單時，資訊量少一些。這種靈活性可以顯著提高上下文的相關性，使答案更精確，減少錯誤。

另一個創新功能是能夠自定義發送給模型的提示模板。這對於根據特定需求調整模型輸出來說，是一個遊戲規則的改變者。無論是調整語氣、格式，還是甚至根據特定行業術語調整內容，這種自定義程度意味著模型的回應可以被精細調整，以滿足各種需求。

這些更新不僅僅是技術提升；它們代表了我們如何與生成式 AI 互動和利用它的一個飛躍。通過允許更詳細的指令和更多背景資訊，Amazon Bedrock 為更精確、有幫助且上下文準確的 AI 生成回應鋪平了道路。無論您是在醫療保健、法律或任何其他領域，這些功能都提供了創建更好地理解回應您特定需求的 AI 應用的工具。

對於任何希望實現這些功能的人來說，Amazon 通過 AWS 管理控制台和 SDK 提供指導，確保無論您是偏好圖形界面還是編碼，都能得到您需要的支持，充分利用這些更新。

Amazon 致力於通過這些更新改善用戶體驗和回應準確度，這明確地表明了 AI 和機器學習技術能力的發展。通過提供更多控制影響 AI 回應的資訊，Amazon Bedrock 正在各個行業中邁向更智能、更可適應和更實用的 AI 工具的重要一步。

[閱讀更多](#)

加速AI發展：NVIDIA與Google Cloud 建立新聯盟

NVIDIA | Google Cloud | AI發展 | 生成性AI | Inception計畫 | Google for Startups Cloud Program

Gemma系列模型 | NVIDIA NeMo框架 | H100 Tensor Core GPU | NVIDIA Blackwell基礎GPU | A3

Mega實例

2024-04-09

在一項旨在加速生成性AI成長的創新舉措中，NVIDIA與Google Cloud宣布了一項戰略合作。這項夥伴關係旨在為全球的新創公司提供重大推動，賦予它們以前所未有的速度和效率開發和部署AI驅動的應用程式與服務的能力。

這次合作的核心在於NVIDIA的Inception計畫與Google for Startups Cloud Program的結合。這種協同作用承諾給予新創公司高達350,000美元的Google Cloud信用額度，同時提供全面的市場進入策略支持和深入的技術專長。目標很明確：幫助這些新創公司比以往任何時候都更快地向他們的客戶交付創新解決方案。

此外，這次合作設定了緩解新創公司在AI領域面臨的財務和技術障礙。通過利用NVIDIA的尖端AI平台，包括Gemma系列模型和NVIDIA NeMo框架，新創公司現在可以訪問最先進的AI工具，這些工具簡化了從概念化到部署的開發過程。

合作夥伴關係還預示著Google Cloud的A3 Mega實例的即將推出，這些實例由NVIDIA H100 Tensor Core GPU提供動力，並引進了革命性的NVIDIA Blackwell基礎GPU。這些技術進步承諾為AI性能設定新的標準，為新創公司提供應對最苛刻的AI、數據分析和高性能計算工作負載所需的計算馬力。

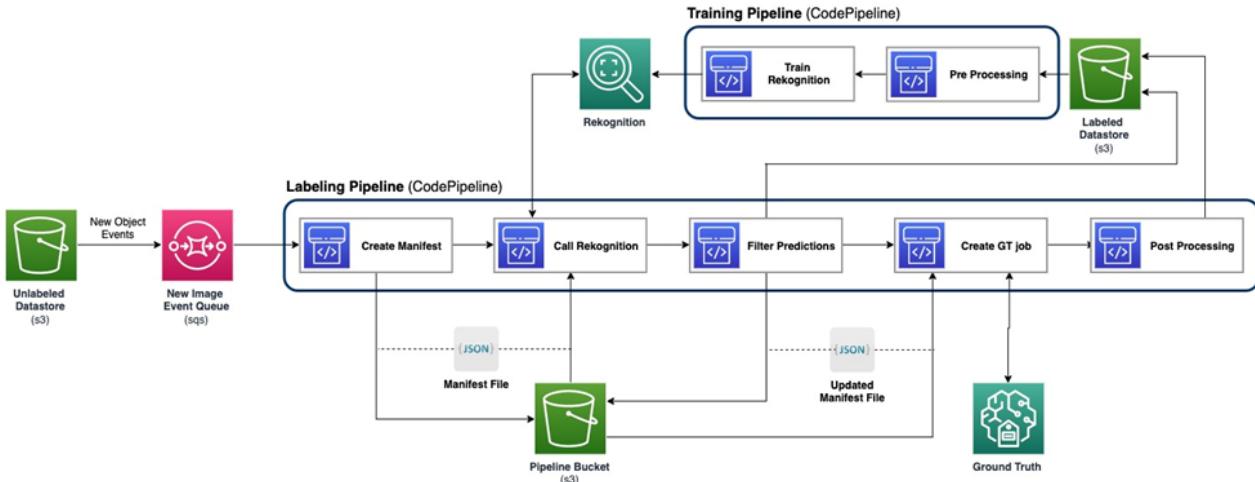
從本質上講，NVIDIA與Google Cloud之間的這次合作不僅僅是提供資源；它是關於創建一個創新蓬勃的生態系統。通過消除AI發展道路上的障礙，這個聯盟為新創公司探索生成性AI及其以外的新領域鋪平了道路，標誌著在利用人工智能的全部潛力的追求中向前邁出的重大一步。

[閱讀更多](#)

利用 AWS 機器學習革新汽車安全技術

AWS | 機器學習 | 汽車安全 | 圖像標註 | Amazon Rekognition | Amazon SageMaker | AWS CodePipeline | AWS Lambda

2024-04-10



利用 AWS 機器學習革新汽車安全技術

在快速發展的汽車技術世界中，透過創新來增強安全性並改善駕駛體驗至關重要。利用 AWS 機器學習服務，已經取得了一項突破性的進展，這項創新推出了一個主動學習管道，專門設計用於自動標註車艙內的圖像，標誌著邁向更智能、更安全車輛的重大飛躍。

這項創新的核心在於主動學習模型，它智能地在機器學習模型和人類專家之間進行迭代，以標註圖像。這不僅大幅降低成本超過 90%，還將標註過程從幾週縮短到僅僅幾小時。其影響深遠，為旨在提升車艙安全功能的汽車公司提供了一個快速、成本效益高、且可擴展的解決方案。

這個解決方案的一個突出組成部分是其使用 Amazon Rekognition Custom Labels 和 Amazon SageMaker Ground Truth。這些工具協同工作以處理圖像，使系統能夠在一小部分手動標註的數據上自行訓練，然後將這些學習應用於更大的數據集。結果是一種更準確、高效且更快速的提升汽車安全系統的方法。

此外，該管道展示了驚人的靈活性和可擴展性，得益於其在 AWS CodePipeline 上的部署以及使用 AWS Lambda 進行數據預處理和後處理。這確保了汽車公司可以適應並擴大其業務以滿足不斷發展的安全標準和要求。

這項技術的影響不僅僅是成本和時間的節省。通過實現更快、更準確的車艙圖像標註，汽車製造商可以顯著提升安全系統的開發，使道路上的每個人都更加安全。這項創新展示了機器學習和雲服務在推動汽車安全未來發展中的力量。

與 AWS 機器學習一起，開啟通往更安全未來的旅程，今天就解鎖您的汽車安全系統的潛力。

閱讀更多

AI 在數據完整性革命中的關鍵角色

AI 數據完整性 機器學習 數據科學 經濟犯罪 腐敗 道德治理

2024-04-10



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

Watch On Demand

在 NVIDIA 最一期的 AI Podcast 中，此節目於 NVIDIA GTC 全球 AI 會議上現場錄製，Cleanlab 的共同創辦人 Curtis Northcutt 與 Berkeley Research Group 的高級數據科學家 Steven Gawthorpe 共享了一種革命性工具的洞見，這工具正在改變數據完整性的遊戲規則。這項工具利用 AI 來精確地識別和糾正數據中的錯誤，顯著提高其可靠性和信任度。他們的對話，由主持人 Noah Kravitz 引導，深入探討了 AI 駕駛的分析怎樣成為對抗經濟犯罪和腐敗的有力盟友，強調了 AI、數據科學和道德治理之間的共生關係，朝著培育一個更公平的社會邁進。

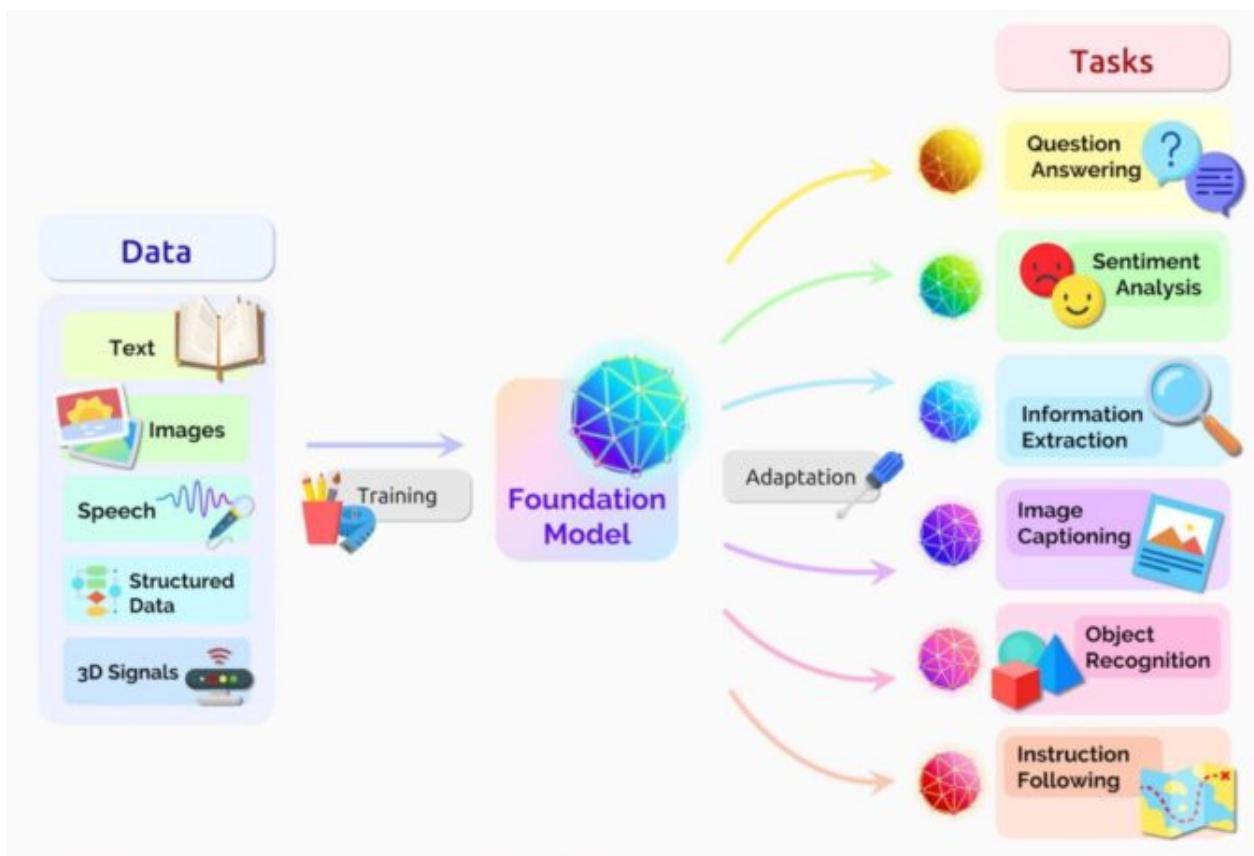
Cleanlab 對數據管理的巧妙方法因其一鍵解決方案而脫穎而出，該方案簡化了機器學習模型的創建過程，展示了機器學習演化所取得的進步及其對數據分析的深遠影響。這項技術不僅僅是關於清理數據；它是為了鋪平一條道路，未來 AI 駕駛的解決方案在打擊犯罪方面發揮關鍵作用，確保數據以前所未有的精確度和可靠性服務於正義事業。

解鎖人工智慧的力量：基礎模型初探

基礎模型 | 神經網絡 | ChatGPT | GPT | NVIDIA | Google Gemma | MistralAI | Meta Llama 2

StabilityAI Stable Diffusion | Microsoft Kosmos 2

2024-04-10



解鎖人工智慧的力量：基礎模型初探

在人工智能的領域中，基礎模型正在成為AI革命的基石，轉變著我們日常與科技互動的方式。想像一個巨大的神經網絡，它經過大海般的數據訓練，能夠理解語言、生成文本、分析圖像，甚至創建影片。這些AI模型就像是讓電腦具備了消化整個圖書館的能力，掌握語言的細微差別，就像人類一樣。

基礎模型迷人之處在於它們的多用途性。它們不僅僅是單一技巧的馬戲表演者；它們可以被微調用於一系列任務，無需每次都從頭開始建立新模型。這代表了時間和資源的顯著節省。無論是生成類人文本、除錯電腦代碼、合成語音，還是創建栩栩如生的圖片和影片，基礎模型都準備就緒。

在這些創新中，ChatGPT璀璨耀眼，證明了OpenAI的GPT模型的力量，現已進入其第四代。但魔法並未止步。NVIDIA進一步推動界限，使這些複雜的模型能夠在個人電腦和工作站上本地運行。這意味著更快、更安全的處理，無需依賴基於雲的服務。

想像一下：在不將任何資料傳送到互聯網的情況下，生成詳細圖像或分析複雜數據。這是一場遊戲規則的變革，尤其是對於那些處理敏感數據或需要快速、實時互動的人來說。

從Google的Gemma，專注於理解和生成文本，到MistralAI在文本生成方面的創造力，以及Meta的Llama 2在文本和代碼產出方面的熟練度，這個領域充滿了潛力。而我們也不應忽視圖像生成領域，StabilityAI的Stable Diffusion模型正在為現實主義和創造力設定新標準。

但或許最令人興奮的發展是多模態模型的出現，如Microsoft的Kosmos 2。這些模型能夠同時處理和理解多種類型的數據，比如文本和圖像，為幾年前還像是科幻小說的應用開闢了新途徑。

隨著這些技術變得越來越易於獲取，我們正站在一個新時代的邊緣，AI的潛力僅受我們想像力的限制。隨著NVIDIA致力於將這些模型帶到你的本地機器，我們不僅僅是在目睹AI的演進；我們正在積極參與其中。

因此，當我們探索基礎模型的廣闊能力時，讓我們擁抱它們所呈現的可能性。AI的未來不僅僅是關於理解數位世界——它是關於重塑它。

[閱讀更多](#)

Microsoft Research 開創語言技術新視野，由 Kalika Bali 領銜

Microsoft Research | 語言技術 | AI 民主化 | 生成式 AI | 教育革新 | 文化包容性

2024-04-11



在 Microsoft Research Podcast 的一次啟發性對談中，傑出的首席研究員 Kalika Bali 揭露了她如何熱情地致力於縮小先進語言技術與世界多樣語言織帶之間的鴻溝。她的職業生涯由“會說話的電腦”之夢點燃，並由對語言學的深刻興趣所推動，Bali 投入了二十多年的時間來協調這些領域。她的使命？讓 AI 技術民主化，使其在每一種語言中都能被取得，特別是對那些全球範圍內被低估和服務不足的社群。

Bali 的工作顯著地由大型語言模型 (LLMs) 的出現和她對 AI 普遍性的堅定信念所支撐。項目如 VeLLM、Kahani 和 Shiksha 成為她的團隊利用生成式 AI 來實現社會利益努力的見證。例如，VeLLM 是一個旨在通過 AI 賦能低資源語言的平台，目標是產生文化和語言包容性的數位內容和工具。

此外，Kahani 作為一項進入視覺故事講述的迷人企劃，旨在通過 AI 生成的圖像創造文化細緻的敘事，使故事對多樣化的受眾群體相關並引人入勝。這項創新不僅為數位內容的包容性開辟了新的維度，也為更豐富、多樣化的故事講述景觀鋪平了道路。

另一方面，Shiksha 正在為印度的教師們革新教育視野。通過大幅減少制定課程計劃所需的時間，Shiksha 賦予教育工作者專注於真正重要的事情——教學和激勵他們的學生。這個項目是 AI 可以為教育場景中有力盟友的閃亮範例，放大教師對學生生活的影響力。

Bali 的旅程，以韌性和創新標誌，展現了一個技術超越語言障礙的未來，促進一個每一個聲音，無論多麼被低估，都能被聽見和理解的世界。通過她的開創性努力，Bali 不僅挑戰現狀，也為所有人啟發了一個更平等和包容的數位未來。

[閱讀更多](#)

美國與日本推進革命性的AI與科技協定

人工智慧 | 量子計算 | 半導體 | AI安全 | 技術協定 | STEM教育

2024-04-11



美國與日本推進革命性的人工智慧與科技協定

在一項突破性的舉措中，美國和日本宣布在人工智慧（AI）、量子計算以及其他關鍵技術領域展開全面合作。這一歷史性的宣布是在日本首相岸田文雄正式訪問白宮之後作出的，與拜登總統一同宣告了科技合作的新時代。

這項夥伴關係的核心是一項1.1億美元的計劃，將華盛頓大學、筑波大學、卡內基梅隆大學和慶應義塾大學的專業知識聚集在一起。在NVIDIA、Arm、Amazon和Microsoft等科技巨頭以及一系列日本企業的支持下，該項目有望將美國和日本推向AI創新的前沿。

這次合作不僅僅是關於開創性研究；它還深深承諾確保AI技術的安全部署。兩國都承諾支持成立國家AI安全研究所，並致力於實現可互操作的AI安全標準和風險管理框架。這包括努力提供AI生成內容的清晰度，確保數位媒體的透明度和真實性。

量子技術，一個有潛力從計算到密碼學革新一切的領域，也是一個主要焦點。美國國家標準與技術研究院（NIST）與日本國立先進工業科學技術研究院（AIST）之間的合作旨在加強量子供應鏈，這是實現這一新興技術全部潛力的關鍵步驟。

此外，該計劃包含半導體發展，計劃探索聯合勞動力發展和技術研討會。這反映了對半導體在現代技術中心腦地位的廣泛認識，從智慧型手機到先進的計算系統。

除了這些技術上的飛躍，該合作強調了培養人才的共同承諾。旨在STEM教育、技術課程開發和人才交流計劃的倡議凸顯了培育新一代創新者和思考者的重要性，這些人將推動數位革命向前發展。

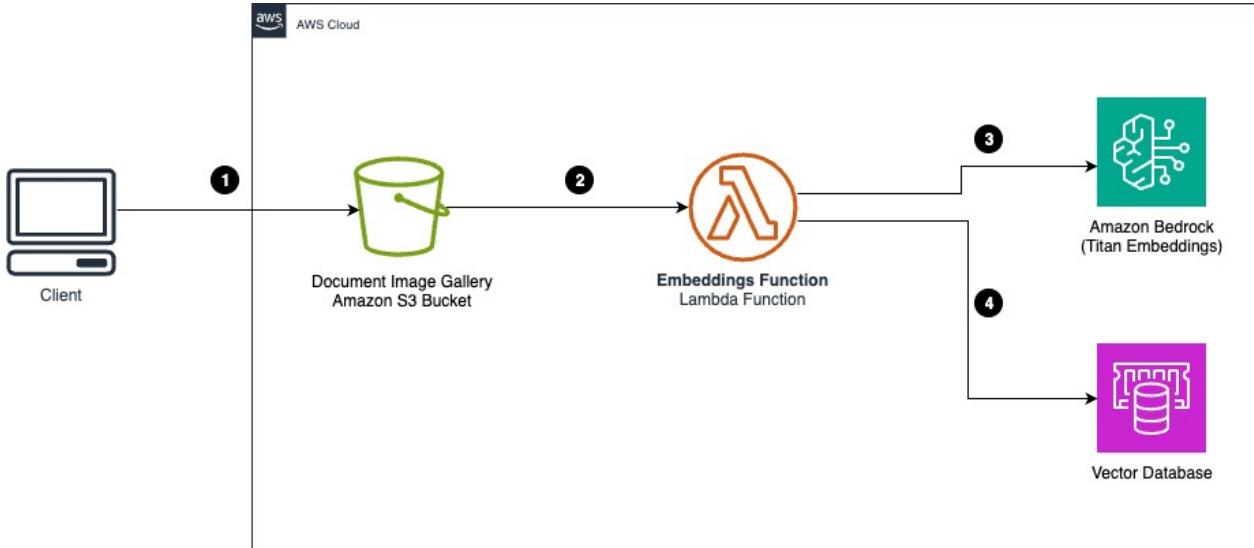
這項美日協定代表著向前邁進的重要一步，不僅在提升技術能力方面，也在塑造一個創新、安全和合作並重的未來方面。這是一個世界的願景，在這個世界中，技術不僅推動了可能性的界限，而且以負責任和包容性的方式進行。

[閱讀更多](#)

轉型文件分類：Amazon Titan多模態 嵌入表示法的革命

Amazon Titan | 多模態嵌入表示法 | 智能文件處理(IDP) | 數位指紋 | 向量表示 | 文件分類 | **Amazon Bedrock** | 上下文搜索 | 歐幾里得L2算法

2024-04-11



轉型文件分類：Amazon Titan多模態嵌入表示法的革命

在今天的數位世界中，各行各業的企業都被各種格式的文件淹沒，使得分類和提取見解成為一項艱鉅的任務。傳統的手工處理方法不僅成本高昂，也容易出錯且難以擴展。Amazon推出了突破性的解決方案：Titan多模態嵌入表示法模型。這項尖端技術，嵌入在Amazon Bedrock之中，為智能文件處理 (IDP) 系統帶來了效率的曙光。

Titan多模態嵌入表示法模型是一個遊戲規則改變者，能夠在沒有事先訓練的情況下理解和分類任何類型的文件。它通過生成“嵌入表示”——可以將其視為數位指紋——來處理圖像和文本，這些嵌入表示隨後可以在新的文件分類工作流程中使用。這些嵌入表示是優化過的向量表示，捕捉文件的本質，無論是掃描成圖像還是由文本組成，都能實現快速索引、上下文搜索和準確分類。

想像一下將新的文件類型整合到您的IDP系統中的便利。透過Amazon Bedrock的API，企業可以動態向量化新模板，增強分類能力，而無需典型的時間和資源投入。這個解決方案不僅僅是有效地分類文件；它是關於以無限可擴展、適應性強和成本效益高的方式來做到這一點。

在這項技術的核心是嵌入表示和向量數據庫，使得基於上下文意義而非僅僅是關鍵字匹配文件的語義搜索成為可能。這種方法允許高度準確的文件分類，利用歐幾里得L2算法實現優化性能。

