企業 AI

適用於企業的生成式 AI 與特定領域模型

利用專門打造的 Intel® AI 軟硬體最佳化訓練與部署，協助您實現業務轉型

intel®

目錄

為什麼要與 Intel 合作？

在 **Intel**，我們有機會為地球上的每個人和每家企業改善生活與成果

但我們並非單槍匹馬！

我們與合作夥伴並肩同行，到處推廣 **AI**，並將部署風險降至最低，為客戶創造真正的價值

與 **Intel** 合作時，您是與完整的 **AI** 生態系統合作

我們琳瑯滿目的 AI 技術組合，以及與軟硬體和系統整合商的合作夥伴關係密切合作，正在打造實際的解決方案，為產業、公司和社群帶來與眾不同的業務成果。

協助您的業務成長茁壯。

加入我們，展開「AI 無所不在」的旅程

利用 Intel® AI 解決方案為客戶創造價值

Intel 的方法能讓廣大的開放式 AI 參與者生態系統提供滿足特定企業 GenAI 需求的解決方案

NAVER

開發強大的大型語言模型（LLM），用於在全球部署從雲端到裝置的進階 AI 服務。NAVER 已確認 Intel® Gaudi® 在為大型變壓器模型執行運算操作方面的基礎能力，
具有優異的每瓦效能。

seekr

值得信賴的 AI 領域的領導者在 Intel® Gaudi® 2、Intel® Data Center GPU Max Series 和 Intel® Tiber™ AI 雲端中的 Intel® Xeon® 處理器上執行生產
工作負載，提供 LLM 開發與生產部署支援。

BOSCH

探索智慧製造的更多機會，包括產生製造異常合成資料集的基礎模型，提供強大、均勻分佈的訓練集（例如自動光學檢測）。

IFF

食品、飲料、氣味和生物科學的全球領導者將利用 GenAI 和數位孿生技術建立整合式的數位生物學工作流程，進而最佳化先進的酶設計和發酵流程。

IBM

利用第 5 代 Intel® Xeon® 處理器作為其 watsonx.data™
資料儲存，並與 Intel® 密切合作，驗證適用於 Intel® Gaudi® 加速器的 watsonx™ 平台。

airtel

Airtel 憑藉 Intel 尖端技術的力量，規劃利用其豐富的電信資料增強 AI 功能，並加速提升客戶的體驗。這項部署符合 Airtel 堅持領航技術創新的承諾，在瞬息萬變的數位
領域協助推動新的收益來源。

OLA

預先訓練及微調其第一個具有 10 種語言生成功能的印度基礎模型，提供業界領先的性價比與市場解決方案。
Krutrim 現在正在 Intel® Gaudi® 2 叢集預先訓練更大的
基礎模型。

Infosys

新一代數位服務與諮詢的全球領導者宣布了一項策略合作，將第 4 代和第 5 代 Intel® Xeon® 處理器、Intel® Gaudi 2 AI 加速器和 Intel® Core™ Ultra 處理器在內的 Intel® 技術引入 Infosys Topaz，這是一套 AI 優先的服務、解決方案和
平台，利用生成式 AI 技術加速商業價值。

企業 AI 價值主張

利用企業 AI 為您實現業務轉型

在現今競爭激烈的環境下，**採用 AI 的企業正引領群雄。**

各行各業的企業都在重新審視營運的每個環節，
瞭解 AI 如何增強甚至自動化工作流程。

將 AI 嵌入企業架構是 Intel 獨一無二的專業。

從提升生產力的 AI 電腦，到瞭解哪些使用案例能帶來
最大價值的專業知識，Intel® 是您值得信賴的夥伴，
可安全且負責地讓 AI 無所不在。

生成式 AI（GenAI）創新技術預計將以比網際網路時代、
行動時代或雲端時代更快的速度由各種規模的企業採用。

下一波 AI 平台將會以經濟實惠且靈活的方式迎接這些
令人興奮的現實。

是時候以不同的方式思考您的企業 AI 了。

此支援套件將協助您瞭解各種市場的企業如何能從生成式 AI（特別是特定領域模型）
獲得重大價值，實現長期成功

什麼是生成式 AI 與大型語言模型？

生成式 AI（GenAI）是 AI 的一個子集，專注於打全新的原創內容。

它涉及訓練及部署 AI 模型以產生資料，例如影像、文字或音訊，這些資料與訓練資料集的實例非常相似。

GenAI 演算法利用深度學習和神經網路等進階技術產生逼真且連貫的輸出，進而支援影像合成、文字生成乃至於創意藝術作品等應用。

大型語言模型（LLM）是一種特定類型的自然語言處理模型，利用深度神經網路處理及產生文字。LLM 接受大量文字資料的訓練，旨在產生連貫且有意義的輸出。

閱讀更多

掌握生成式 AI 的威力

進一步瞭解

企業將如何使用 GenAI？

消費品與零售

- 虛擬試衣間
- 送貨與安裝
- 協助尋找店內產品
- 需求預測與庫存規劃
- 新型產品設計

醫療保健與醫藥

- 協助忙碌的前線人員
- 轉錄及總結醫療記錄
- 聊天機器人回答醫療問題
- 預測性分析為診斷和治療提供資訊

製造

- 協助技術人員的專家 Copilot
- 與機器對話互動
- 規範性和主動性的現場服務
- 自然語言疑難排解
- 保固狀態與說明文件
- 瞭解流程瓶頸、設計復原策略