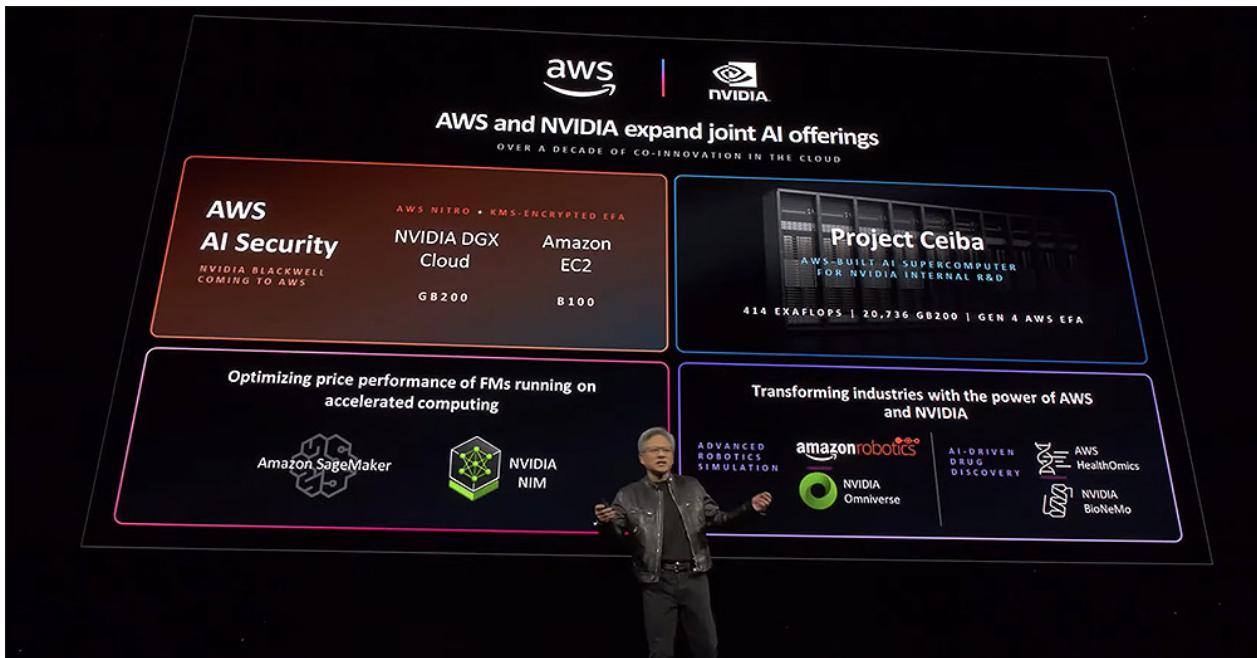
無論您在處理發票、銀行對賬單還是合同，Amazon Titan多模態嵌入表示法模型都能簡化過程，提供一種可擴展、低延遲的解決方案來管理多樣化的文件景觀。這一創新不僅僅是對IDP工作流程的升級；它完全重新想像了文件分類，承諾一個未來——在這個未來中，企業可以更多地關注洞察，而少關注文件管理的複雜性。

[閱讀更多](#)

在 NVIDIA GTC 2024 上的精彩合作與 創新：AWS 與 NVIDIA 領導生成式 AI 革命

生成式 AI | NVIDIA | AWS | GTC 2024 | AI 工作負載 | 大型語言模型 | Project Ceiba | AI 超級電腦 | 生成式 AI Competency

2024-04-11



在 NVIDIA GTC 2024 上的激動人心的合作與創新：AWS 與 NVIDIA 領航生成式 AI 革命

在加州聖荷西的中心地帶，今年三月，AWS 在 NVIDIA 的全球 AI 會議 GTC 2024 中標誌著其顯著的存在。這是一個重大的時刻，吸引了超過 18,000 名現場熱情參與者和令人印象深刻的 267,000 名虛擬與會者。這個混合型活動凸顯了 AWS 與 NVIDIA 深厚的合作夥伴關係，這種合作可以追溯到 13 年前，並且不斷地推動創新的界限，特別是在生成式 AI 的領域。

AWS 與 NVIDIA 之間的協同作用是有形的，兩個巨頭討論他們共同的努力，旨在為全球的開發者民主化進入最先進技術的門檻。他們合作的核心是 NVIDIA 的尖端加速器和 AWS 的強大雲基礎設施，這個組合承諾將革命化 AI 工作負載，從先進的圖形到機器學習。

GTC 2024 的亮點之一是宣佈將新的 NVIDIA Blackwell 平台整合到 Amazon Elastic Compute Cloud (Amazon EC2) 實例中。這一舉措將提高大型語言模型 (LLMs) 的性能，提供安全的 AI 能力和對訓練數據及模型權重的端到端控制。

但這還不是全部。AWS 還揭幕了 Project Ceiba，一個完全由 AWS 與 NVIDIA DGX Cloud 助力的 AI 超級電腦，其驚人的能力達到 414 exaflops。僅此一項目就足以加快 AI 創新的步伐，使更複雜的模型和以前無法達到的突破成為可能。

除了這些技術奇觀，AWS 在 GTC 上展示了其生成式 AI Competency 計畫，為渴望探索並在生成式 AI 解決方案中取得成功的合作夥伴提供資源和最佳實踐。

AWS 與 NVIDIA 之間的這次合作不僅僅是關於技術進步；這是他們致力於推動創新並賦能企業及開發者探索 AI 新領域的證明。展望未來，跨行業產生變革性變化的潛力是巨大的，從生物學和藥物發現到先進的模擬等等。

[閱讀更多](#)

OpenAI在日本推出針對日語的尖端GPT-4定製模型

OpenAI **GPT-4** 日本東京 日語優化 **Speak**英語學習應用程式 **Daikin** **Rakuten** **TOYOTA Connected** **ChatGPT Enterprise** 社會挑戰

2024-04-14

OpenAI選擇日本東京作為其在亞洲的首個辦公室，標誌著其全球擴張的重要一步。這一戰略舉措受到了致力於開發符合日本獨特需求的AI工具的承諾的驅動，促進了與政府、當地企業和研究機構的合作。

在一項令人振奮的發展中，OpenAI推出了為日語優化的GPT-4定製版本。這一定製模型承諾在翻譯和摘要日文文本方面提供增強的表現，擁有比其前身高達三倍的速度。它不僅更高效，而且更具成本效益，預示著日本AI可及性和實用性的新時代。

首批受益者之一的Speak，一個日本頂尖的英語學習應用程式，報告說用戶錯誤的導師解釋速度增加了2.8倍，並伴隨著顯著的成本降低。這一創新為各個領域的學習體驗和生產力的提高鋪平了道路。

OpenAI在日本的存在還加強了與Daikin、Rakuten和TOYOTA Connected等領先公司的合作關係，利用ChatGPT Enterprise來自動化複雜的商業流程並增強數據分析和報告。此外，當地政府已經通過整合ChatGPT體驗到生產力的提升，如在橫須賀市，該技術提高了公共服務的效率。

這項冒險不僅僅是關於技術；它還關於社區。OpenAI熱衷於在日本市場深入整合，由新任OpenAI日本總裁長崎忠夫領導。他們一起旨在通過AI解決方案解決包括鄉村人口減少和勞動力短缺在內的社會挑戰。

隨著OpenAI在全球的足跡擴大，它繼續擁抱多樣化的觀點，這對於其確保AI的好處惠及全人類的使命至關重要。OpenAI日本代表了技術與文化遺產之間的橋樑，承諾一個充滿創新與合作的激動人心的未來。

敬請關注，隨著OpenAI的定製GPT-4模型的推出，承諾將革新行業並增強日本及其他地區的日常生活。

[閱讀更多](#)

NVIDIA的願景：AI作為社會進步的催化劑

NVIDIA | 人工智慧 | 社會進步 | 高性能計算 | 教育 | 超級計算機 | 氣候科學 | 清潔能源

2024-04-15



在俄勒岡州立大學對學生的一次引人入勝的演講中，NVIDIA的CEO Jensen Huang闡述了人工智慧（AI）的變革力量，將其標榜為「技術對社會提升最偉大的貢獻」。在這關鍵時刻，Huang與大學一同慶祝一個價值2.13億美元的研究綜合體的開幕，這是對AI深遠潛力的證明。

Huang不僅將AI視為一項技術奇蹟，更視其為人類的合作夥伴，為社會和經濟提升提供前所未有的機會。這項技術承諾打破歷史上限制獲取數字革命利益的障礙，使編程和數據分析對各個領域更為普及。

以Huang及其妻子Lori命名的新研究綜合體，突顯了俄勒岡州立大學結合高性能計算與各學科領域的雄心。這項倡議旨在利用美國最強大的NVIDIA超級計算機之一，應對全球氣候科學、清潔能源等挑戰。

此外，NVIDIA對教育和勞動力準備的承諾也很明確。通過諸如與大學合作的1.1億美元夥伴關係以增強AI技能，以及在喬治亞理工學院建立AI超級計算中心等舉措，NVIDIA為下一波創新者鋪平了道路。

正如Huang所描述的，這個時代類似於一場新的工業革命，一場在大規模上製造智慧的革命。AI在這場革命中的角色超越了單純的計算；它作為一個導師、一個合作者，讓每一個人，無論其領域為何，都能利用計算的力量來推動他們的工作和創新。

當我們站在這個新世界的邊緣時，NVIDIA在教育和AI領域的倡議預示著向一個包容、技術賦能的社會邁進的充滿希望的旅程。

[閱讀更多](#)

Microsoft 在 2024 年 NSDI 的開創性貢獻：塑造未來網絡系統

Autothrottle | Zipper | Spectrumize | Max-Min Fair Resource Allocation | NetVigil | 5G | IoT | CPU
資源 | SLO | 網絡安全 | 資源管理 | 微服務

2024-04-16



在塑造網絡與分散式系統未來的重要一步中，Microsoft 在 NSDI '24 研討會上呈現了一系列顯著的研究成果。在眾多突出貢獻中，獲獎論文 Autothrottle 介紹了一種新穎的資源管理框架，專為微服務設計，確保在維持服務水平目標 (SLOs) 的同時有效使用 CPU 資源。

創新成果揭曉：

- Autothrottle：這個開創性的框架智慧地管理雲端應用的資源，透過獨特的雙層方法平衡應用的延遲與資源使用，帶來顯著的 CPU 節省和減少 SLO 違規。
- Zipper：一個針對 5G RAN 切片的新系統，Zipper 承諾提供特定應用的吞吐量和延遲改進。使用精密的模型預測控制方法，確保為每個應用的需求提供最佳的頻寬分配。
- Spectrumize：為 IoT 連接帶來革命性創新，Spectrumize 利用衛星網絡中的多普勒偏移來增強數據包的檢測和解碼，為低功耗 IoT 設備提供效率和可靠性的顯著提升。
- Max-Min Fair Resource Allocation：這項研究提出了一個針對大網路資源分配的優化解決方案，確保公平和效率，特別重要的是對於流量工程和集群排程。

- NetVigil：為數據中心安全量身定做的強大異常檢測系統，NetVigil 使用先進的機器學習技術來防範複雜的威脅，在準確性和成本效益上都顯著優越。

這些由 Microsoft 研究人員帶來的創新橫跨廣泛的應用範圍，從提升 5G 功能和衛星網絡到改善雲服務的安全和資源管理。通過這些發展，Microsoft 繼續引領推進我們日益連接的社會所依賴的基礎設施。

[閱讀更多](#)

以Spectrumize革命化全球連接：物聯網（IoT）的未來

物聯網 | 全球連接性 | **Spectrumize** | 衛星網絡 | 多普勒效應 | 頻譜效率

2024-04-16

在令人興奮的前進中，Microsoft Research介紹了一種突破性解決方案，針對數位時代最迫切的挑戰之一：全球連接性。他們的創新論文「Spectrumize：針對物聯網的頻譜效率衛星網絡」，在2024年USENIX研討會上發表，承諾將改變我們在世界最偏遠角落連接的方式。

想像自己是一位在偏僻地區的農民，由於網路連接性差而難以監測作物。Microsoft的研究提供了一線希望，提出了一種方法，使得連接全球的裝置變得簡單如按下按鈕。這不僅僅是一個遙不可及的目標；通過他們在直接對衛星連接技術上的先驅性工作，這個現實被拉近了。這項技術允許裝置與衛星通信，然後衛星將數據回傳地球，確保無論地點如何都能實現無縫連接。

這項創新背後的魔法在於解決頻譜限制 - 這是廣泛衛星通信的一大障礙。Microsoft的團隊利用了移動衛星的一項獨特特性——多普勒效應，來區分信號。這種方法使得多個裝置和衛星可以同時使用同一頻率通信，而不會互相干擾。

這項研究不僅為農民開啟了大門，也增強了能源和供應鏈管理等關鍵行業，使全球連接性民主化。以可負擔性為核心，Microsoft致力於確保世界上每個人，無論身在何處，都能訪問可靠的連接性。

隨著我們展望未來，擴展衛星網絡和物聯網的追求持續進行。下一個重大挑戰之一是在相同頻譜內地面網絡和衛星網絡的共存。Microsoft Research準備好迎接這一挑戰，他們在連接世界的使命中進一步推動界限。

隨著我們進入這個新的連接時代，距離和孤立不再阻礙我們在全球社區中溝通和繁榮的能力。

[閱讀更多](#)

介紹 Idefics2：Hugging Face 的視覺語言模型未來

Idefics2 | 視覺語言模型 | Hugging Face | AI | Transformers | 多模態互動 | OCR

2024-04-16



介紹 Idefics2：Hugging Face 的視覺語言模型未來

發現 AI 技術的最新突破，Hugging Face 推出的 Idefics2。這款最先進的視覺語言模型正在革新機器理解並與周遭世界互動的方式。憑藉著從圖像和文字中理解並生成文字回應的能力，Idefics2 正在為 AI 能力設定新的標準。

想像一下，對一張圖片提出問題並獲得準確、有文本脈絡的答案。或者，讓一個裝置從一系列圖像中講述一個故事，從視覺文件中提取重要資訊，甚至根據它看到的圖像解決算術問題。Idefics2 讓這一切成為可能，推動了 AI 對我們視覺世界理解的界限。

Idefics2 脫穎而出的不僅是其以精準性回答視覺問題的能力，還有它與較大型模型如 LLaVa-Next-34B 和 MM1-30B-chat 相比，儘管只有八十億參數，卻有著有利的比較。它與 Hugging Face 的 Transformers 的整合確保了它可以輕鬆地為各種應用進行微調，使其成為開發人員和研究人員的多功能工具。

通過對包括網絡文件、圖片-標題配對和 OCR 數據在內的多樣化數據集進行培訓，Idefics2 已經掌握了通過改進的光學字符識別（OCR）能力解讀複雜圖表和數據的技能。這與其創新的架構相結合，保持了圖像完整性，避免了常規的調整大小，從而保留了視覺輸入的原始品質和脈絡。

對於那些對將視覺和文字數據融合以創建先進 AI 系統感到好奇的人來說，Idefics2 代表了一次飛躍。其將視覺特徵無縫整合到語言分析中的能力，為開發具有脈絡感知的 AI 應用開啟了新的可能性領域。

這一突破證明了 AI 在增強我們與數位內容互動方面的無窮潛力，使得像 Idefics2 這樣的技術成為未來必不可少的工具。Hugging Face 不僅邀請社群探索 Idefics2 的能力，還為那些有興趣利用其強大功能於他們項目的人提供了全面的教程。

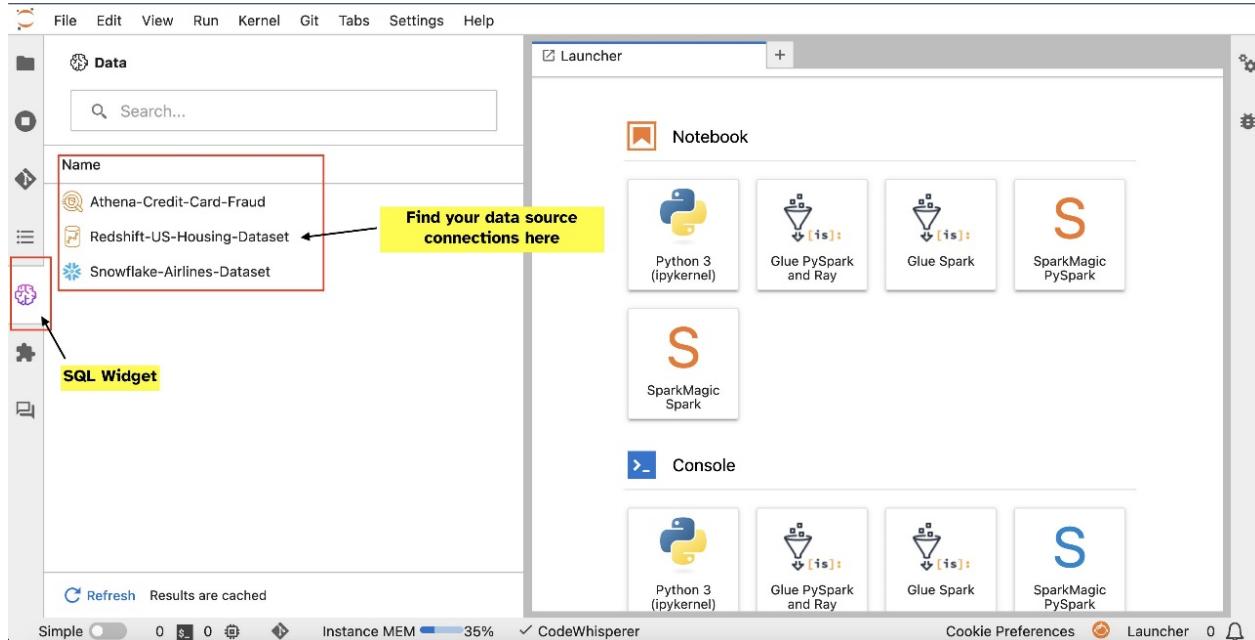
隨著 Idefics2 為下一代 AI 驅動的多模態互動鋪平道路，請持續關注其展開的可能性。

[閱讀更多](#)

SageMaker Studio 透過 Text-to-SQL 及直接資料庫連線革新 SQL 查詢



2024-04-16



SageMaker Studio 以 Text-to-SQL 和直接資料庫連線革新 SQL 查詢在資料科學家和機器學習愛好者感到興奮的發展中，Amazon SageMaker Studio 推出了一項新功能，旨在顯著簡化分析和利用數據的工作流程。SageMaker Studio，以其全面的環境著稱，用於建立、訓練和部署機器學習模型，現在直接在 JupyterLab 筆記本中整合了 SQL 支援，使得與各種資料源連接、探索和操作數據變得前所未有的容易。輕鬆實現直接資料庫連線在機器學習項目中切換多種工具來管理 SQL 查詢的日子已經一去不復返了。隨著內建 SQL 支援的引入，使用者可以輕鬆連線到流行的數據服務，如 Amazon Athena、Amazon Redshift、Amazon DataZone 和 Snowflake。這種整合允許在 JupyterLab 介面內無縫瀏覽、搜索和預覽資料庫、架構、表格和視圖。此外，這項功能支援在同一筆記本中混合 SQL 和 Python 代碼，增強了機器學習任務中數據探索和轉換的效率。透過開發者友好功能提升您的生產力為了幫助快速開發和調試 SQL 查詢，SageMaker Studio 的 JupyterLab 筆記本提供了開發者生產力增強功能，如 SQL 命令完成、代碼格式化幫助和語法高亮。這些功能不僅節省時間，還使用者更容易編寫和完善他們的 SQL 代碼，提高整體生產力。利用 Text-to-SQL 功能的力量除了直接資料庫連線外，SageMaker Studio 還利用先進的大型語言模型 (LLMs) 的能力引入了 Text-to-SQL 功能。這一開創性功能使使用者能夠從自然語言描述生成 SQL 查詢，即使是那些對 SQL 經驗最少的人也能輕鬆編寫復雜的查詢。通過簡化查詢編寫過程，Text-to-SQL 為更廣泛的受眾開放了數據探索和分析，促進了更包容和高效的數據驅動決策。部署 Text-to-SQL 模型以實現更廣泛的使用對於希望擴大使用 Text-to-SQL 功能的團隊，SageMaker Studio 提供了將這些

模型作為 SageMaker 端點部署的選項。這種方法允許靈活且可擴展地託管自定義模型，使更廣泛的用戶群能夠無縫地從自然語言輸入生成 SQL 查詢。數據科學生產力的一大飛躍Amazon SageMaker Studio 整合 SQL 支援和 Text-to-SQL 功能，標誌著數據科學和機器學習領域的一個重大進步。通過消除對數據訪問和查詢生成的障礙，SageMaker Studio 使資料科學家能夠更多地專注於他們的核心任務——構建、訓練和部署模型——同時享受更流暢和高效的工作流程。今天就在 SageMaker Studio 探索這些新功能，改變您與數據和機器學習項目合作的方式。

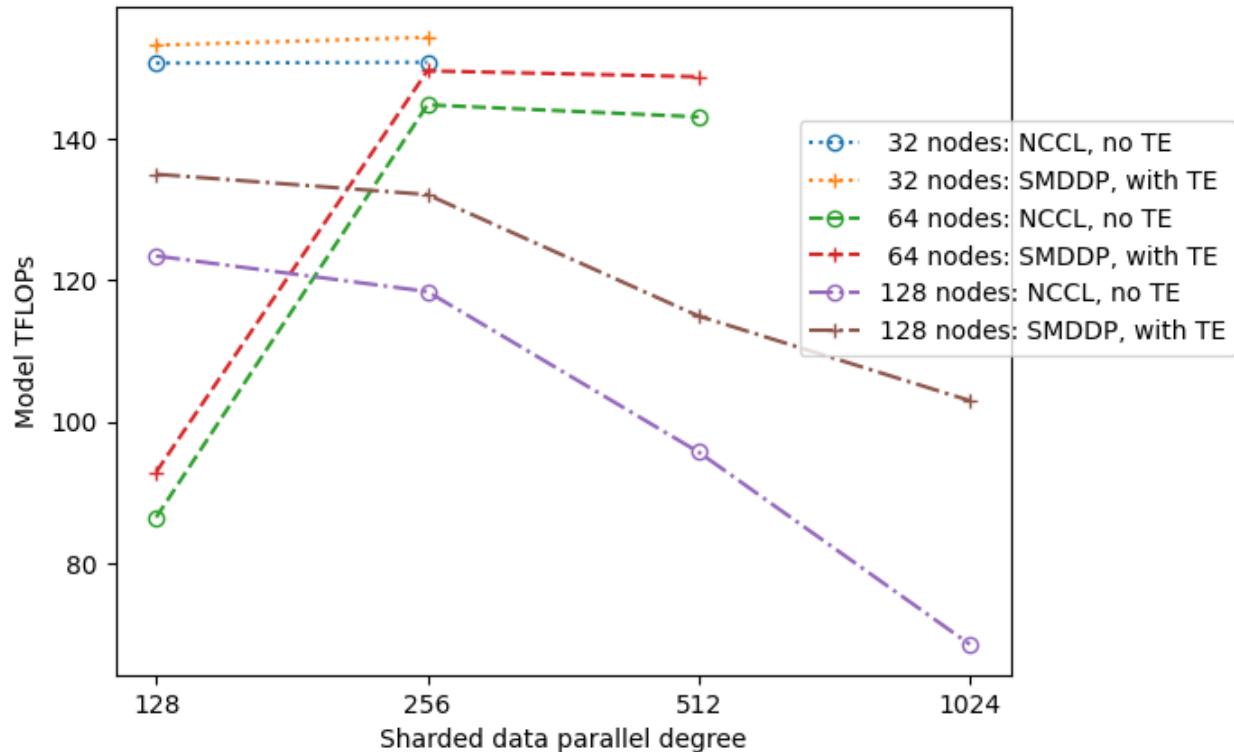
[閱讀更多](#)

Amazon SageMaker 透過新函式庫革新大型語言模型訓練

Amazon SageMaker	大型語言模型	訓練	模型並行庫 2.0	分散式資料並行庫	效率	可擴展
Transformer Engine	張量並行技術	吞吐量	性能			

2024-04-16

Llama 2: 70B, sequence length 4096



Amazon SageMaker 透過新函式庫革新大型語言模型訓練

在一項突破性的發展中，Amazon 推出了其最新創新，以解決訓練大型語言模型（LLMs）的複雜性 - SageMaker 模型並行庫 2.0（SMP）和 SageMaker 分散式資料並行庫（SMDDP）。這些技術旨在使包含數十億甚至數萬億參數的龐大模型的訓練更加高效和可擴展。

傳統上，訓練如此巨大的模型需要在數百或數千個 GPU 上分配工作負載，因為單個設備上的記憶體限制。這在分散式訓練效率和可擴展性方面帶來了重大挑戰。然而，Amazon 的新發布標誌著正面解決這些挑戰的轉折點。

SageMaker 模型並行庫 2.0 與開源 PyTorch 完全分片資料並行（FSDP）APIs 整合，為訓練大型模型提供了一個熟悉的介面。這種兼容性擴展到了 Transformer Engine（TE），首次結合了 FSDP 的張量並行技術。SMP 和 SMDDP 之間的這種協同作用揭示了大型模型訓練的最先進效率，展示了近線性的擴展效率。

實際上，這意味著現在在 Amazon SageMaker 上訓練數十億參數的模型可以實現顯著的吞吐量和性能。例如，訓練 Llama 2 模型，其大小範圍從 70 億到 70 億參數，已經在各種配置中展示了穩定的擴展效率，大大減少了整體訓練時間。

此外，與 TE 的整合和對所有收集集體操作的優化推動了分散式訓練可能性的界限。這些進步不僅緩解了通信瓶頸，而且通過激活卸載和支持長序列長度，特別是在低資源設置中提高了吞吐量。

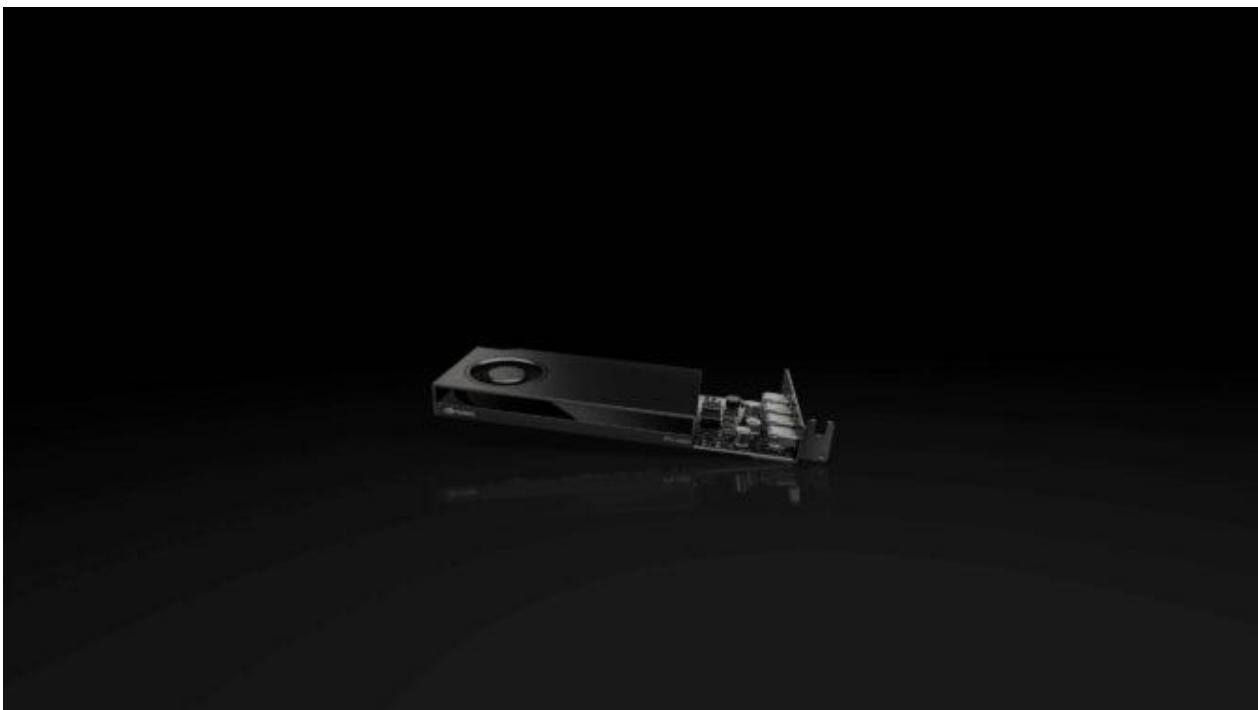
本質上，Amazon 的 SageMaker 模型並行庫 2.0 和分散式資料並行庫正在重新定義 LLM 訓練的格局。通過提供可擴展的解決方案和卓越的性能，這些技術賦予了研究人員和實踐者同樣的權力，為下一代 AI 進步鋪平了道路。

[閱讀更多](#)

透過 NVIDIA 最新 GPU 釋放創造力與效率

NVIDIA | GPU | AI | RTX A400 | RTX A1000 | 光線追蹤 | 生成式 AI | Ampere 架構 | 編碼 | 解碼

2024-04-16



NVIDIA 正將專業圖形處理提升至新的水平，推出了 RTX A400 與 A1000 GPU，旨在為 AI 驅動的設計和生產力加速。這些強大的設備基於尖端的 NVIDIA Ampere 架構，為每一個工作站帶來即時光線追蹤和生成式 AI 工具的魔法，無論項目規模如何。

RTX A400 GPU 是一款改變遊戲規則的設備，擁有 24 個 Tensor 核心用於 AI，使得在您的桌面上就能執行先進的 AI 應用程序。它在提供即時光線追蹤方面表現出色，使創作者能夠產生生動、物理精確的 3D 渲染圖像，模糊了數位與現實之間的界限。它支援同時連接四個顯示器，非常適合高密度環境，為從金融到運輸的各行各業帶來革命。

另一方面，RTX A1000 GPU 憑藉其 72 個 Tensor 核心，代表了一個重大的飛躍，提供了三倍以上的生成式 AI 處理速度。它是創意人士和專業人員的夢想，加速了 2D 和 3D CAD、視頻編輯等任務，同時其視頻處理能力無與倫比，處理更多的編碼流並提供了比前代產品雙倍的解碼性能。

這些 GPU 不僅在力量上出色，它們還以效率設計。它們的單插槽、50W 設計意味著它們可以適配於緊湊、節能的工作站，而不會犧牲性能。

通過為用戶配備這些先進工具，NVIDIA 不僅提升了生產力，更是為新的創造力領域開啟了大門。不論是針對工業規劃、內容創建、建築或健康護理，這些 GPU 正在為可能的新標準設定新的水平，使工作流程中的真實度和效率達到驚人的水平。

專業圖形的未來已經到來，憑藉 RTX A400 和 A1000 GPU，NVIDIA 再次站在了前沿，推動著行業前進。

[閱讀更多](#)

革命性的影片編輯：DaVinci Resolve 19 與 AI 的力量

DaVinci Resolve | **AI** | **IntelliTrack** | **UltraNR** | **NVIDIA** | **RTX** | **GPU** | **TensorRT** | **SketchUp** | **DirectX**
12 | **Unreal Engine** | **Adobe Premiere Pro**

2024-04-16



革命性的影片編輯：DaVinci Resolve 19 與 AI 的力量

在影片內容稱霸的世界裡，影片編輯工具的發展速度讓人嘆為觀止。引領潮流的 Blackmagic Design 推出了 DaVinci Resolve 的第 19 版，內含許多新功能，定將重新定義編輯格局。

焦點投向兩項突破性的 AI 驅動功能：IntelliTrack 與 UltraNR。IntelliTrack 利用人工智慧進行精確的物件追蹤、穩定以及甚至音頻定位，讓編輯者更輕鬆地精煉其影片。與此同時，UltraNR 使用 AI 進行空間降噪，相較於 Mac M2 Ultra，在強大的 GeForce RTX 4090 上達到了前所未有的三倍速完成度。

這些 AI 效果，藉由 NVIDIA 的 TensorRT 在 RTX GPU 上加速，承諾提供高達兩倍的 AI 效能，讓美容增強、邊緣偵測和添加水彩效果等任務流暢如絲。這種創意與技術的融合確保了 DaVinci Resolve 繼續是影片編輯專業人士和愛好者們不可或缺的工具。

但創新不僅停留在影片編輯上。SketchUp 2024 以全新的由 DirectX 12 駕駛的圖形引擎亮相，顯著增強了建築師和設計師的設計流程。這次更新不僅使得導航及繞行複雜模型變得輕而易舉，而且在 NVIDIA RTX 4090 GPU 上顯示出顯著的性能提升。

作為故事講述和技術力量的見證，藝術家 Rakesh 揭幕了一幕令人著迷的 3D 场景「The Rooted Vault」，該場景利用 NVIDIA 的 RTX 加速技術精心打造。他從收集參考資料到在 Unreal Engine 5 和 Adobe Premiere Pro 中做最後修飾的旅程，凸顯了 GPU 加速對渲染和實時編輯的轉變性影響。

隨著科技持續突破可能的界限，由 NVIDIA RTX 技術驅動的 DaVinci Resolve 19 和 SketchUp 2024 等工具，不僅促進了創造力；它們為世界各地的故事講述者、設計師和建築師開啟了新的大門。

[閱讀更多](#)

本週聚焦Microsoft Research的創新

AI可靠性 | 雲計算效能 | 生成式AI | 大型語言模型 | 電源管理 | AI模型壓縮 | 機器翻譯 | 混合會議 | 圖像解析度

2024-04-17



本週聚焦於Microsoft Research的創新

Microsoft Research在科技領域內展現了多元的進步，從提升AI可靠性到優化雲計算效能。以下是他們開創性工作的簡要概述：

- 導航生成式AI中的信任問題：Microsoft深入探討了使用者如何平衡對AI生成內容的信任。他們的研究強調了辨別準確AI建議與誤導性建議的重要性，旨在減少過度依賴的風險，這可能導致決策失誤甚至不再使用AI技術。
- 雲端中大型語言模型(LLMs)的電源管理：隨著對AI驅動語言模型需求的飆升，Microsoft提出了一種新框架POLCA，幫助雲服務更有效率地運行更多伺服器，而不會消耗過多能源。這項創新承諾提供更綠色、更可持續的方法來滿足日益增長的AI計算需求。
- 利用LLMLingua-2加速AI：Microsoft與清華大學合作，引入了一種高效的方法來壓縮AI模型提示，顯著加快處理速度而不損失關鍵信息。這項進步可以使AI技術更易於獲得且更具成本效益。

4. 橋接語言差距，透過AfrimTE與AfricomET：Microsoft與幾家學術機構合作，旨在通過改進機器翻譯支援資源不足的非洲語言。他們的工作旨在使AI更具包容性，將其好處擴展到更廣泛的語言和文化。
5. 通過Hybridge增強混合會議：Microsoft探索3D介面，為混合會議中的遠程參與者提供平等的競爭條件。他們的發現表明，相比傳統的2D視頻會議，3D設置可以創造更具參與感和包容性的環境。
6. FeatUp：提高AI分析的圖像解析度：Microsoft推出FeatUp，一種多功能工具，可提高AI分析的圖像解析度，確保在過程中不會丟失細節。這可能會革命化從醫療成像到監控的任務，其中清晰度至關重要。

Microsoft Research的每個項目都展示了對推動技術達到新高度的承諾，確保它以負責任且包容的方式服務於社會。

[閱讀更多](#)

SAS 透過為所有技能水平打包模型革命化 AI 的可及性

AI | SAS | 模型 | 整合性 | 民主化 | 行業解決方案

2024-04-17



在一項讓人工智慧 (AI) 更容易被大眾接觸的舉措中，SAS 推出了一系列針對特定行業的 AI 模型，使用者無需成為資料科學家就能部署它們。這項倡議將改變各種規模的組織如何將 AI 融入他們的運作，使得以 AI 的力量來解決商業挑戰變得更快速且更有效率。

想像一下，即使您這輩子從未寫過一行程式碼，也能在 AI 助理的幫助下優化您的倉庫空間。這就是 SAS 帶來的簡單力量。這些輕量級、即插即用的模型旨在加速實際應用，如欺詐檢測、供應鏈優化和醫療保健支付完整性等。

這些模型的突出之處在於它們的易於整合性。它們可以迅速部署，使組織能夠在沒有傳統複雜性和延長時間線的情況下受益於 AI。這在當今快節奏的市場中尤其關鍵，敏捷和快速決策可以成為顯著的競爭優勢。

SAS 致力於使 AI 更容易被接觸，是一項更廣泛的 10 億美元投資的一部分，旨在提供以 AI 驅動的行業解決方案。這些模型建立在 SAS 豐富的經驗上，並為企業需求進行了微調，確保企業能夠利用 AI 獲得可信的結果和具體的好處，無論他們的技術專長如何。

這項倡議不僅僅是簡化 AI，也是使其民主化。通過使這些強大的工具可用且易於使用，SAS 正在幫助各種規模的企業和擁有不同技能集的個人發揮 AI 的潛力，為各個領域的創新鋪平了道路。

請密切關注，因為這些開創性的 SAS 模型預計將在今年晚些時候推出，標誌著使 AI 更容易被日常商業場景應用和可及的重要里程碑。

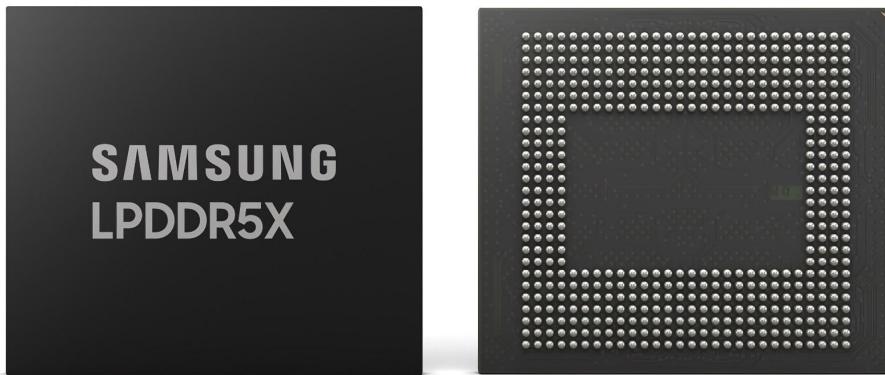
對這項發展感到興奮嗎？這種方法可能會重新定義我們如何看待和使用我們行業和工作場所中的 AI，這是一個值得密切關注的話題。

[閱讀更多](#)

三星以LPDDR5X DRAM設立新標準， 提升AI性能

LPDDR5X | 三星 | AI性能 | 節能技術 | 高性能記憶體

2024-04-17



在技術上取得重大飛躍的三星，推出了LPDDR5X DRAM，這是一款在記憶體晶片領域中的奇蹟，擁有高達10.7 Gbps的驚人速度。這項創新採用先進的12奈米級工藝技術設計，不僅超越了先前的性能基準，還在更小的晶片尺寸上實現了這一壯舉。

這一發展尤其令人興奮，因為它預示著在設備上人工智能（AI）的新時代即將到來。三星記憶體產品規劃部門的執行副總裁YongCheol Bae表示，焦點從移動設備轉向了包括PC、汽車、服務器和加速器在內的高性能領域。LPDDR5X正準備滿足對既高效又強大的記憶體日益增長的需求，使我們的設備能夠運行更複雜的AI應用。

但這對日常用戶和行業來說意味著什麼呢？首先，與其前身相比，這款新晶片不僅性能提高了逾25%，容量增加也超過了30%，這意味著我們的設備將變得更智能、更快速且更高效。想像一下，你的智能手機在運行複雜的AI應用時不會耗盡電池，或者數據中心的服務器在處理大量數據時能更迅速地運行，同時消耗更少的電力。

此外，LPDDR5X融合了最先進的節能技術，能夠根據工作負載需求動態調整能源使用。這不僅延長了移動設備的電池壽命，還降低了像服務器這樣的大規模應用中的能源使用，最終減少了運營成本。

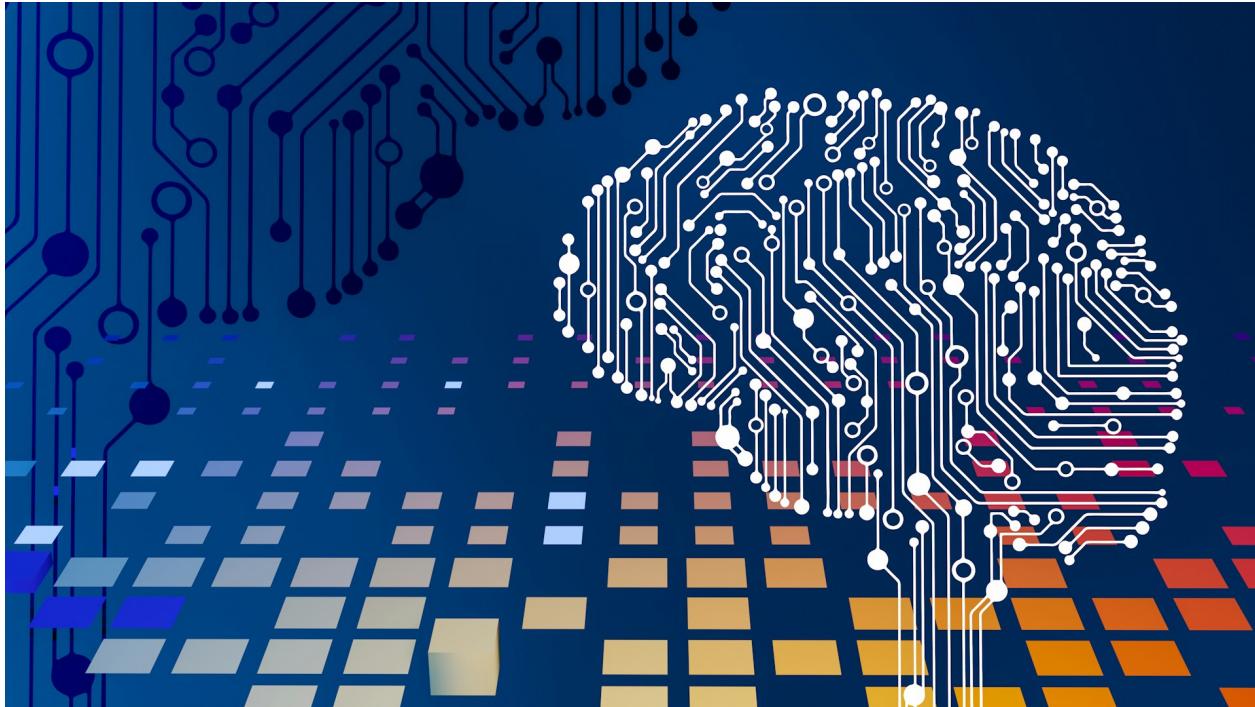
三星對LPDDR5X DRAM的遠景旨在革新我們與技術的互動方式，使設備上的AI變得比以往任何時候都更強大和可訪問。隨著我們熱切期待這些晶片在2024年下半年的大規模生產，顯然，多虧了這些創新，AI和移動技術的未來是光明的。

[閱讀更多](#)

透過生成式AI轉型業務：來自Ikigai Labs的Kamal Ahluwalia洞察

生成式AI 大型語言模型 大型圖形模型 業務轉型 數據安全 道德考慮

2024-04-17



在迅速進化的人工智慧領域中，Ikigai Labs的Kamal Ahluwalia提供了一窺企業創新未來的機會，透過生成式AI。Ikigai Labs站在最前沿，將零散且結構化的數據挑戰轉化為預測性和可行性的洞察力。這種方法對於那些數據成堆卻缺乏洞察力的企業尤其關鍵。

生成式AI，特別是其最受關注的形式 - 大型語言模型(LLMs) - 已經在科技界引起了轟動。然而，LLMs主要針對非結構化數據，當涉及到構成企業決策支柱的結構化、表格化數據時，就出現了缺口。Ikigai Labs用其專有的大型圖形模型(LGMS)填補了這一缺口，承諾將如何在銷售、產品開發、員工管理和財務規劃等關鍵領域進行預測和規劃的方式帶來革命。

將生成式AI整合到業務運營中的旅程並非沒有障礙。成本、數據安全和對準確性的需求構成了強大的障礙。然而，Ikigai採用創新的方法，專注於時間序列數據並利用LGMS，提供了一條強調成本效益、安全性和精確性的前進道路。

對於希望利用生成式AI的公司來說，道路在於利用這項技術來解決獨特的業務挑戰，而不是追求通用的應用。從優化供應鏈到預測市場需求，甚至是應對勞動力管理的複雜性，潛在的應用既多樣又具有影響力。