媒體與娛樂

- 智慧搜尋、量身打造的內容探索
- 制定標題與文案
- 針對內容品質提供即時的意見回饋
- 個人化播放清單、新聞摘要、建議
- 透過觀眾的選擇進行互動敘事
- 專屬優惠、訂閱方案

金融服務

- 發現交易訊號，提醒交易員注意脆弱的頭寸
- 加速承保決策
- 最佳化與重建舊式系統
- 逆向工程銀行與保險模型
- 監測潛在的金融犯罪和詐騙
- 自動化資料收集以滿足監管法規要求
- 從企業披露的內容擷取深入解析

資料來源：由《麻省理工科技評論》彙編，依據《生成式 AI 時代的零售業》、9《大解鎖：製造業中的大型語言模型》、10《生成式 AI 無所不在、同時啟動》和《媒體與娛樂中的大型語言模型》12 Databricks，2023 年 4 月至 6 月。

生成式 AI 和大型語言模型使用案例

聊天機器人與虛擬助理

客戶支援

程式碼生成與偵錯 LLM

接受公司文件培訓

情緒分析

評估客戶滿意度

文字分類與叢集

將大量資料分門別類以識別趨勢

語言轉譯

將公司網頁轉換成其他語言

總結與釋義

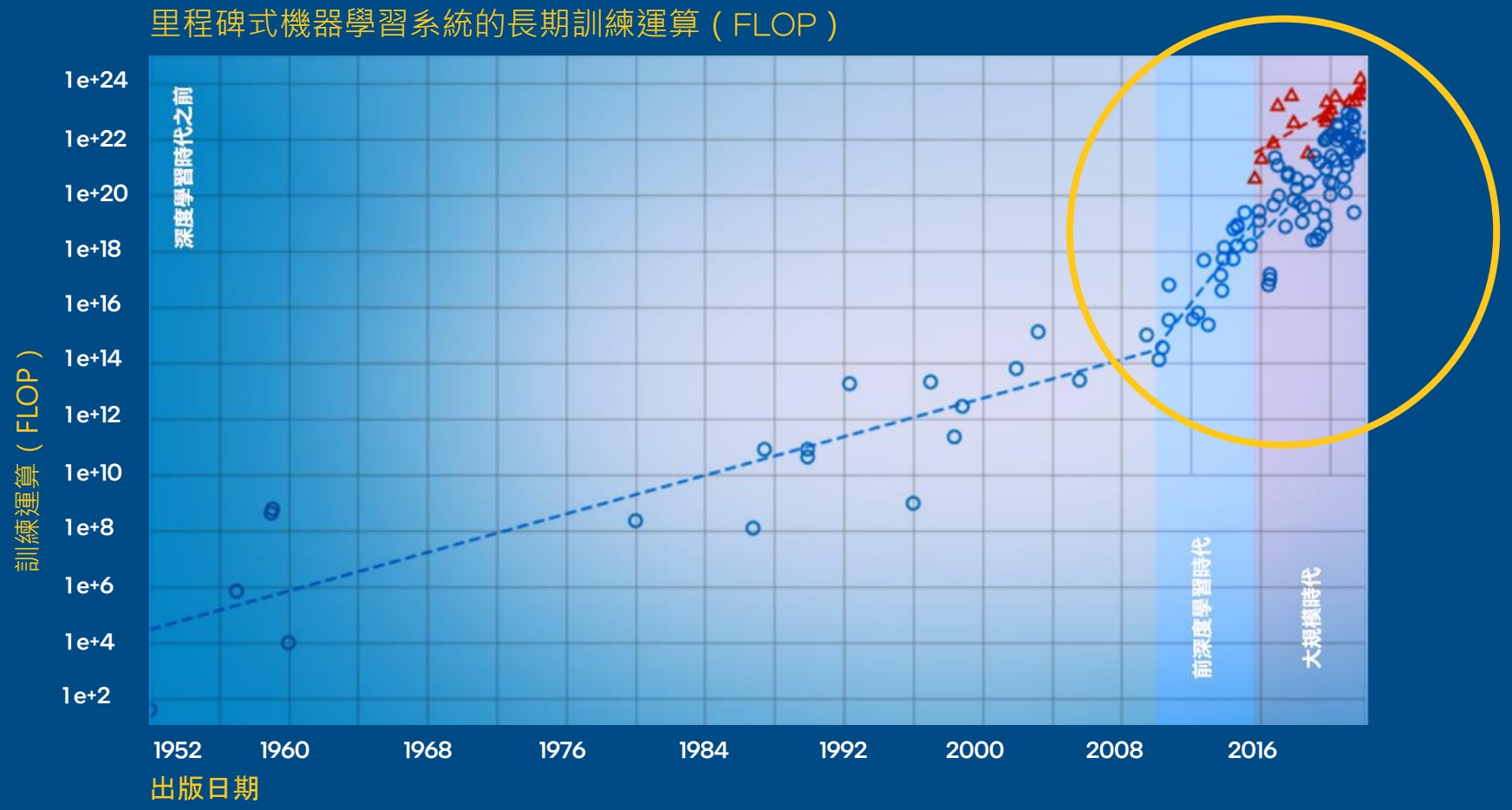
會議記錄總結

內容、影像、影片生成