然而，擁生成式AI不僅僅是採納新技術；它還需要組織內部的文化轉變。領導層的承諾、持續的教育和願意實驗是解鎖AI變革潛力的關鍵。

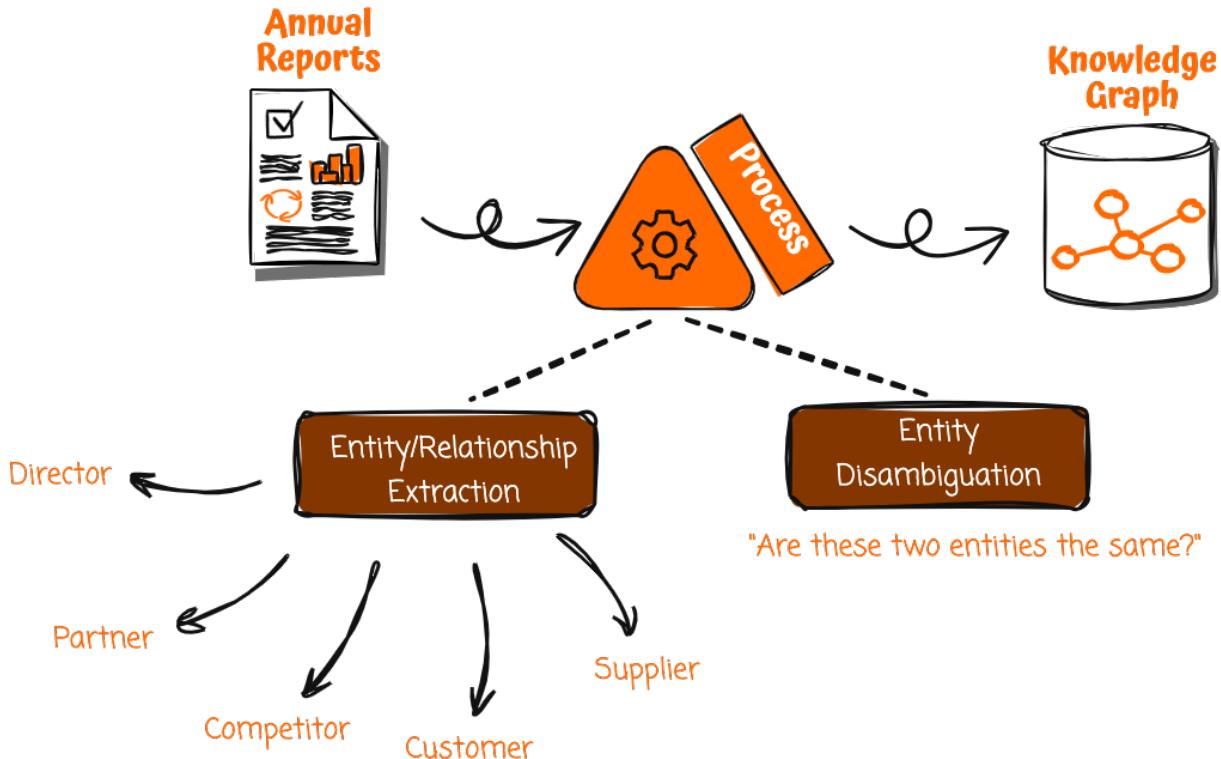
隨著AI不斷成熟，對於企業來說，問題不再是是否應該採用生成式AI，而是他們能多快地做到這一點，同時確保道德考慮始終處於前沿。有了Ikigai Labs的榜樣，由生成式AI驅動的業務未來不僅看起來智能，而且富有洞察力且在道德上有根基。

[閱讀更多](#)

創新分析：透過 Amazon Bedrock 和 Neptune 提升投資策略

[Amazon Bedrock](#) [Amazon Neptune](#) [投資策略](#) [知識圖譜](#) [生成式人工智能](#) [自然語言處理](#) [風險管理](#)

2024-04-17



創新分析：透過 Amazon Bedrock 和 Neptune 提升投資策略

在快節奏的資產管理世界中，保持領先不僅僅是關注明顯的市場觸發因素，如財報或公司降級。那些隱藏的連結，那些公司更廣泛生態系統內事件的第二和第三層次影響，往往才是識別風險和機會的關鍵，這些風險和機會可能不會成為頭條新聞。

Amazon 推出了一個尖端解決方案，旨在改變投資組合經理人揭露這些關鍵洞見的方式。利用知識圖譜和生成式人工智能（AI）的力量，這種先進的分析工具能夠發現傳統方法可能錯過的潛在市場變動。

這項創新的核心是兩種 Amazon Web Services (AWS) 產品：Amazon Neptune 和 Amazon Bedrock。Neptune 提供了一個強大的、完全管理的圖形資料庫服務，非常適合處理高度連接的數據集。這使它非常適合繪製公司及其供應商、客戶和合作夥伴之間複雜的關係網。另一方面，

Amazon Bedrock 則是自然語言處理的動力源泉，提供了來自領先 AI 公司的高性能基礎模型的訪問。這使得從公司年度報告等非結構化來源自動提取和結構化有價值數據成為可能。

想像一下，通過自動化、無伺服器和可擴展的架構，自動識別供應商運營中斷可能對下游製造商造成的影響的優勢。這項技術不僅簡化了從密集年度報告建立全面知識圖譜的過程，還豐富了投資組合經理人的新聞摘要，用文章精確指出與他們的投資興趣相關的潛在影響發展。

通過連接不同資訊之間的點，投資專業人士可以獲得更清晰、無噪音的市場景觀視圖。這不僅有助於風險管理，還揭露了可能否則一直隱藏的機會。隨著市場動態的演變，這樣的能力將變得越來越關鍵，以維持競爭優勢。

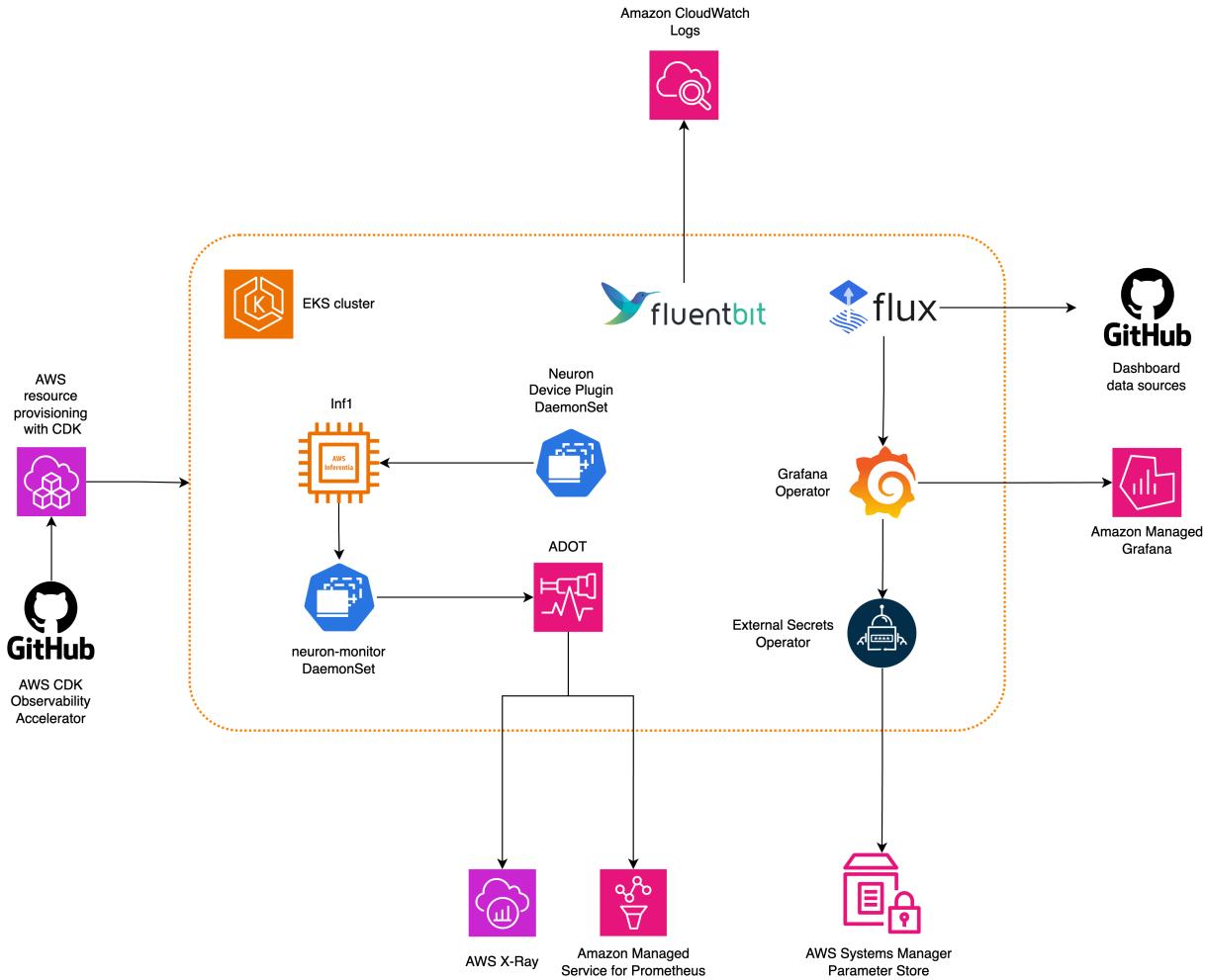
Amazon 的 Neptune 和 Bedrock 技術的結合，是投資分析未來的典範，深層的、AI 驅動的洞察使得更加明智的決策成為可能，從而潛在地導致優越的投資結果。

[閱讀更多](#)

利用 AWS 上的開源可觀察性解鎖機器學習的未來

AWS 可觀察性 機器學習 Amazon EKS AWS Inferentia Amazon Managed Grafana Amazon Managed Service for Prometheus AWS Neuron

2024-04-17



利用 AWS 上的開源可觀察性解鎖機器學習的未來

在快速發展的機器學習（ML）領域中，模型的複雜性和大小達到了新的高度，這需要前所未有的計算資源規模。為了應對這些挑戰，Amazon Web Services (AWS) 在 Amazon Elastic Kubernetes Service (EKS) 集群內的 AWS Inferentia 節點中引入了一種突破性的可觀察性方法。

這項創新的本質在於能夠實時監控和優化 ML 晶片的性能。這是通過為 AWS Inferentia（一款專為機器學習設計的晶片）專門設計的開源可觀察性模式實現的。這種模式為無與倫比地洞察 ML 晶片如何在 Amazon EKS 集群內被利用和管理鋪平了道路，最終導致模型性能和成本效率的提升。

這個解決方案的核心是一套工具，包括 Amazon Managed Grafana 和 Amazon Managed Service for Prometheus，再加上 AWS Neuron 裝置插件。這些組件無縫地協同工作，收集、儲存並可視化來自 ML 晶片的度量，為您的 ML 基礎設施提供了全面的視角。這使得數據科學家和工程師能夠微調資源，識別和處理異常，並就容量規劃做出明智的決策。

這種方法區別於其他的不僅僅是其技術實力，還有其可接近性。通過利用開源模式和 AWS CDK Observability Accelerator，AWS 提供了一種流暢高效的方式來將可觀察性實施到在 EC2 Inf1 實例上運行的 EKS 集群中。這使得各種規模的團隊都能夠實現其 ML 努力中的最先進性能，民主化了對先進監控功能的訪問。

通過這個創新解決方案，AWS 不僅提高了機器學習模型的效率和效果，還賦予了組織推動 ML 可能性邊界的能力。這是對開源和雲計算在推動技術未來發展中力量的證明。

[閱讀更多](#)

解鎖隱形光譜：Living Optics 提升高光譜成像技術

高光譜成像 | Living Optics | NVIDIA Inception | 監測植物健康 | 橋樑結構弱點檢測

2024-04-17



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

Watch On Demand

解鎖隱形光譜：Living Optics 將高光譜成像技術提升到新的高度

在最近一次與 AI Podcast 主持人 Noah Kravitz 的精彩對話中，創新初創公司 Living Optics 的 CEO Robin Wang 分享了他們的突破性高光譜成像相機如何在各個行業改變遊戲規則的見解。與傳統成像技術不同，Living Optics 的相機能夠捕捉到令人印象深刻的 96 種顏色光譜數據，遠超人眼所能察覺的範圍。這項先進技術為世界帶來了無限可能，從通過監測植物健康提高農業生產力，到通過早期檢測橋樑結構弱點來提升公共安全。

如 Wang 所解釋，高光譜成像提供了更豐富、更詳細的數據集，讓使用者能夠揭示以前看不見的洞察。Living Optics，作為 NVIDIA Inception 計劃的自豪成員，旨在使這項強大的技術大眾化，讓不同領域的創新應用變得觸手可及。這一倡議可能會革新我們解決問題和制定決策的方式，利用無與倫比的細節和準確性來指導我們的策略。

請繼續關注 Living Optics 在技術與發現交匯處持續發展的進展，因為 Living Optics 將繼續為高光譜成像的新應用鋪平道路。

[閱讀更多](#)

與NVIDIA的Instant NeRF一起踏進未來：將2D圖像轉變為3D世界

NVIDIA Instant NeRF 3D場景 神經輻射場 RTX技術 藝術家 設計師

2024-04-17



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

Watch On Demand

在數位創意領域裡，NVIDIA推出的Instant NeRF是一項激動人心的技術飛躍，它能在幾秒內將靜態圖像變為生動的3D場景。這項創新為藝術家、設計師以及任何渴望探索想像界限的人開啟了新的維度。

Instant NeRF的核心理念是神經輻射場（NeRF），一種AI模型，它接收一系列2D圖像並將其賦予生命，創造出完全實現的3D環境。借助NVIDIA RTX技術的力量，包括Tensor Cores的實力，Instant NeRF迅速且高效地執行這項魔法，使得曾經在計算上令人生畏的任務現在變得令人印象深刻的可及。

Instant NeRF的特別之處在於它能夠把握並渲染光線和空間的複雜性，生成具有深度和細節的場景，邀請人們探索。從將一系列簡單的照片轉化為動態3D風景，到為虛擬遺產之旅復興文化文物，其應用範圍廣泛且令人興奮。

這項技術不僅僅適用於技術嫻熟之人；憑藉設計簡易的工具和指南，任何人都可以深入3D場景創建的世界。無論你是開發者、藝術家，或只是對數位創新的下一波感到好奇的人，Instant NeRF為你提供了體驗到目前為止難以想像的經歷的大門。

想像一下，在不進入房產內部的情況下進行導覽，或在購買前以全3D形式體驗產品。Instant NeRF正在為這些體驗鋪平道路，轉變我們與數位內容互動的方式。

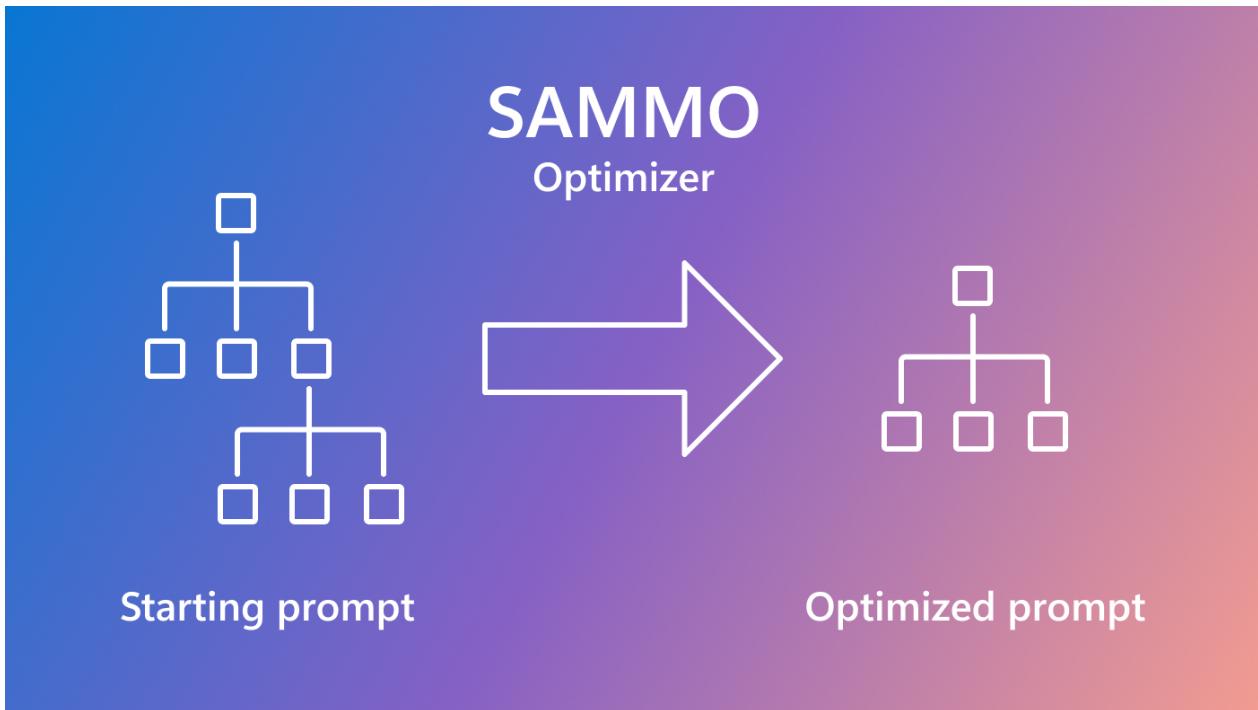
與Instant NeRF一同潛入未來，看看你的創造力能帶你到哪裡。無論你是在製作沉浸式數位藝術、以新維度保存歷史遺址，還是僅僅探索3D環境的潛力，Instant NeRF承諾的旅程既豐富又革命性。

[閱讀更多](#)

Microsoft Research 推出 SAMMO：透過提示優化轉變 AI

Microsoft Research | SAMMO | AI | 提示優化 | 大型語言模型 | LLMs | GPT-4 | Mixtral 8x7B | 結構感知的多目標元提示優化

2024-04-18



Microsoft Research 推出 SAMMO：透過提示優化轉變 AI

在人工智能領域中，Microsoft Research 已經透過推出 SAMMO 這一開創性框架，向前邁出了重要一步，專門用於AI提示的優化。這一創新工具將改變我們與大型語言模型（LLMs）的互動方式，使得根據特定任務量身定製 AI 回應變得更加容易和高效。

提示優化一直是充分利用 AI 潛力的挑戰。隨著像 GPT-4 和 Mixtral 8x7B 這樣可以處理廣泛輸入的語言模型的進步，更精緻的提示設計方法的需求已經變得明顯。SAMMO（結構感知的多目標元提示優化）是 Microsoft 對這一挑戰的回應，提供了一種以最小的手動努力來微調 AI 提示的方法。

SAMMO 的獨特之處在於它能夠將提示不僅僅視為文本字符串，而是作為動態的、可編程的實體。通過將提示表示為函數圖，SAMMO 允許修改個別組件和子結構以優化性能。這可能意味著從刪除不必要的部分到納入特定領域知識，所有這些都旨在使 AI 在執行任務時更加準確和高效。

SAMMO 的關鍵特性包括其結構化的優化方法，關注於提示的架構而不僅僅是文本，以及其多目標搜尋能力，同時解決多個優化目標。此外，SAMMO 已經展示了顯著的多樣性，跨越從指令調整到檢索增強生成（RAG）的應用範圍顯示出顯著的性能提升。

SAMMO 的推出不僅是一項技術成就；它代表了我們如何進行 AI 提示設計的範式轉變。通過使提示的自定義和優化變得更加容易，SAMMO 為更有效和用戶友好的 AI 應用鋪平了道路。這一突破承諾將增強 AI 助手的實用性，使它們對特定用戶需求更加響應和適應。

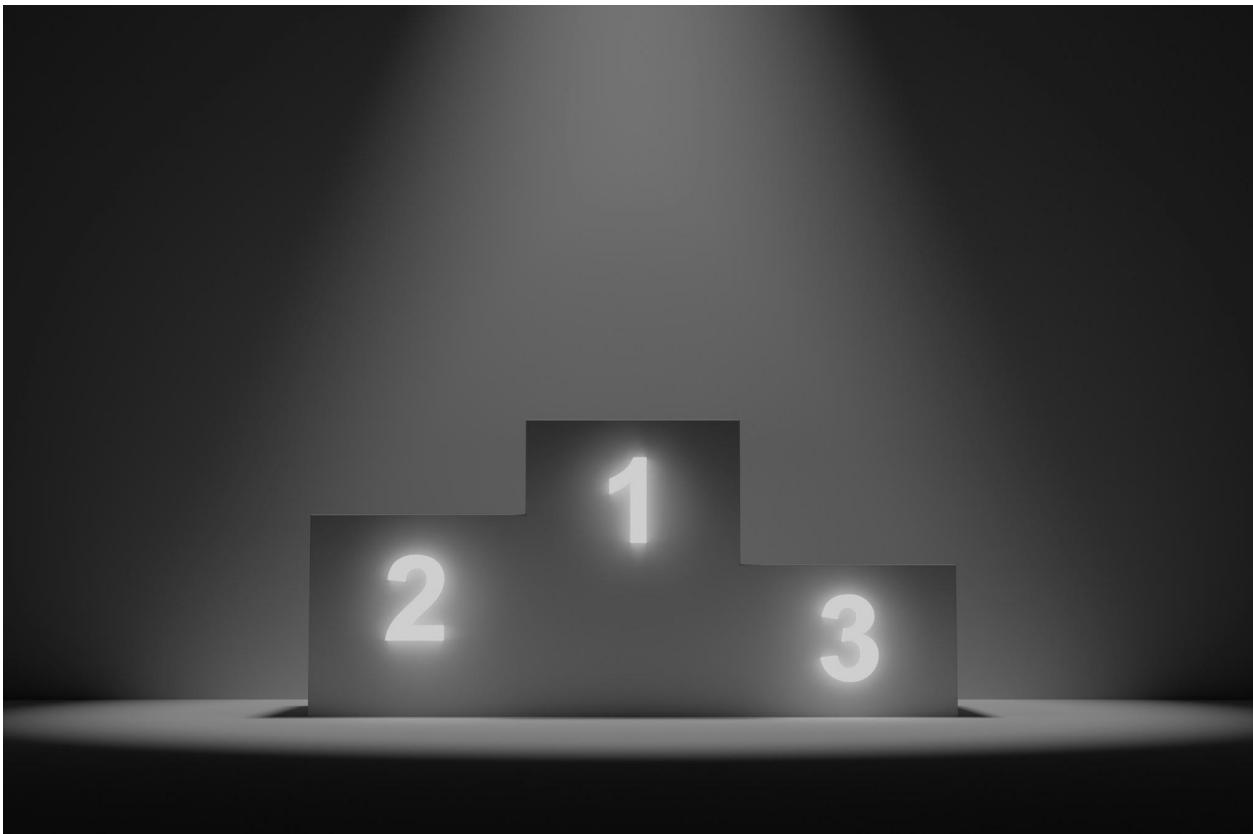
隨著 Microsoft Research 繼續探索和開發 SAMMO，圍繞這一工具的用戶驅動社區的潛力浮現，為分享最佳實踐和擴展其能力提供了一個平台。隨著 SAMMO 的出現，AI 提示優化的未來看起來光明，預示著人工智能效率和精確度的新時代。

[閱讀更多](#)

介紹 Mixtral 8x22B：開源 AI 模型的重大突破

Mixtral 8x22B | 開源 AI | SMoE 架構 | 多語言 | 數學和編碼 | Apache 2.0

2024-04-18



在最近的一項突破中，Mistral AI 揭露了 Mixtral 8x22B，為開源人工智慧（AI）模型設定了新的標準。這個創新的模型不僅僅是一步，而是在 AI 的性能和效率上邁出了一大步。

Mixtral 8x22B 以其稀疏專家混合（SMoE）架構而著稱，使其成為一個擁有僅 390 億個活躍參數的 1410 億參數運作的強大力量。這意味著什麼？就像你能在圖書館中瞬間找到你需要的那本書，而不必翻遍每一個書架。這樣的效率使 Mixtral 8x22B 能夠更快、更準確地處理信息和做出決策。

但這還不是全部。這個模型是一個多語言高手，流利掌握包括英語、法語、意大利語、德語和西班牙語在內的多種主要語言。它不僅僅是談論語言；Mixtral 8x22B 也是一個在數學和編碼方面的神童，具有優越的技術能力。其「限制輸出模式」和對原生函數調用的支持，為大規模應用開發和技術進步開啟了新的視野。

憑藉 64K 代幣的上下文窗口，Mixtral 8x22B 在從廣泛文件中回憶精確信息方面表現卓越，對於處理大型數據集的企業來說，它是一個寶貴的工具。它在標準行業基準測試中的表現超越了許多現有模型，從常識和推理到編碼和數學的專業知識。

在高度寬鬆的 Apache 2.0 許可證下發布，Mixtral 8x22B 體現了 AI 社群合作和創新的精神，提供無限制的使用並鼓勵廣泛採用。

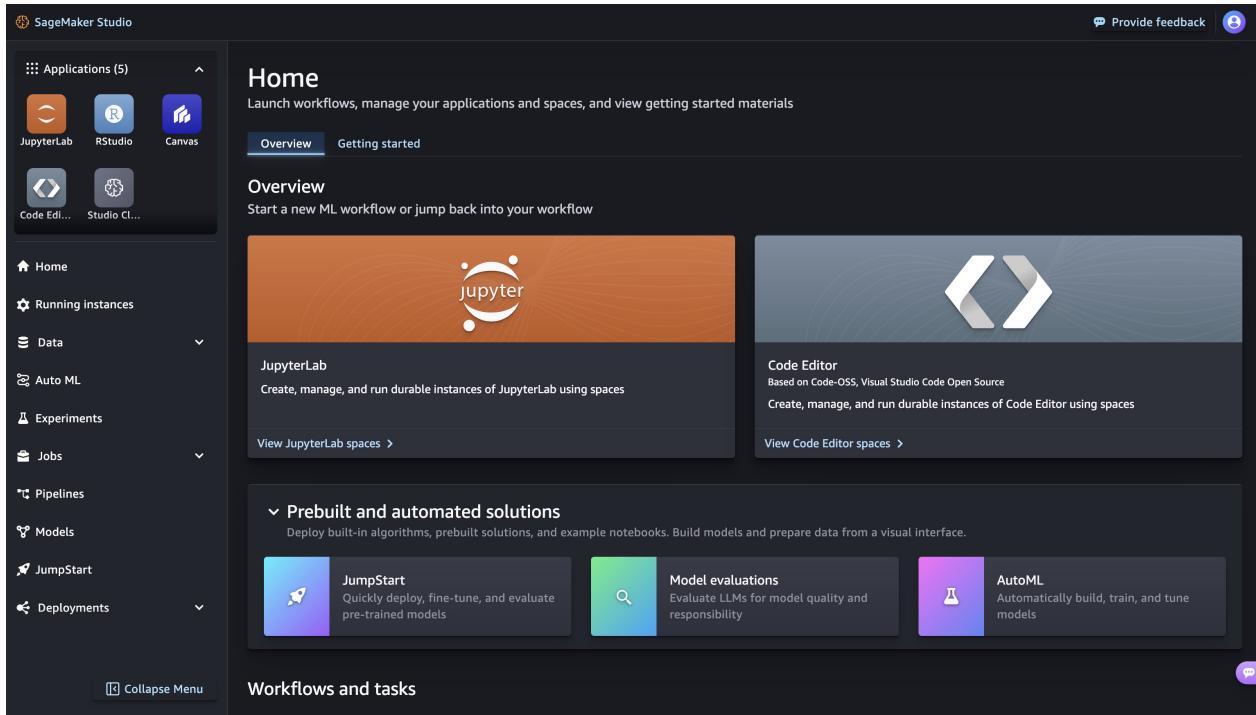
這個模型代表了先進 AI 工具民主化的一個重要里程碑，為開發者和企業開啟了新的可能性。 Mistral AI 的 Mixtral 8x22B 不僅僅是 AI 研究的一項成就；它是開源 AI 開發未來的一個指標。

[閱讀更多](#)

探索Amazon SageMaker JumpStart中 Meta Llama 3模型的力量

Meta Llama 3 | Amazon SageMaker JumpStart | 生成式文字模型 | 機器學習 | 人工智慧
SageMaker Pipelines | SageMaker Debugger

2024-04-18



在充滿科技進步的世界中，隨著Meta Llama 3基礎模型的介紹，一個新的篇章在Amazon SageMaker JumpStart中展開。這些模型旨在革新我們與生成式文字模型的互動方式，Llama 3作為創新的燈塔，提供一系列預訓練且微調良好的模型，可輕鬆部署並執行推理。

Meta Llama 3有何特別之處？

Meta Llama 3模型有兩種大小，分別為8B和70B，每種都支援8k的上下文長度。這種多樣性使得它能夠應用於從增強推論和代碼生成到改善遵循指令的能力等廣泛的應用領域。Llama 3的實力核心在於其僅解碼器的Transformer架構和創新的分詞器，確保了模型的卓越性能。

增強功能，為了更順暢的體驗

訓練後的改進顯著降低了錯誤拒絕率，同時也提升了對齊性並使模型反應更為多樣化。這不僅提升了使用者體驗，也為更準確和多樣化的輸出鋪平了道路。藉由使用Amazon SageMaker的特色功能，如SageMaker Pipelines和Debugger，部署Llama 3模型從未如此安全或高效。

SageMaker JumpStart：你的創新之門

SageMaker JumpStart讓發現和部署基礎模型的過程變得簡單明了。無論是透過Amazon SageMaker Studio還是透過SageMaker Python SDK以程式化方式，只需幾次點擊或一行代碼即可輕鬆存取Llama 3模型的強大功能。這種易於存取性，再加上AWS安全環境下的數據安全保障，使得部署更加令人放心。

文本生成的無限可能

想像一下，執行從回答複雜查詢到翻譯語言甚至生成創意文本的任務。Llama 3模型提供了這些以及更多功能，允許技術的精密與實用性的無縫結合。無論是進行對話、洞察見解還是編寫故事，潛在的應用範圍與你的想像力一樣無限。

總之，將Meta Llama 3模型整合至Amazon SageMaker JumpStart，標誌著機器學習與人工智慧旅程中的一個重要里程碑。這證明了當創新遇到易用性時會出現無窮的可能性，同時確保了安全性和效率。開始探索Llama 3模型的廣闊潛力，讓AI的力量轉變你的世界。

[閱讀更多](#)

NVIDIA 以 Meta Llama 3 加速人工智慧發展

NVIDIA | Meta Llama 3 | AI | 大型語言模型 | NVIDIA H100 Tensor Core GPU | NVIDIA NeMo | NVIDIA
TensorRT-LLM | NVIDIA Triton Inference Server | NVIDIA Jetson Orin | GeForce RTX GPU

2024-04-18



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

Watch On Demand

NVIDIA 以 Meta Llama 3 加速 AI 發展

NVIDIA 揭幕了在人工智慧 (AI) 領域的一大飛躍，透過提升 Meta Llama 3 這個尖端的大型語言模型 (LLM) 的性能。這次合作結合了 Meta 先進的 LLM 與 NVIDIA 強大的運算能力，為 AI 效率與速度設定了新的標準。

Meta Llama 3 在 24,576 NVIDIA H100 Tensor Core GPU 的驚人配置下進行訓練，展現了其開發背後的強大計算力。該模型旨在橫跨各種平台表現卓越，從雲服務和數據中心到邊緣設備和個人電腦，確保廣泛的可及性和多樣性。

對於希望深入生成式 AI 世界的開發者和企業，NVIDIA 提供了諸如 NVIDIA NeMo 之類的工具，用於以特定數據微調 Meta Llama 3，以及 NVIDIA TensorRT-LLM 和 NVIDIA Triton Inference Server 以實現最佳部署。這一生態系統支持創建符合獨特商業需求的定制 AI 模型，推動了 AI 可能性的界限。

此外，NVIDIA 強調了針對設備和個人電腦的優化重要性。通過利用 NVIDIA Jetson Orin 和 GeForce RTX GPU，Meta Llama 3 能夠在機器人和邊緣計算設備以及工作站和個人電腦上高效運行。這種廣泛的兼容性確保開發者擁有在各種應用中實現其 AI 構想所需的工具。

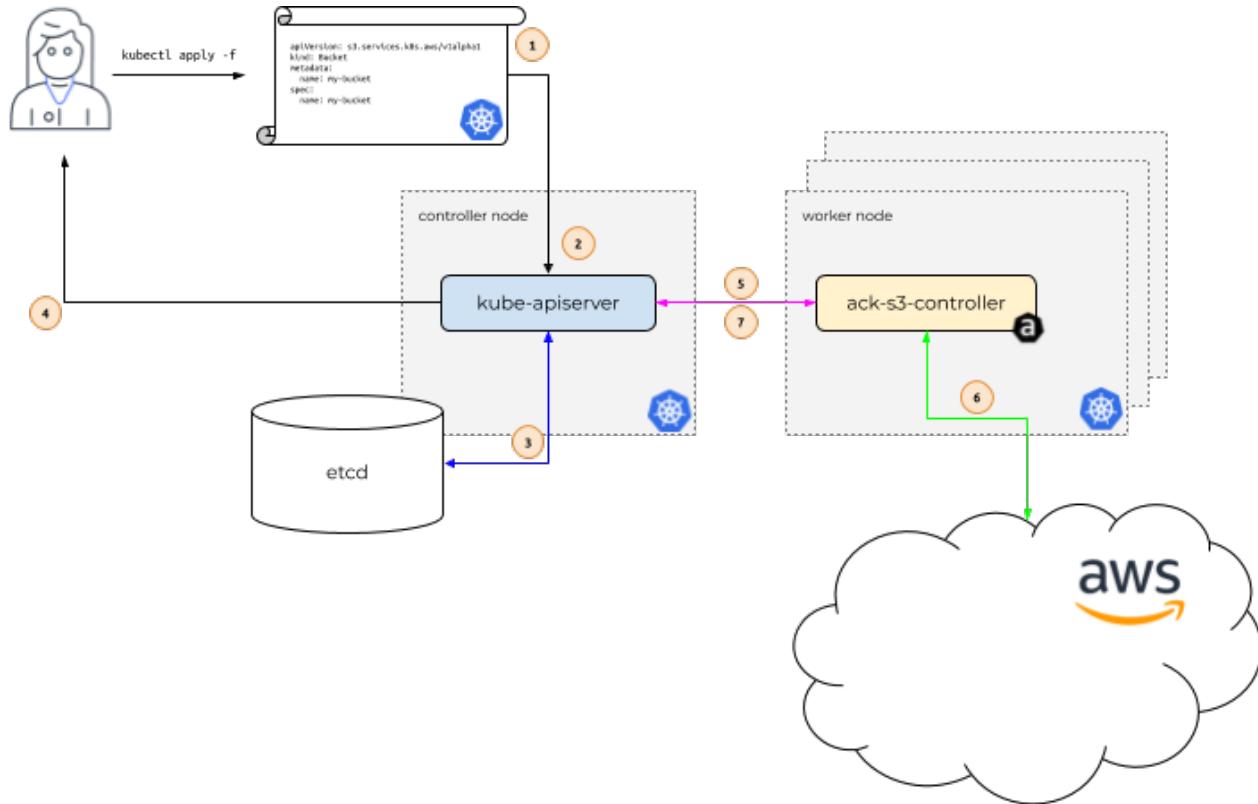
由 NVIDIA 的加速計算和 Meta 的 Llama 3 驅動的這個新 AI 時代，承諾將改變我們與技術的互動方式，使 AI 比以往任何時候都更加可及、高效和強大。

[閱讀更多](#)

革新機器學習部署：Amazon SageMaker 遇上 Kubernetes

Amazon SageMaker **Kubernetes** 機器學習模型部署 成本削減 大型語言模型 推理能力 資源利用率

2024-04-19



顛覆機器學習部署：Amazon SageMaker 與 Kubernetes 的結合

在一次改變遊戲規則的更新中，Amazon 推出了一種創新的機器學習模型部署方式，承諾將大幅削減成本。透過利用 Kubernetes Operators 的力量，Amazon SageMaker 現在提供了新的推理能力，旨在將大型語言模型 (LLMs) 的部署成本平均降低 50%。

這些增強功能是通過 Amazon SageMaker Operators for Kubernetes 實現的，它使用了 AWS Controllers for Kubernetes (ACK) 框架。這種設置允許直接透過 Kubernetes API 順暢地創建 AWS 資源，將 Amazon SageMaker 的強大功能更接近 Kubernetes 使用者。

最新版本 v1.2.9 引入了對推理組件的支持，使得能夠在 Amazon SageMaker 端點上部署基礎模型，並對資源分配進行精細控制。這不僅優化了資源利用率，還能根據不同用例有效地擴展端點，最終平均將模型部署成本降低一半。

該更新使使用者能夠在單一端點上部署多個基礎模型，根據需要調整加速器和記憶體的分配。這種靈活性提高了性能，降低了延遲，並確保了機器學習操作的成本效益擴展。

對於那些已經使用 Kubernetes 作為其控制平面的人來說，這一整合帶來了部署 SageMaker 推理組件的額外優勢，豐富了 Kubernetes 體驗，增添了強大的機器學習能力。

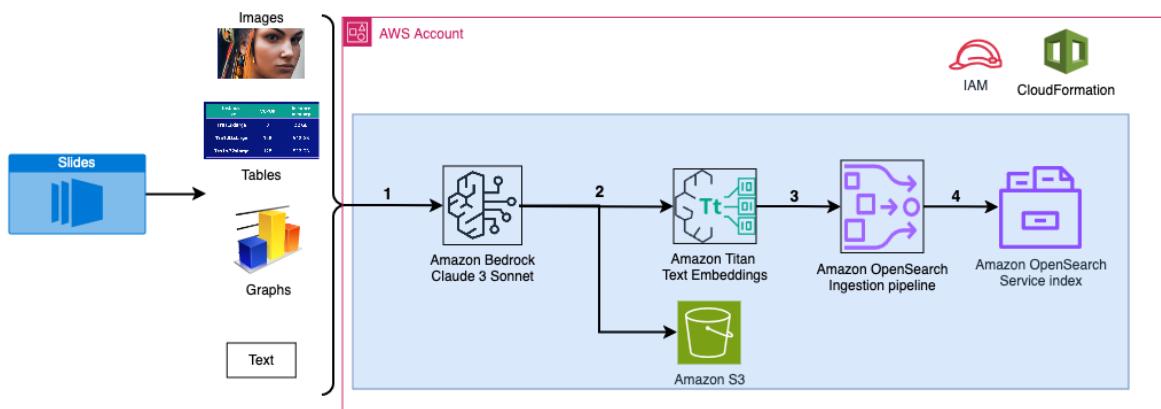
Amazon SageMaker 的這一戰略舉措標誌著向著讓機器學習更加易於使用和負擔得起的方向邁出了一大步，邀請使用者探索新的推理能力，並親身體驗成本優勢。

[閱讀更多](#)

利用 Amazon Bedrock 先進 AI 革新您的簡報方式

Amazon Bedrock | **AWS 機器學習** | **幻燈片簡報** | **多模態基礎模型** | **Anthropic Claude 3 Sonnet**
Amazon Titan Text Embeddings | **文字嵌入** | **檢索增強生成**

2024-04-19



用 Amazon Bedrock 的先進 AI 革新您的簡報

在創新的一大步中，Amazon Web Services (AWS) 展示了一種轉型的方法來透過先進的機器學習增強幻燈片簡報。利用寄存於 Amazon Bedrock 上的多模態基礎模型的力量，使用者現在可以以前所未有的方式與他們的幻燈片互動，讓檢索資訊就像提問一樣簡單。

這個改變遊戲規則的解決方案採用了 Anthropic Claude 3 Sonnet 模型以及 Amazon Titan Text Embeddings 模型來分析幻燈片內容。通過為每張幻燈片生成詳細的文字描述並將這些轉換成文字嵌入，系統創建了一個豐富的、可搜索的您簡報內容的資料庫。無論您的幻燈片包含密集的文字、錯綜複雜的圖表，還是複雜的圖像，這種方法確保通過一個簡單的查詢就能訪問到每一個細節。

當提出一個問題時，系統無縫地從資料庫中識別出最相關的幻燈片，並利用 Claude 3 Sonnet 模型來製作一個簡潔、有信息量的回答。這個過程不僅加深了對所呈現材料的理解，同時也通過使得特定資訊的定位變得更加容易，從而使簡報的準備和交付變得更加高效。

這種方法設計簡單，將檢索增強生成 (Retrieval Augmented Generation, RAG) 的能力擴展到包含視覺元素，橋接了文字數據和視覺簡報豐富情境之間的差距。無論是用於內部會議、客戶簡報，還是學術講座，這項技術都承諾將改變我們與幻燈片互動並從中獲益的方式。

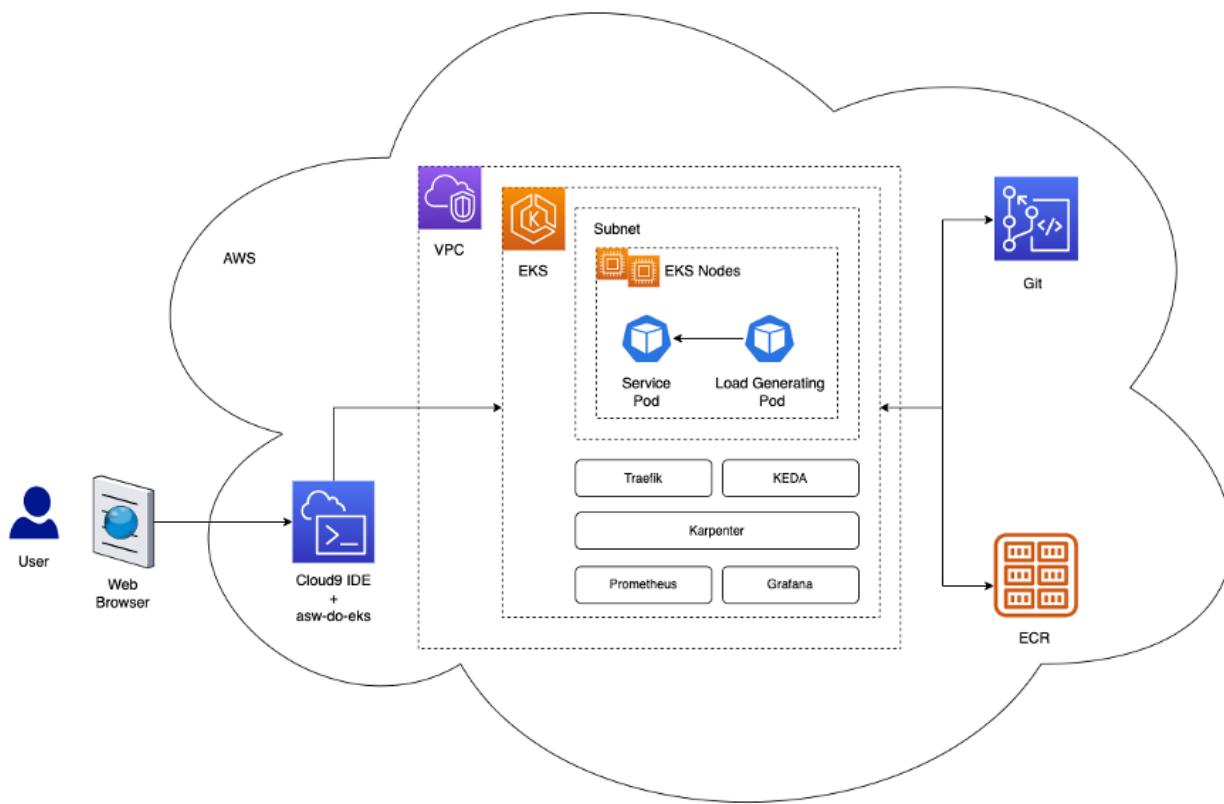
請繼續關注這一激動人心的系列進展，因為 AWS 繼續推動機器學習和人工智能的界限，使複雜的技術對每個人都變得可訪問。

閱讀更多

利用 AI 加速藥物發現：Iambic Therapeutics 的重大突破

AI | 藥物發現 | Iambic Therapeutics | 生成式 AI | 預測式 AI | Amazon EKS | Karpenter | 雲計算

2024-04-19



在一項劃時代的發展中，新興的藥物發現初創公司 Iambic Therapeutics 正在革新我們對於癌症治療藥物創建方法的看法。利用先進的 AI 技術，他們設計了一種方法，能夠快速且有效地篩選數百萬潛在的藥物分子。這項創新設定了將生命挽救的治療快速交付給需要的患者的快軌。

通過利用生成式和預測式 AI 的力量，Iambic 的平台可以探索廣闊的化學空間，開啟一個藥物發現更快更有效的新時代。他們的整合平台將 AI 軟體與基於雲的數據和可擴展的計算基礎設施結合在一起，並通過高通量的化學和生物學功能進行增強。這種協同作用不僅為 AI 提供了寶貴的數據，還實現了自動化的決策制定和處理，使整個藥物開發過程變得更加高效。

他們成功的一個顯著證明是，他們的項目從開始到臨床候選品在短短 24 個月的時間框架內加速，為業界樹立了新的標杆。這一壯舉是通過使用 Amazon Elastic Kubernetes Service (EKS) 和 Karpenter 實現的，這允許對於他們發現平台至關重要的可擴展 AI 訓練和推理。

在 Amazon EKS 上部署 Karpenter 提供了一種解決方案，動態調整計算資源，確保 AI 模型能夠訪問必要的 GPU 力量進行實時處理和分析。這種方法不僅優化了計算效率，還大大降低了成本，展示了一個可擴展的模型，用於未來的 AI 驅動藥物發現努力。

通過推動 AI 在製藥領域可能性的界限，Iambic Therapeutics 正在引領一條將更好、更有效的藥物比以往任何時候都更快地推向市場的道路。他們的工作體現了將人工智能與雲計算技術整合的轉型潛力，為藥物發現創新設定了新的標準。

[閱讀更多](#)

利用 AWS 自動訓練實現個人化的革命

AWS | Amazon Personalize | 自動訓練 | 機器學習 | 個人化推薦 | 數據處理 | 特徵識別 | 算法選擇 | 模型訓練 | 數據隱私 | 安全

2024-04-20

Step 1
Specify solution details

Step 2
Training configuration

Step 3
Review and create

Training configuration

Customize the solution to meet your specific business requirements. You can't change these configurations after you create your solution.

▼ Automatic training - new Info

Configure whether the solution automatically creates new solution versions.

Turn on - recommended

Automatically create new solution versions at a configurable frequency. This removes the manual training required for the solution to learn from your most recent data. The first training starts within an hour after you create this solution

Turn off

Manually control solution version creation at a frequency that aligns with your use case and how often you import new data.

Training frequency

Specify the training frequency in days. For example, it can create a solution version every 5 days. Training starts when you create the solution. With each solution version, you incur training costs.

7

Minimum of 1 day. Maximum of 30 days. Default set to 7 days.

► Columns for training - optional

► Hyperparameter configuration - optional

► Additional configuration - optional

Cancel

Previous

Next

在一項激動人心的發展中，Amazon Web Services (AWS) 為 Amazon Personalize 推出了一項突破性功能，旨在讓機器學習 (ML) 驅動的個人化變得更加動態和有效。這個新能力，被稱為自動訓練，承諾保持你的個性化推薦與用戶不斷變化的偏好和行為完美同步，確保你的服務隨著時間保持相關性和吸引力。

自動訓練允許 Amazon Personalize 根據最新數據自動更新和精煉其模型。這意味著，隨著你的用戶與你的內容、產品或服務進行互動，Amazon Personalize 從這些互動中學習，調整其推薦以更好地匹配當前趨勢和用戶偏好。它就像擁有一個自我改進的引擎，確保你的推薦不會過時。

這是它的工作原理：一旦你啟用自動訓練，Amazon Personalize 就會負擔起重任，根據你指定的頻率（從每日到每月）重新訓練其推薦模型。這個過程納入了最新的用戶數據，確保推薦始終與最新的用戶行為和項目更新保持一致。

這個系統的美在於其簡單性。Amazon Personalize 自動化了整個 ML 流程 - 從數據處理和特徵識別到算法選擇和模型訓練。這意味著即使是沒有深度 ML 專業知識的人也可以部署複雜的個性化策略，增強用戶參與度和滿意度。

此外，Amazon Personalize 通過加密來確保你的數據的隱私和安全，讓你在為你的觀眾提供定制體驗時安心。

無論你是將個性化推薦整合到你的網站、移動應用程序還是電子郵件營銷活動中，Amazon Personalize 配備自動訓練提供了一種強大的方式來適應你用戶的不斷變化的需求，保持你的服務新鮮和相關。

今天就擁抱個性化的未來，與 AWS 一起，讓每一次用戶互動都計算在內。

[閱讀更多](#)

03 資訊安全

英美聯手在AI安全領域取得重大進展

AI安全 英國 美國 跨國合作 技術標準 研究所 OpenAI

2024-04-02

英國與美國在AI安全上的重大飛躍

在前所未有的舉動中，英國與美國聯手，共同提升AI安全標準。這次合作標誌著英國科技大臣 Michelle Donelan與美國商務部長Gina Raimondo之間簽署了一份諒解備忘錄，旨在為下一代AI技術建立嚴格的測試程序。

這項協議為英國新成立的AI安全研究所和即將出現的美國對應機構交換專長鋪平了道路，以期預先解決人工智慧可能帶來的潛在危險。這種夥伴關係不僅僅是關於知識共享；它還關於對抗私人開發AI模型所帶來的風險，包括來自OpenAI等行業巨頭的模型。

這對日常人意味著什麼？想像一下一個AI更加融入我們日常生活的世界，在從駕駛到網上購物等一切方面都變得更加順暢和安全。然而，強大的力量伴隨著巨大的責任，這個英美倡議全都是為了確保推動未來創新的AI能夠保持安全、可靠並且對各方都有益。

通過聯合測試活動和共享研究，兩國都在承諾創造一個更安全的AI未來。這項協定不只是為了防止潛在的AI事故；它是邁向全球AI安全方法的基礎步驟，確保曾經看似科幻的技術能夠以最小的風險和最大的益處融入我們的世界。

本質上，這個英美協議不僅僅是跨越大西洋的握手；它是一個承諾，為那些增強而非危害我們生活方式的AI技術鋪路。

[閱讀更多](#)

深偽困境中的航向：保護企業於數位幻象之中

深偽 AI 網絡安全 身份驗證 立法

2024-04-03

深偽困境中的航向：保護企業於數位幻象之中

在一個看見不再代表相信的世界裡，由AI生成的深偽影像（deepfakes）威脅著企業的大門，不僅威脅到企業資產，也威脅到信任與安全。近期，有跨國公司就因深偽影片中的高管被仿冒，而被騙走2560萬美元，這鮮明地提醒我們，利用這項技術進行惡意活動的網絡罪犯之精密程度。

深偽影像，這種數位變色龍，透過AI的幫助，令人聲音和臉龐變得逼真無比，可以假扮任何人，從CEO到總統。它們的應用範圍從無害娛樂到可能動搖政治的錯誤資訊，以及複雜的財務詐騙。网络安全專家強調，識別並對抗這些AI騙局的迫切性，這些騙局已經從基本的釣魚郵件演變至包含逼真度極高的影片、圖像和音頻。

因應此，像是美國的聯邦通信委員會和聯邦貿易委員會這樣的監管機構正在加強力度，制定法律以限制未經授權使用AI模仿聲音和冒充實體。然而，僅靠立法行動是不夠的。公司必須挺身而出，教育其勞動力關於深偽的危險，升級釣魚防禦以識別AI生成的詐騙，並增強身份驗證過程。

這是一場與時間和技術的賽跑。隨著AI持續進化，創建和檢測深偽將只會變得更加困難。企業必須保持警惕，採取包括法律、技術和教育策略在內的多面向方法，以防範深偽威脅。這樣做，他們不僅保護了自己的財務利益，也保護了品牌的完整性以及員工和客戶的信任。

保持資訊更新，確保安全。

[閱讀更多](#)

航向數位假訊息的新時代

假訊息 | AI | 選舉 | 數位戰爭 | 民主 | 網絡安全

2024-04-05

在Microsoft的威脅情報團隊近期的一項揭露中，來自中國的一種令人擔憂的策略浮出水面：利用先進的人工智能生成的假訊息運動，目的是為了影響2024年幾場關鍵選舉。焦點集中在美國、南韓和印度的總統及立法選舉，這些地方的國家支持的中國網絡組織，可能會與北韓行動者合作，部署AI創建的內容，以偏向他們的地緣政治議程來誘導公眾輿論。

這種新型別的網絡操縱利用人工智能來創造和散播假訊息通過社交媒體，包括從迷因和影片到音頻剪輯的一切，都旨在誤導和操縱公眾觀念。一個顯著的例子是在台灣1月總統選舉期間觀察到的「乾演練」，其中AI合成的內容虛假地呈現政治背書和指控，標誌著數位戰爭策略的一次重大演進。

部署AI渲染的「新聞主播」和捏造的背書凸顯了對旨在破壞民主選舉完整性的技術的日益投資。雖然目前對公眾輿論的影響被評估為低，但Microsoft的洞察表明一個策略性的長期威脅，可能會日益挑戰民主對話和全球選舉完整性的基礎。

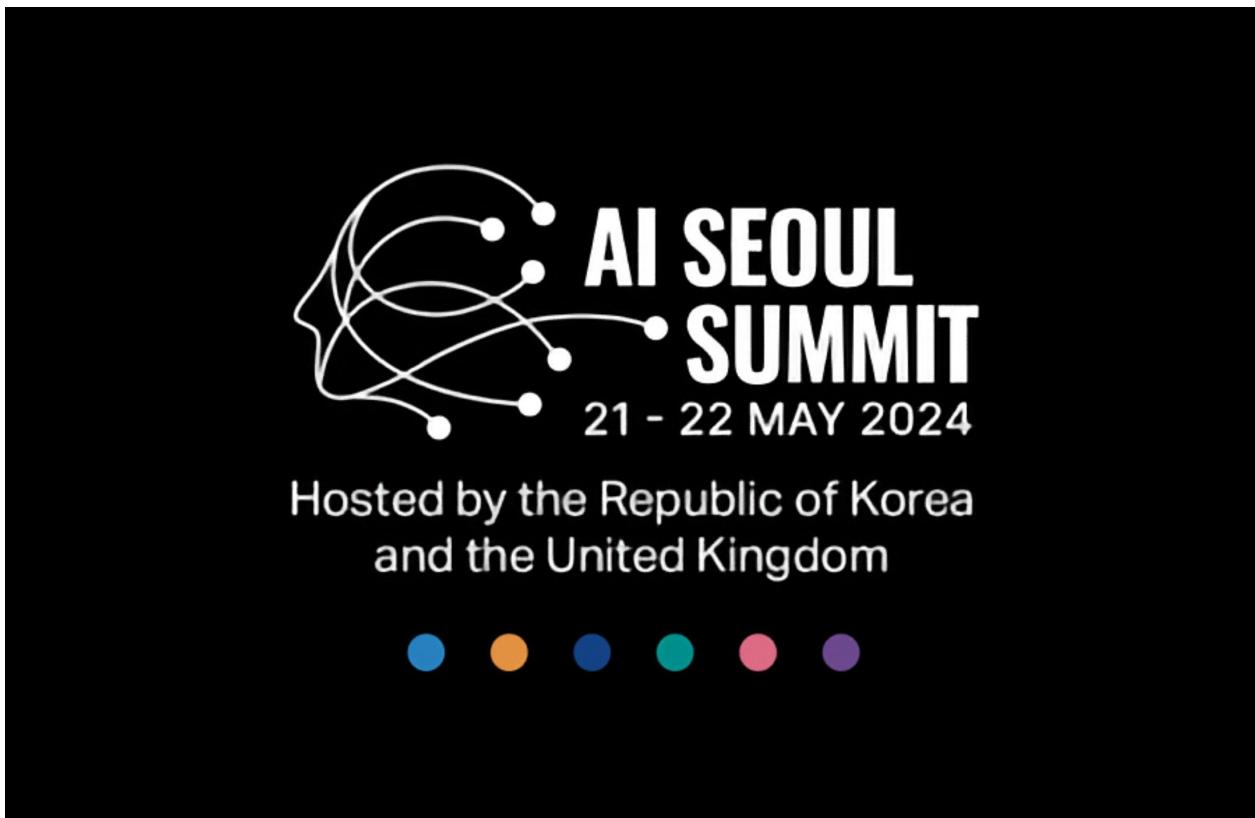
當我們站在這數位懸崖的邊緣時，全球社會必須保持警惕，促進網絡安全和數位素養的發展，以捍衛民主的基石：自由和公正的選舉。

[閱讀更多](#)

AI首爾峰會：AI安全與創新的全球對話

AI首爾峰會 | AI安全 | AI治理 | 國際合作 | Yoshua Bengio | 科學報告

2024-04-12



在一個開創性的事件中，定於5月21日和22日，英國與南韓將聯手主辦AI首爾峰會。這場峰會不僅僅是一次會議；它是對人工智慧（AI）未來的一個願景——目標是創造一個AI技術被安全負責地開發的世界。

在AI安全、包容性及創新的主題下，這場峰會是歷史性的布萊切利公園討論所啟動努力的延續。在英國首相Rishi Sunak和南韓總統Yoon Suk Yeol的領導下，峰會承諾將成為AI治理的一個關鍵時刻。它將以一場虛擬領導人會議開場，隨後是一場數位部長們的面對面會議，展示了在利用AI轉型潛力的同時，防範其風險方面的團結一致。

什麼讓AI首爾峰會脫穎而出？那就是對可行見解的承諾。峰會期間，除了討論之外，將發布由圖靈獎得主Yoshua Bengio領導的關於先進AI安全的國際科學報告。這份報告旨在整合關於AI安全的最佳科學研究，標誌著峰會不僅僅是一個討論論壇，而是一個朝向安全AI發展的全球行動催化劑。

隨著AI重塑產業和社會，AI首爾峰會體現了一項共同努力，以確保這項強大技術推動人類福祉、包容性和繁榮的進步。這是對國際合作和廣泛聲音能夠引導AI走向一個造福所有人未來的信念的一份證明。

使用 AWS 的尖端 Nitro 系統確保生成式 AI 工作流的安全

AWS | Nitro 系統 | 生成式 AI | 安全 | 加密 | Amazon EC2 | ML 加速器 | GPU | NVIDIA
Blackwell 架構

2024-04-16



在人工智慧 (AI) 迅速演進的領域中，AWS 正在開創安全的生成式 AI 應用，確保寶貴的數據保持受保護。AWS 了解到保密的重要性，特別是當涉及到個人和財務數據等敏感信息時，因此正在加強其 Nitro 系統的應用。

Nitro 系統是 AWS 運算的核心，以安全和性能為其核心設計。它保證沒有人，甚至 AWS 員工，可以訪問您在 Amazon Elastic Compute Cloud (Amazon EC2) 實例上的數據或操作。自 2017 年以來運作的這一高級別保護意味著您的敏感數據、模型權重和由 AI 驅動的洞察保持遠離外部訪問或漏洞。

此外，Nitro 系統引入了一種創新方法，透過其 Elastic Fabric Adapter (EFA) 和 AWS Scalable Reliable Datagram (SRD) 協議，在雲規模的大規模分散式訓練中確保安全通訊。這一設置確保您的數據在傳輸過程中保持加密，同時不影響性能。

AWS 也在增強其產品，計劃包括使用 ML 加速器和 GPU 的端到端加密工作流程，提供更強大的安全層。這一升級將支持最新的 AI 模型和複雜的處理需求，同時維持您的 AI 數據和操作的嚴格保密性。

隨著 AWS 和 NVIDIA 緊密合作，將這些先進的安全特性與 NVIDIA 即將推出的 Blackwell 架構整合，AWS 客戶可以期待一個未來，其中生成式 AI 應用不僅具有變革性，而且在 AWS 的領先基礎設施中安全地錨定。

AWS 通過在生成式 AI 工作流程的所有層面上優先考慮安全性，確保您的創新 AI 項目不僅具有突破性，而且對未經授權的訪問有所防護，讓您可以自信地專注於創建和擴展您的 AI 解決方案。

[閱讀更多](#)

在人工智慧領域，資料隱私和安全成為主要關切

資料隱私 | 安全性 | 生成式AI | 法規遵守 | 技術整合

2024-04-17



在不斷演進的人工智慧（AI）領域中，最近的一項研究揭示了AI決策者之間的一個重大擔憂：資料隱私和安全性。有高達80%的塑造策略和在其組織中使用AI技術的人士對這些關鍵問題發出了警報。

這種擔憂源於組織在嘗試利用生成式AI提升生產力和創新力時面臨的一系列挑戰。儘管對AI能夠實現的成就感到熱情，但缺乏戰略規劃和技術人才短缺是顯著的障礙。此外，將生成式AI整合到現有系統中以及預測相關成本帶來了額外的複雜性。

由Coleman Parkes Research進行並由SAS贊助的研究突顯出當前做法中令人擔憂的差距：只有少數組織採取措施評估大型語言模型（LLMs）中的偏見和隱私風險。這一差距因缺乏生成式AI的全面治理框架而加劇，使許多企業面臨著未能遵守法規的風險。

為了應對這些挑戰，建議組織不應將生成式AI視為快速解決方案，而應將其視為增強現有流程的工具。這需要一種深思熟慮的方法，強調制定堅實的策略、技術整合和AI模型的可解釋性。

這些發現強調了創建不僅推動價值，而且以可持續和可擴展的方式解決人類需求的實際AI應用的重要性。隨著AI技術的不斷發展，要在這個競爭激烈的環境中保持領先，將取決於組織能夠如何深思熟慮且有效地適應和回應這些新興的擔憂。

[閱讀更多](#)

04 應用

深入探索 AWS 與 Hugging Face 在北美的 AI 未來巡迴演講

AWS | Hugging Face | 生成式 AI | 巡迴演講 | Amazon SageMaker | AWS Trainium | AWS Inferentia
開源模型

2024-04-02

深入了解 AWS 與 Hugging Face 在北美巡迴演講中的 AI 未來

對於科技愛好者和開發人員來說，有一則令人興奮的更新消息：AWS 正與 Hugging Face 聯手，在北美帶來一場令人振奮的生成式 AI 巡迴演講。這次合作旨在加速生成式 AI 之旅，讓開發者以更快速、更經濟的方式，將他們的 AI 應用帶入現實。

對那些尚未瞭解的人來說，Hugging Face 是 AI 領域的巨人，擁有超過 500,000 個開源模型和 100,000 個以上的數據集。這次合作承諾將透過使用 Amazon SageMaker、AWS Trainium 和 AWS Inferentia，簡化這些模型的訓練、微調和部署過程。這意味著開發者現在可以創建性能更佳、成本更低的生成式 AI 應用。

這場巡迴演講是您與 Hugging Face 和 AWS 專家會面、學習最新生成式 AI、以及親手微調和部署基礎模型的絕佳機會。Hugging Face 的首席傳道士 Julien Simon 將從四月至六月，巡訪八個 AWS 總部，深入探討特定用例，並展示 AWS 與 Hugging Face 如何助您的項目一飛沖天。

巡迴城市包括西雅圖、三藩市、聖克拉拉、洛杉磯、波士頓、紐約市、奧斯汀以及華盛頓特區的阿靈頓。與會者有兩種精彩選擇：請求與專家進行一對一會面，或在西雅圖、聖克拉拉、紐約市或奧斯汀註冊參加 hands-on 開發者工作坊。

這場巡迴演講不僅是交流和學習的機會，更是站在生成式 AI 革命前沿的機會，瞭解如何高效部署開源模型，同時降低成本。不要錯過成為 AI 未來一部分的機會。欲瞭解更多詳情或註冊，請聯繫 AMER Hugging Face 巡迴演講的組織者。

我們迫不及待想在那裡見到您，並一起探索 AWS 與 Hugging Face 合作為 AI 世界帶來的無限可能性！

[閱讀更多](#)

簡化機器學習存取：Amazon SageMaker Canvas 與 AWS IAM Identity Center

[Amazon SageMaker Canvas](#) [AWS IAM Identity Center](#) [機器學習](#) [AutoML](#) [單一登入](#) [使用者驗證](#) [群組權限控制](#)

2024-04-02

在機器學習 (ML) 的領域中，Amazon SageMaker Canvas 正在打破障礙，允許使用者不需撰寫一行代碼即可生成預測並管理ML工作流程。從數據準備到部署ML模型，SageMaker Canvas 是您全方位ML體驗的首選，包括 AutoML、管理型端點，甚至配置生成式AI的基礎模型。

但是存取性和安全性呢？這就是 AWS IAM Identity Center 登場的時刻，使生活變得更加輕鬆和安全。通過整合單一登入 (SSO) 功能，使用者現在可以僅憑其 IAM Identity Center 憑證，就能直接進入 SageMaker Canvas，省去了通過 AWS Management Console 導航的常見障礙。

過程很簡單。首先，啟用 IAM Identity Center 並將其與 SageMaker Studio 域連接以進行使用者驗證。然後，在 IAM Identity Center 中創建或整合使用者和群組，並將它們分配到 SageMaker Studio 域。瞧！您的使用者現在可以直接憑藉其憑證訪問 SageMaker Canvas，開始他們的ML之旅，同時維護一個安全的環境。

有三種存取選項可供選擇，迎合不同的需求。無論是通過 IAM Identity Center 直接存取 SageMaker Canvas、通過 SageMaker Studio 導航，還是使用直接連結立即訪問 Canvas，靈活性隨手可得。

這種方法不僅通過利用群組進行權限控制來簡化用戶管理，還符合雲端工程中要求的嚴格安全標準，同時不影響開發團隊至關重要的敏捷性和獨立性。

所以，就這樣——一個無縫、安全且簡化的進入無代碼ML世界的途徑，通過 Amazon SageMaker Canvas 和 AWS IAM Identity Center，ML 的未來實際上只是一點之遙。

[閱讀更多](#)

Google 利用 AI 為開發中國家鋪路

Google 開發中國家 AI Sprinters 人工智能 技能訓練 經濟發展 雲計算 數據系統

2024-04-03



Google 以 AI 為開發中國家鋪路

在一項開創性的舉措中，Google 發布了一份題為「AI Sprinters」的報告，為開發中國家利用人工智能（AI）的潛力提供了一個策略性藍圖。同時，Google 透過 Google.org 承諾提供1500萬美元，特別是針對服務不足的社群，來加強 AI 技能訓練。

AI 技術在解決開發中國家面臨的一些最迫切挑戰方面發揮了重要作用，這些挑戰遍及拉丁美洲、中東、亞洲和非洲。從優化里約熱內盧的交通到協助非洲農民早期發現蝗蟲爆發，AI 的多功能性正在證明它是一個遊戲改變者。此外，它在擴展業務和提升醫療保健成果方面的應用，預示著一個更光明、更高效的未來。

然而，從潛力轉化為實際成長需要政府、私營部門和民間社會的共同努力。「AI Sprinters」報告強調了進行策略性投資和制定政策以推動經濟發展的必要性。主要建議包括擁抱雲計算、提升工人的 AI 技能、現代化國家數據系統，以及培養對 AI 友好的規範。

這項倡議在一個關鍵時刻到來，當全球金融領袖為世界銀行/國際貨幣基金組織春季會議聚集在一起，敦促他們將 AI 納入國家發展戰略。Google 的遠見不僅限於即時的技術進步，旨在橋接數字鴻溝，並在開發中國家顯著改善數百萬人的生活。

這個關鍵時刻可能標誌著一個新時代的開始，其中 AI 不僅僅是一項技術進步，而是全球經濟和社會發展的一個基本支柱。

[閱讀更多](#)

以創新照亮道路：Dotlumen 在輔助技術上的飛躍

Dotlumen | 視障 | **AI眼鏡** | 人工智慧 | 輔助技術 | **NVIDIA** | 流動性 | 觸覺回饋

2024-04-03

在使世界對視障者更易於導航的征途中，Dotlumen 引進了一項突破性解決方案：Dotlumen 眼鏡。正如在 NVIDIA 的 AI Podcast 上揭露的那樣，Dotlumen 的 CEO Cornel Amariei 分享了這些 AI 驅動的眼鏡的見解，這些眼鏡旨在為視覺障礙者的流動性帶來革命性的變革。

Dotlumen 眼鏡配備了先進的感應器，並利用人工智慧來規劃出安全的行走路徑。它們透過觸覺回饋與使用者溝通，這意味著它們使用溫和的振動來引導佩戴者安全地繞過障礙物並通過他們的環境。

這種創新的方法不僅承諾將增強視障者的獨立性和自信，還代表了在輔助技術方面的重大飛躍。作為 NVIDIA Inception 計劃的一部分，Dotlumen 站在使用 AI 進行社會影響的最前沿，展示了技術轉變生活的巨大潛力。

在挑戰和技術突破之中，Dotlumen 的旅程體現了結合 AI 與同理心和創新來解決現實世界問題的力量。隨著我們期待 Dotlumen 眼鏡的發布，這項技術為使世界對每個人都更加可及的道路前進照亮了方向。

[閱讀更多](#)

在數位時代，資料品質對於GenAI在行銷領域的關鍵作用

生成式人工智能 | 行銷 | 資料品質 | 個人化 | 客戶資料基礎

2024-04-04

在今天的數位時代，生成式人工智能（Generative AI，簡稱GenAI）正在革新行銷領域，資料的品質從未像今天這般重要。最近的一項調查顯示GenAI在行銷領域的未來光明，有70%的行銷長（CMOs）已經在使用它並探索個人化、內容創建和市場細分等領域。然而，期望與實際之間仍存在顯著差距，主要是由於底層資料品質的問題。

資料品質不佳可能導致令人失望的AI驅動的行銷體驗。想像一下，你期待一個個人化的購物體驗，最終卻因為AI沒有關於你與品牌間歷史的清晰、統一的視角而獲得不相關的建議。這種情況鮮明地提醒我們，沒有高品質的資料，即使是最先進的AI也無法交付它所承諾的奇蹟。

另一方面，當AI由準確、全面的資料驅動時，結果可能令人驚訝。個人化達到新高度，操作效率提高，行銷活動變得更加有效，同時降低了計算成本。高品質的資料使AI能夠創建高度個人化和方便的購物體驗，將一次性買家轉化為忠誠客戶。

解鎖這種潛力的關鍵在於建立統一的客戶資料基礎。傳統的資料統一方法往往難以達到預期效果，但AI模型提供了一個解決方案，通過智能地連接多個渠道和接觸點的資料點。這種方法結果形成了一個詳細的客戶檔案，為更有效的AI驅動行銷策略提供燃料。

隨著GenAI持續發展，向其提供高品質資料的重要性不言而喻。對行銷人員而言，這意味著要優先考慮資料的品質和全面性，以發揮AI的全部潛力，確保個人化行銷在大規模實現成為現實，而不僅僅是一個承諾。

[閱讀更多](#)

以 Amazon Personalize 革新新聞消費方式

Amazon Personalize 新聞推薦 機器學習 個性化 實時數據

2024-04-04

以 Amazon Personalize 革新新聞消費方式

在今日數位時代，每日產生的新聞內容量龐大，想要尋找對您而言重要的新聞，就如同大海撈針。為了解決這個挑戰，Amazon 推出了一項創新解決方案——Amazon Personalize，這將改變我們發現和消費新聞的方式。

Amazon Personalize 允許創建一個根據個人偏好量身定制的新聞推薦應用，克服了用戶興趣多樣化、閱讀歷史有限和新聞週期不斷變化等障礙。它利用先進的機器學習算法分析每位用戶的參與度，找出最與他們相關的新聞故事和來源。值得注意的是，它設有一個「當下趨勢」的配方，用於識別實時趨勢和熱門新聞，確保用戶也能接觸到及時和廣泛感興趣的內容。

該技術設計用於實時工作，提供個性化新聞摘要，適應不斷變化的興趣，並突出顯示趨勢新聞。對於那些在外的人來說，它甚至可以編譯個性化新聞摘要，直接通過電子郵件發送，確保您總是以最小的努力保持在循環中。

在幕後，Amazon Personalize 擁有一個複雜的架構，能夠處理歷史和實時數據。它巧妙地平衡了個性化與內容的新鮮度，加入新發布的文章。這個系統受益於 Amazon 的雲基礎設施，使其能夠無縫擴展以滿足需求，並維護一個響應式、引人入勝的用戶體驗。

通過利用 Amazon Personalize，新聞平台現在可以提供一項服務，這項服務不僅讓讀者保持知情，而且以一種獨特定制於每個人的方式進行，使得新聞消費比以往任何時候都更加相關、高效和愉悅。

在繁忙的新聞世界中，Amazon Personalize 可能是不被信息過載壓垮、保持知情的關鍵，真正是資訊過載時代創新的一盞明燈。

[閱讀更多](#)

Gramener利用Amazon SageMaker對抗都市熱島效應

Amazon SageMaker 都市熱島效應 GeoBox 地理空間數據 機器學習 環境挑戰 城市規劃

2024-04-05

Gramener利用Amazon SageMaker對抗都市熱島效應

在應對環境挑戰方面邁出了令人印象深刻的一步，Gramener利用了Amazon SageMaker的地理空間能力，開發了先進的解決方案，用於理解和緩解都市熱島(UHIs)。UHIs是指都市區域比起周圍鄉村地區顯著溫暖的地區，導致了許多環境和健康問題，從能源消耗增加到熱相關疾病的增加。

Gramener的創新工具GeoBox讓使用者能夠輕鬆分析公共地理空間數據，提供了一種轉型的方法來識別和處理UHI熱點。通過將複雜的空間數據轉換成容易理解的洞見，GeoBox促進了有效的UHI緩解策略的實施，為可持續的城市發展鋪平了道路。

通過全面分析衛星圖像和戰略性應用機器學習模型，Gramener的解決方案深入探討了導致UHIs的因素。這種方法不僅能夠以驚人的準確度預測地表溫度，還有助於規劃更綠色、更涼爽的城市空間。通過整合基礎設施和人口數據，GeoBox模型提供了一個細緻的視角，展示了城市規劃如何影響UHI效應，使得能夠做出支持氣候適應努力的明智決策。

這個解決方案的核心在於其快速處理大量數據的能力，將分析時間從數週縮短到僅幾小時。這種效率對於城市規劃者和環保人士來說是一個遊戲改變者，提供了迅速的洞見，支持積極的城市和環境規劃。

Gramener的開創性工作展示了結合地理空間分析與機器學習來有效對抗和緩解城市環境挑戰的潛力。通過利用Amazon SageMaker的強大功能，Gramener不僅僅是在預測都市熱島的未來，而是積極塑造一個更涼爽、更可持續的城市未來。

[閱讀更多](#)

🌟 深入機器學習的世界，參加 ML 奧林匹亞競賽！🌟

ML 奧林匹亞競賽 | **機器學習** | **Kaggle** | **健康護理** | **可持續性** | **電腦視覺** | **生物信號 ML 模型** | **預測地震損害** | **預報天氣條件** | **識別線上有害語言** | **Google for Developers**

2024-04-08

🌟 深入機器學習世界，參與 ML 奧林匹亞競賽！🌟

你是否對機器學習解決現實世界問題的能力感到著迷？ML 奧林匹亞競賽回來了，為愛好者和開發者帶來超過 20 個激動人心的挑戰！這場獨特的競賽在 Kaggle 上舉辦，邀請來自全球的參賽者在廣泛的領域內展示他們的技能——健康護理、可持續性、電腦視覺等等。

不論你是對使用生物信號 ML 模型檢測吸煙模式的挑戰感到好奇，還是有興趣區分我們的海洋中的水母與塑膠汙染，ML 奧林匹亞競賽為每個人都準備了一些東西。從預測地震損害到預報天氣條件，甚至識別線上有害語言，這些挑戰旨在推動機器學習的邊界並促進創新。

由 ML GDE 和 TFUG 等充滿熱情的社群以及全球開發人員組織的，ML 奧林匹亞競賽不僅僅是一場競賽。這是一個學習、成長和與志同道合的人連接的實踐機會，他們熱衷於通過技術製造差異。在 Google 透過其 Google for Developers 計劃的支持下，這個活動承諾將是一次開創性的體驗。

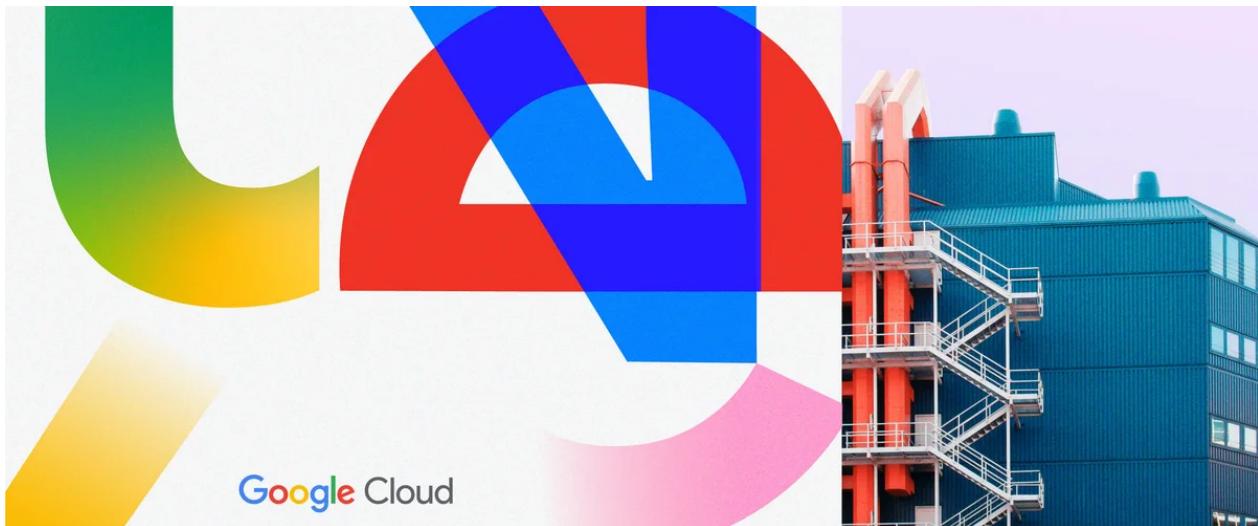
因此，如果你熱衷於應用你的機器學習技能來解決現實世界的問題，或僅僅對 AI 的最新進展感到好奇，ML 奧林匹亞競賽在等著你。深入這個激動人心的世界，參與激發你興趣的挑戰，並成為一個通過技術推動變革的社群的一部分。不要錯過這個機會來測試你的技能，學習新技能，並為有意義的項目做出貢獻。讓我們用 ML 奧林匹亞競賽為更好的明天創新吧！🚀

[閱讀更多](#)

利用AI創新：Google Cloud對企業的 變革性影響

Google Cloud | **AI創新** | **企業變革** | **客服機器人** | **Vertex AI** | **生成性AI** | **數據分析** | **客戶體驗** | **智能文
件摘要** | **聊天機器人** | **電子商務** | **營銷**

2024-04-09



在人工智慧 (AI) 重新定義技術邊界的時代，各行各業的企業都在利用Google Cloud的AI創新，不僅提高生產力和效率，還創造了新穎的客戶體驗和工作場所賦能工具。

作為這場技術革命的前鋒，Google Cloud的年度活動Google Cloud Next展示了超過300家企業如何整合AI技術，革新他們的運營。從客服機器人和AI驅動的員工輔助工具，到安全和編碼的進步，AI的多功能性在整個範圍內都很明顯。

跨行業的創新應用：

- Bayer 通過一個放射學平台重新定義醫療保健，該平台使創建AI驅動的應用程序成為可能。這些應用程序幫助放射科醫生在數據分析和文件記錄方面，簡化診斷過程並加速病人護理。
- Best Buy 通過其Gemini模型提升零售體驗，引入了一個生成性AI虛擬助手，用於無縫客戶互動。這個夏天，預計將帶來一種現代化的購物體驗，完備AI支持的產品故障排除和服務排程。
- Cintas 利用Google Cloud的Vertex AI Search優化服務效率，打造了一個生成性AI驅動的搜索引擎，用於快速訪問重要文件和合同，從而提升客戶服務標準。
- Discover Financial 通過Vertex AI賦能其聯絡中心代理，利用智能文件摘要大大減少通話時間，提升客戶支援效率。

- IHG Hotels and Resorts 通過IHG One Rewards應用中的生成性AI聊天機器人，即將革新旅遊規劃，提供指尖下的個性化假期規劃。
- Mercedes-Benz 將AI整合到其電子商務中，提供一個更智能的在線購物旅程，擁有AI驅動的銷售助理和個性化營銷活動，承諾提供精緻的客戶體驗。
- WPP 與Google Cloud的生成性AI一同重新想像營銷，將Gemini 1.5 Pro模型整合入WPP Open，以前所未有的準確性預測內容表現，並探索實時視頻描述和敘事增強，為迷人的故事講述增添魅力。

這些Google Cloud AI技術的開創性應用例子展示了AI在從醫療保健到零售，從金融服務到汽車等多個不同領域的變革潛力。隨著公司繼續探索和實施這些先進工具，AI的可能性視野不斷擴大，承諾未來將技術和人類創造力融合，解鎖創新和效率的新領域。

敬請期待更多塑造我們世界的技術進步故事。

[閱讀更多](#)

探索AI的全球影響：Google AI Podcast系列洞見

AI | Google AI Podcast | 政策 | 經濟 | 科學 | 民主 | 可持續性 | 生成式AI | 氣候變化 | 全球可持續發展
目標

2024-04-11



探索AI的全球影響：Google AI Podcast 系列洞見

在人工智慧（AI）影響社會的每一個層面的時代，Google 開始了一項任務，透過與世界上一些最具洞察力的思想家進行引人入勝的對話，來揭開AI效應的神秘面紗。這家科技巨頭最近推出了一個限定系列播客，題為「與全球領袖對談AI與社會的6次對話」，旨在闡明AI如何與政策、經濟、科學、民主、可持續性等多方面交匯。

該系列擁有與全球專家的深入討論，包括：
- 政策與AI：James Manyika 和 Ngaire Woods 探討如何透過策略性政策利用AI的優勢，同時限制其風險。
- 透過生成式AI推動經濟進步：Kent Walker 和 Andrew McAfee 深入討論生成式AI如何成為經濟增長和生產力提升的催化劑。
- AI在科學和醫學中的應用：與Anna Greka 的對話揭示了AI在理解遺傳疾病和制定精確治療方案中的關鍵作用。
- AI與民主：Yasmin Green 和 Rory Stewart 考察AI對民主過程的影響，以及在指導AI倫理發展中政策的重要性。
- 加速AI在氣候行動中的應用：Lord Nicholas Stern 討論AI在推動更綠色過渡和對抗氣候變化效應中的工具作用。
- AI與全球可持續性：Matt Brittin 和 Kate Garvey 探索AI實現聯合國全球可持續發展目標的潛力。

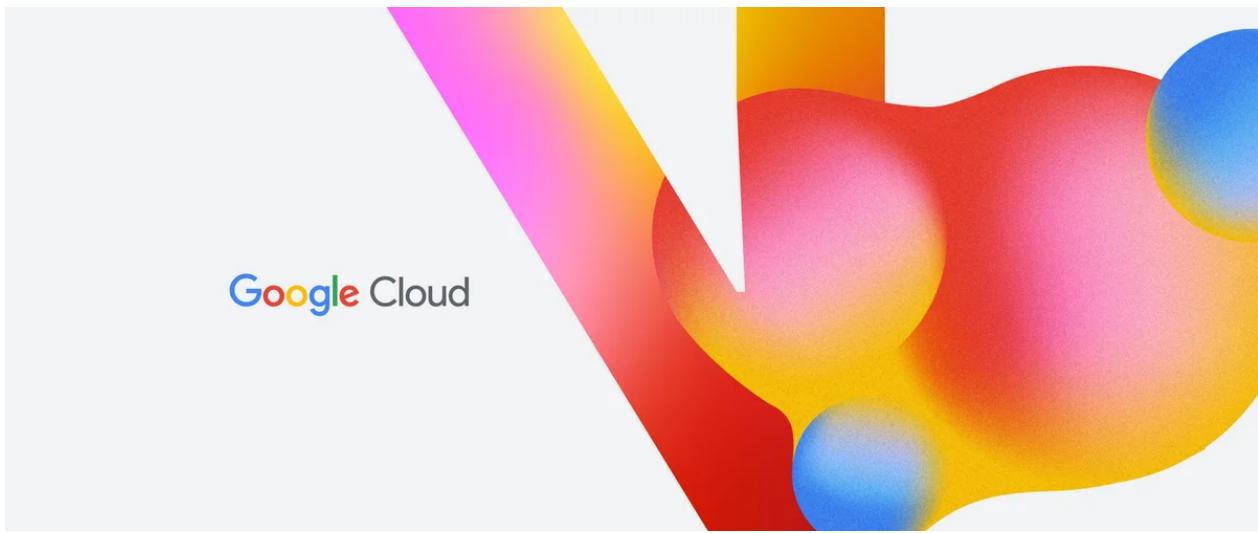
這些啟發性的討論旨在呈現AI創新的平衡觀點，突顯其轉型的好處以及內在風險。這個播客系列可在YouTube、Spotify和Apple Podcasts等主要平台上收聽，為從政策制定者到普羅大眾提供洞見，以使AI更好地服務於全球社會的進步。

[閱讀更多](#)

解鎖未來：在 Google Cloud Next '24 上實現的實際 AI 轉型

Google Cloud | 生成式 AI | 客戶服務 | 員工生產力 | 數據分析 | 編碼 | 網絡安全

2024-04-12



解鎖未來：在 Google Cloud Next '24 上實現的實際 AI 轉型

在 Google Cloud Next '24 的激動人心的展示中，我們見證了生成式 AI 的力量如何正在革新公司、政府、初創企業和研究人員的工作方式。從使客戶服務互動更加順暢到增強員工生產力，這些進步預示著效率和創新的新時代。

重新想像客戶服務

想像與客戶代理互動，他們不僅完美地理解您的需求，還能在各種平台上無縫引導您找到最佳解決方案。像 Best Buy 和 Alaska Airlines 這樣的公司正在使用 Google Cloud AI 創建超個性化的客戶體驗，而 ADT 則在開發一個 AI 客戶代理，以革命性地改變家庭安全設置。

前所未有的賦能員工

員工代理將通過管理日常任務和簡化溝通來轉變工作場所，從而提升協作和生產力。像 Avery Dennison 和 Bank of New York Mellon 這樣的組織正在利用生成式 AI 創建靈活、安全的協作平台和虛擬助手，使信息檢索變得輕而易舉。

創意解決方案觸手可及

AI 代理正在增強營銷和生產團隊的創造力，能夠以前所未有的速度和效率生成吸引人的內容，從產品描述到動態活動。Carrefour 和 Canva 只是幾個例子，AI 正在讓任何人都能挖掘內心的藝術家或設計師。

數據分析變得簡單

數據代理為組織提供了相當於隨時擁有一支數據分析師團隊的能力，隨時準備回答複雜問題並在廣泛的數據集中揭示洞見。這種能力正在革新像 Anthropic 和 Essential AI 這樣的企業利用信息進行戰略決策的方式。

使用 AI 進行編碼：開發者的夢想

軟件開發領域正因代碼代理而在生產力和代碼質量上獲得顯著提升。這些由 AI 驅動的助手幫助開發者熟悉新的編程語言，提高代碼質量，並加快開發週期，正如 Capgemini 和 Commerzbank 展示的那樣。

通過 AI 強化安全

最後，由 AI 驅動的安全代理在監控、威脅檢測和合規性方面提供了前所未有的速度和效率，重塑了網絡安全格局。例如，BBVA 和 Pfizer 的創新正在為保護數位資產設定新的標準。

當我們期待這些 AI 進步塑造的未來時，很明顯，Google Cloud 的技術套件正處於這一轉型旅程的最前沿。

[閱讀更多](#)

用 Google 的生成式 AI 革新您的廣告視覺故事講述

生成式 AI | 廣告 | 視覺故事講述 | Google | Demand Gen | 數位水印 | SynthID | 轉換率 | 數位故事講述

2024-04-16



用 Google 的生成式 AI 革新您的廣告視覺故事講述

在視覺內容稱霸的數位時代，Google 正將廣告推向新的高度，其最新科技突破應用於 Demand Gen 活動中。想像一下，只需幾下點擊，就能為您的廣告創建令人驚艷的高質量圖片。多虧了 Google 的生成式 AI，這現在對全球的廣告商而言已成為現實。

生成式 AI 正在改變品牌講述故事的方式，使創建迎合您特定需求的引人入勝的圖像資產成為可能。無論是從頭開始，還是尋求在現有成功視覺上擴展，Google 的 AI 都能使用您提供的簡單文字提示，生成大量獨特的圖片。想像一下：您正在推銷露營裝備，想要在迷人的北極光下展示您的產品。有了 Google 的 AI，現在就能夠將您色彩鮮豔的帳篷描繪在如此迷人的背景下，以前所未有的方式吸引潛在客戶。

這些進步不僅僅是圖像生成。Google 致力於負責任地使用 AI，確保每個生成的圖片都是獨一無二的，並且可識別為 AI 生成。這項倡議包括一個 SynthID，一個數位水印，使每個圖像都可被驗證，保持品牌的真實性和信任。

此外，這些工具的整合到 Demand Gen 活動中已顯示出有希望的結果，使用視頻和圖像廣告的廣告商與那些僅依賴圖像的廣告商相比，每美元轉換率增加了 6%。Google 的指導不僅止於圖像創建；它還擴展到跨格式優化您的創意策略，確保您的內容無論出現在哪裡，從 YouTube Shorts 到 Gmail，都感覺原生且吸引人。

在擁抱 Google 的生成式 AI 工具時，廣告商不僅是在增強他們的創意能力；他們正踏入數位故事講述的新領域，想像力是唯一的限制。敬請關注更多洞察，並確保在您的日曆上標記 Google Marketing Live 2024，以發現這些創新如何革新您的廣告策略。

透過Bing和GenAI革新競爭情報與銷售策略

Bing | 生成式人工智能 | **GenAI** | 競爭情報 | 銷售策略 | 市場分析 | 數據收集 | 市場動態 | **Vodafone**
TOBi

2024-04-18



透過Bing和GenAI革新競爭情報與銷售策略

在不斷變化的數位領域中，Microsoft透過整合Bing和生成式人工智能（GenAI），提供了前所未有的銷售策略和競爭分析工具。這種創新的方法正在重塑企業理解市場動態、識別機會以及導航競爭格局的方式。

在這場轉變的核心，是能夠高效地收集和分析來自互聯網的大量數據，豐富內部資料庫，提供對營運和市場環境的全面視角。通過區分內部數據的結構化、洞察性質和外部數據的策略性、趨勢導向，企業現在可以利用Bing獲取更廣泛的信息範疇。

這種策略不僅僅是關於數據收集；它是一次深入了解市場份額、顧客情緒和競爭對手策略的深潛，通過先進的生成式AI模型來實現。Bing作為一個從互聯網提取豐富數據的重要工具，當與GenAI結合使用時，可以解碼複雜的市場動態，提供市場份額分佈、顧客偏好和品牌感知的洞察。這種雙技術方法使企業能夠精煉他們的銷售策略和競爭分析，確保它們不僅是時下的，也是前瞻性的。

這項技術的一個現實世界應用例子可以在Vodafone的AI驅動數位助理TOBi的開發中看到，它展示了整合Bing和AI以增強客戶服務和運營效率的實際益處。同樣地，透過Bing和GenAI分析競爭對手的定價策略、產品供應和品牌感知，為企業提供了創新和在競爭賽跑中保持領先的知識。

此外，這種整合支持了一種持續學習和技術素養的文化，鼓勵員工適應並與技術進步共同成長。這種方法標誌著朝向更以人為本的技術的轉變，增強了勞動力和企業營運策略的潛力。

Microsoft透過Bing和GenAI的創新使用，不僅是競爭情報和銷售策略的遊戲規則改變者，也是技術推動商業成功力量的證明。通過擁抱這些工具，公司可以解鎖新的效率、創造力和成長水平，確保他們在數位時代保持競爭力和創新性。

[閱讀更多](#)

用亞馬遜的即時會議助理革新您的會議體驗

即時會議助理 | 亞馬遜Transcribe | 亞馬遜Bedrock | 實時轉錄 | 即時翻譯 | 隱私 | 合規性 | 大型語言模型 | 開源 | AWS

2024-04-18

Bob Strahan 03:31.4 - 03:37.2
Um, I'll look into Dynamo DB. And that's plan to be group next Friday with Chris and Kishore does that work?

Babu Srinivasan 03:37.7 - 03:42.0
Sounds good. I will have a kinesis prototype ready by next week.

用亞馬遜的即時會議助理革新您的會議體驗

在忙碌的商業世界中，會議是日常的一部分。然而，它們帶來了自己的一套挑戰，比如在參與對話時進行筆記、快速查找信息、保持遲到者更新，以及管理語言障礙。亞馬遜的即時會議助理(LMA)正在這裡為您的會議體驗帶來轉變，用尖端技術解決所有這些問題。

由亞馬遜Transcribe、亞馬遜Bedrock及亞馬遜Bedrock的知識庫所驅動，LMA為任何希望簡化會議工作流程的人提供了改變遊戲規則的功能。它從您的基於瀏覽器的會議（目前支持Zoom和Chime）中捕獲音頻和元數據，提供實時轉錄、75種語言的即時翻譯和情境感知輔助等功能。它可以使用即時轉錄作為上下文，從您信賴的來源提供答案，確保您不會錯過任何重要信息。

LMA不止於此；它還提供隨需摘要，讓任何遲到的人迅速追上進度變得輕而易舉。這些摘要由亞馬遜Bedrock生成，確保您能夠隨手掌握所有行動項目、負責人和截止日期。一旦會議結束，LMA會自動生成一份全面的摘要並使用大型語言模型提取洞察力。

對於那些關心隱私和合規性的人，LMA為您提供了保障。它確保您負責遵守有關錄製會議的法律和公司指南。

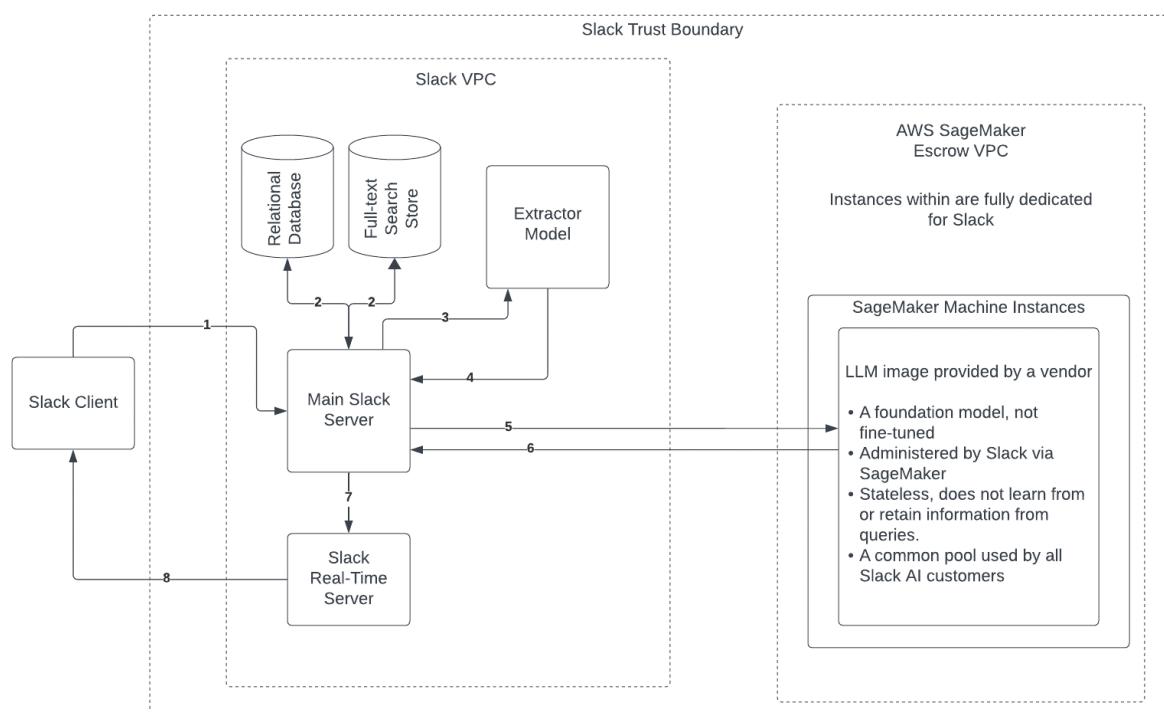
設置LMA很簡單，一切都作為開源提供在GitHub上。您所需要的只是一個AWS賬戶，就可以在您的工作空間部署這一奇蹟。向會議中的多任務壓力說再見，用亞馬遜的即時會議助理迎接效率和生產力的新時代。

[閱讀更多](#)

Slack 透過 AI 提高生產力：工作效率的巨大飛躍

Slack | AI | 生產力 | Amazon SageMaker JumpStart | 數據安全 | 用戶隱私 | 工作效率

2024-04-18



Slack 透過 AI 提升生產力：工作效率的飛躍進步

在數位工作場所的激動人心進展中，Slack 引入了創新的生成式 AI 功能，革命性地改變了專業人士與工作空間互動的方式。這一突破，由 Amazon SageMaker JumpStart 驅動，承諾將單調任務轉變為一個無縫、高效的過程，使用者能夠毫不費力地利用他們對話中的集體智慧。

Slack AI 不僅僅是另一個工具；它是一個遊戲規則改變者。通過將先進的 AI 直接整合到平台中，Slack 現在提供了一套旨在最大化生產力的功能。從快速篩選大量信息的智慧搜尋功能，到將冗長討論精煉為簡明要點的對話摘要，以及個人化的每日回顧，Slack AI 確保用戶始終了解情況，而不會感到不知所措。

這項創新的核心是 SageMaker JumpStart，一個引擎，為 Slack 提供訪問最先進的 AI 模型，同時優先考慮數據安全和用戶隱私。Slack AI 的特別之處在於其承諾將客戶數據保留在平台內，遵守用戶所期望的最高標準的安全和合規性。這一舉措證明了 Slack 對維護信任和完整性的承諾，確保 AI 驅動的功能不僅高效，而且在設計上是安全的。

對於組織和用戶來說，這意味著朝著更有意義和更集中的工作方向邁出了重要一步。Slack AI 即將重新定義生產力的概念，使團隊更容易挖掘他們的集體知識，簡化他們的工作流程，並以前所未有的效率實現他們的目標。

隨著我們擁抱這個新時代的工作場所技術，Slack AI 作為創新的一盞明燈，承諾一個未來，工作不僅僅關於我們做什麼，而是關於我們如何智能和安全地做。

[閱讀更多](#)

NVIDIA 在 EMEA 地區慶祝 AI 成就，頒發合作夥伴獎項

NVIDIA | AI | EMEA | 合作夥伴獎項 | 技術應用 | 市場創新 | 產業變革 | 教育 | 醫療保健 | 生成式AI | 機器人手術平台 | 虛擬助理

2024-04-18



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

Watch On Demand

在對人工智慧創新與進步的顯著致敬中，NVIDIA 最近表彰了其在歐洲、中東和非洲 (EMEA) 的合作夥伴所取得的成就。這次表彰活動是在年度 EMEA 合作夥伴日期間舉行，是 NVIDIA 合作夥伴網絡 (NPN) 倡議的一部分，著重展示 AI 在多元產業中的變革力量。

今年的活動突出了 18 家傑出合作夥伴的寶貴貢獻，這些合作夥伴在七個獎項類別中獲得認可，這些類別凸顯了這些組織如何利用 NVIDIA 的 AI 技術在各自領域開創創新道路。

在眾多獎項中，Rising Star Awards 頒給了三家展現出卓越成長和創新的公司。Vesper Technologies 在北歐因其顯著的營收增長和廣泛的客戶基礎脫穎而出，利用 NVIDIA AI 解決方案於數據中心。AMBER AI & Data Science Solutions GmbH 在中歐展現了值得讚揚的營收增長，超過 100%，成為德國 NVIDIA 合作夥伴生態系統中的關鍵玩家。HIPER Global Enterprise Ltd. 獲得了南歐及中東地區的獎項，因其優異的服務和在利用 NVIDIA 計算技術的客戶項目上的影響而受到慶祝。

Star Performer Awards 頒給了 Boston Limited、DELTA Computer Products GmbH 和 COMMIT DMCC，每家公司都因在各個領域部署 NVIDIA 技術的卓越成就而受到認可，從產業和教育到在阿聯酋的策略性解決方案實施。

PNY Technologies 和 TD Synnex 因其分銷卓越而獲表彰，PNY 連續第三年贏得年度分銷商，TD Synnex 則因其網絡能力和技術專長而受到認可。

在市場進入卓越方面，Bynet Data Communications Ltd. 因其在以色列市場的策略性舉措而獲得表彰，Vesper Technologies 和 M Computers s.r.o. 也因其在 AI 參與的市場創意和領導力而獲得肯定。

WPP 因其基於 NVIDIA Omniverse 平台的生成式 AI 驅動內容引擎榮獲行業創新獎，這一平台正在革新市場營銷和廣告中的內容創造。Ascon Systems 和 Gcore 因其在工業流程和語言技術方面的開創性應用而受到高度讚揚。

Pioneer Award 表彰了 Arrow Electronics – Intelligent Business Solutions 推廣 NVIDIA IGX Orin 於醫療保健領域，這標誌著在機器人手術平台方面的重大進展。

年度諮詢合作夥伴獎頒給了 SoftServe，因其全面的培訓計劃和 NVIDIA 技術的專業知識，Deloitte 和 Data Monsters 也因其對 AI 轉型和虛擬助理開發的貢獻而受到認可。

這些獎項凸顯了 NVIDIA 及其合作夥伴在推動 AI 採用方面的協作精神和共同願景，展現了不僅定義了當今技術景觀，也為未來創新鋪平了道路的顯著成就。

[閱讀更多](#)

05 服務

立即體驗 ChatGPT，開啟 AI 世界的探索之旅

ChatGPT **OpenAI** **AI 民主化** **無需註冊** **增強的內容保護措施** **保存聊天歷史記錄** **語音聊天** **定制命令**

2024-04-01

OpenAI 剛剛宣布了一項令 ChatGPT 愛好者和新手都感到興奮的更新：無需註冊即可即時使用 ChatGPT。這一突破性舉措旨在將 AI 民主化，讓每個人、每個地方都能輕鬆接觸到 AI。已經有超過 1 億用戶在 185 個國家透過 ChatGPT 探索新知識、尋找創造性靈感，並找到問題的答案，OpenAI 正在拓展其視野，使這些體驗變得更加易於獲得。

從今天開始，可訪問性是遊戲的名字。您可以立即深入使用 ChatGPT 並開始您的 AI 之旅。這項舉措是 OpenAI 確保 AI 巨大潛力對所有人都可用，促進全球學習和創新環境的承諾的一部分。

除了無障礙訪問外，OpenAI 還引入了增強的內容保護措施，以及讓用戶有機會影響他們的互動如何幫助改善所有人的模型—無論有無帳戶。此外，對於選擇註冊的人來說，還有許多額外功能等待著，包括保存聊天歷史記錄、分享對話，甚至進行語音聊天和定制命令。

如果您對 AI 領域感到好奇，但設置帳戶的過程似乎令人生畏，現在是您無拘無束探索的機會。OpenAI 邀請您今天就體驗 ChatGPT，加入數百萬人發現 AI 持有的無限可能性。

OpenAI：搭建好奇心與技術之間的橋樑。

[閱讀更多](#)

以NVIDIA GeForce NOW遊戲陣容邁入未來

NVIDIA | GeForce NOW | 雲端遊戲 | Fallout系列 | 串流 | 遊戲體驗

2024-04-11



踏入未來，體驗NVIDIA GeForce NOW遊戲陣容

在令玩家興奮的更新中，NVIDIA通過將Bethesda標誌性的「Fallout」系列，包括Fallout 4和Fallout 76，引入其GeForce NOW服務，擴展了雲端遊戲的視野。這次新增引領了本週10款新遊戲的大軍，承諾在任何設備上帶來在廣闊的核戰荒地中的冒險旅程。

進入雲端保險庫

Fallout 4邀請玩家探索聯邦的廢墟，每一個決定都為新的未來鋪路。另一方面，Fallout 76則開放早期核後的阿巴拉契亞，讓玩家在不斷發展的線上世界中結盟或形成對立。通過GeForce NOW，這些沉浸式體驗現在可以在各種設備上無縫串流，即使是使用終極會員資格以4K質量，也確保您的遊戲冒險永不因硬件限制而停止。

此外，GeForce NOW還不僅僅止步於此。服務還引入了一系列新遊戲，包括備受期待的Gigantic: Rampage Edition，承諾帶來史詩般的5v5戰鬥，以及如Inkbound 1.0和Broken Roads等激動人心的標題。無論您是在MOBA比賽中制定策略，還是開始敘事豐富的旅程，NVIDIA都確保提供高品質的遊戲體驗，隨時隨地都可以訪問。

一週的新發現

這次更新與Fallout系列電視改編的發布完美對齊，為粉絲們帶來了一場Fallout盛宴。除了Fallout標題之外，GeForce NOW會員還有大量選擇，從Ghostrunner的快節奏動作到Terra Invicta的策略遊戲，直接串流到他們選擇的屏幕上。

加入雲端革命

NVIDIA的GeForce NOW革新了遊戲體驗，打破了高品質遊戲體驗與全球玩家之間的障礙。通過串流這些引人入勝的標題，包括Fallout 4和Fallout 76的沉浸式世界，NVIDIA不僅增強了對頂級遊戲的訪問，也確保了無論您身在何處，使用何種設備，冒險永遠不會停止。

與GeForce NOW一起開始您的下一場遊戲冒險，發現雲遊戲的無限潛力。

[閱讀更多](#)

利用 AWS CloudFormation 簡化 Amazon Lex 聊天機器人的部署

AWS CloudFormation | **Amazon Lex** | **聊天機器人** | **自然語言理解** | **自動化部署** | **版本控制** | **可擴展性** | **AWS Lambda** | **CloudWatch**

2024-04-16

Alias name	Created	Associated version
TestBotAlias	6 minutes ago	Draft version

利用 AWS CloudFormation 簡化 Amazon Lex 聊天機器人的部署

在不斷進化的人工智慧領域中，聊天機器人和虛擬助理已成為提升用戶體驗的不可或缺的一環。AWS 提供的著名服務 Amazon Lex，利用先進的自然語言理解技術，賦予開發人員創建複雜對話介面的能力。最近的進展引入了一種使用 AWS CloudFormation 模板管理 Amazon Lex 機器人的方法，這在自動化和簡化部署過程方面是一個重大進步。

AWS CloudFormation 為您的 Amazon Lex 機器人提供了一個藍圖，涵蓋了所有必要的 AWS 資源，從而消除了部署不一致性和人為錯誤。這種方法確保了在開發、測試和生產等各種環境中均能實現統一的部署機制，大大減少了設置時間和努力。

主要優點包括：

- 一致性和自動化：在所有環境中一致部署您的聊天機器人，極小化手動干預。
- 版本控制：使用像 Git 這樣的系統，輕鬆管理並回滾您機器人的版本，確保可靠性。
- 可擴展性：輕鬆管理複雜的聊天機器人配置，使擴展變得輕而易舉。
- 整合和自動化：無縫整合 AWS Lambda 函數以實現自定義邏輯，並自動化整個部署生命週期。

該過程涉及創建一個 CloudFormation 模板來定義 Amazon Lex 機器人，包括意圖、槽位和機器人運作所需的其他組件。這種方法不僅簡化了最初的部署，還支持版本控制和別名，允許通過藍/綠部署有效管理生產和開發環境。

此外，將 AWS Lambda 函數整合到 CloudFormation 模板中引入了自定義層，使開發人員能夠直接在聊天機器人的工作流中添加驗證、初始化和履行邏輯。條件分支功能進一步增強了機器人的對話能力，允許根據特定業務邏輯或用戶輸入動態響應。

通過在 CloudFormation 模板中實現 Amazon CloudWatch 日誌，提供了機器人互動的透明度和洞察力，對於持續改進和維護至關重要。

總之，使用 AWS CloudFormation 管理 Amazon Lex 機器人標誌著簡化聊天機器人部署和管理的重要一步，承諾了一致性、可擴展性和自定義能力。這種方法不僅節省了寶貴的開發時間，還為創建更具吸引力和智能的對話介面開啟了新的可能性。

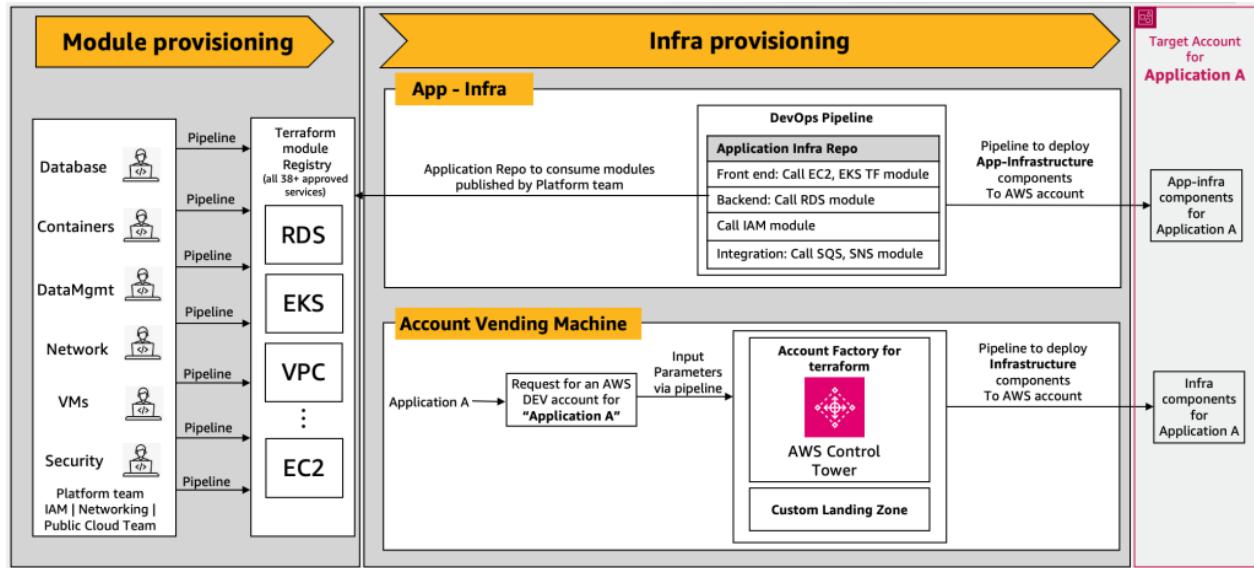
這篇緊湊的敘述旨在概括利用 AWS CloudFormation 管理 Amazon Lex 機器人的精髓，使技術細節對非專業觀眾來說變得容易理解和接近，從而釋放自動化聊天機器人部署和管理的全部潛力。

[閱讀更多](#)

利用 Amazon Bedrock 的 AI 驅動 IaC 腳本革新雲端遷移

Amazon Bedrock | AI | IaC | 雲端遷移 | AWS | Terraform | CloudFormation | 安全標準 | 合規要求

2024-04-18



利用 Amazon Bedrock 的 AI 驅動 IaC 腳本革新雲端遷移

在快速進化的雲計算世界中，組織不斷尋找方法來簡化他們的遷移過程，同時確保遵守合規和安全性。Amazon Web Services (AWS) 引入了一個創新的解決方案來應對這一挑戰：Amazon Bedrock。此服務利用生成式人工智能 (AI) 自動創建基礎設施即代碼 (IaC) 腳本，專為 AWS Landing Zone 設計，使雲端轉換更加順暢和高效。

Amazon Bedrock 簡化了 Terraform 和 CloudFormation 腳本的生成，這些工具對於定義和管理雲環境至關重要。Bedrock 的區別在於它能夠產生符合組織特定需求、安全標準和合規要求的定制腳本。這是通過利用來自領先 AI 公司的高性能基礎模型實現的，可通過單一 API 訪問。

對於雲工程師，無論是資深的還是初學者，這意味著在學習 IaC 細節和手動編碼上花費的時間大大減少。Bedrock 讓團隊能夠輸入高層次的架構描述，從中生成基線的 Terraform 腳本。這個腳本不僅適合組織的獨特要求，還納入了行業標準的安全和合規措施。

此外，AWS Landing Zone 與 Amazon Bedrock 實現了無縫整合，提供了部署 AWS 資源的標準化方法。這確保組織能夠從一開始就建立一個安全、合規、高效的雲基礎。

Amazon Bedrock 不僅加速了雲端遷移之旅，還有助於持續優化雲基礎設施。對於旨在利用 AWS 力量的組織來說，它是一個減少複雜性和提高安全性的遊戲改變者。

與 Amazon Bedrock 一同深入雲端遷移的未來，體驗前所未有的簡單和效能。這是雲計算新紀元的開始，AI 驅動的解決方案將技術挑戰轉化為成長和創新的機會。

[閱讀更多](#)