電子郵件初稿、創意產生、行銷視覺效果、短片

資料來源：大型語言模型的小巧指南，Databricks，2023 年。

隨著模型尺寸的增長，運算也隨之增長



Epoch、亞伯丁大學、AI 治理中心、聖安德魯斯大學、麻省理工學院、圖賓根大學、馬德里康普頓斯大學進行的研究

不只涉及巨型模型

	巨型（第三方）	比	**小巧靈活**（提升 10-100 倍）
可解釋性	**專有模型**	比	根據開放原始碼的模型
準確度	全方位一體式通用	比	專屬、特定領域、自訂
地點	雲端型（即服務）	比	在本機執行推斷；邊緣、用戶端與內部部署
成本	永久性擴充成本	比	成本管理
上市速度	快速設定（秒）	比	建構時間（小時/天）

許多小型模型的成長

6 個月內將數百億參數減少至不到 20 億參數

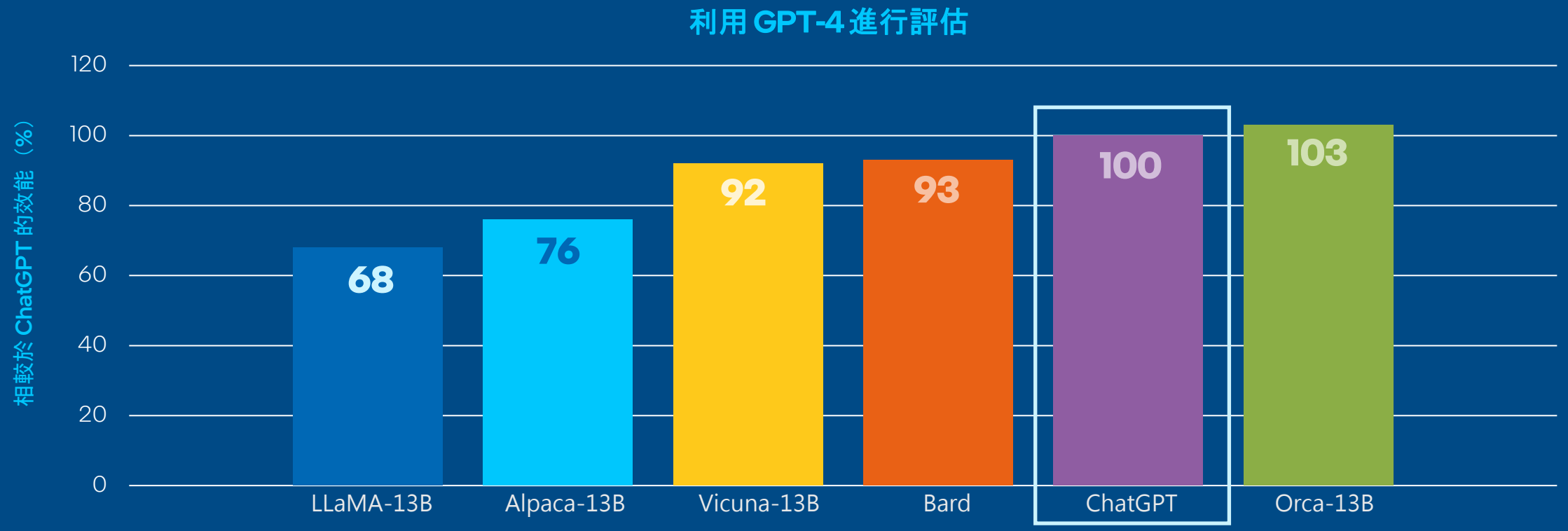
databricks

- 每週出現數十種小型模型
- 商業與開放原始碼授權
- 表示如果利用精心挑選的資料進行訓練，那麼小型模型可以複製大型模型的準確性

- 數千種特定領域的商業模型和 AI 平台正在展示
- 可以在特定領域資料的一些處理器上微調模型

相較於 ChatGPT，小型模型表現良好

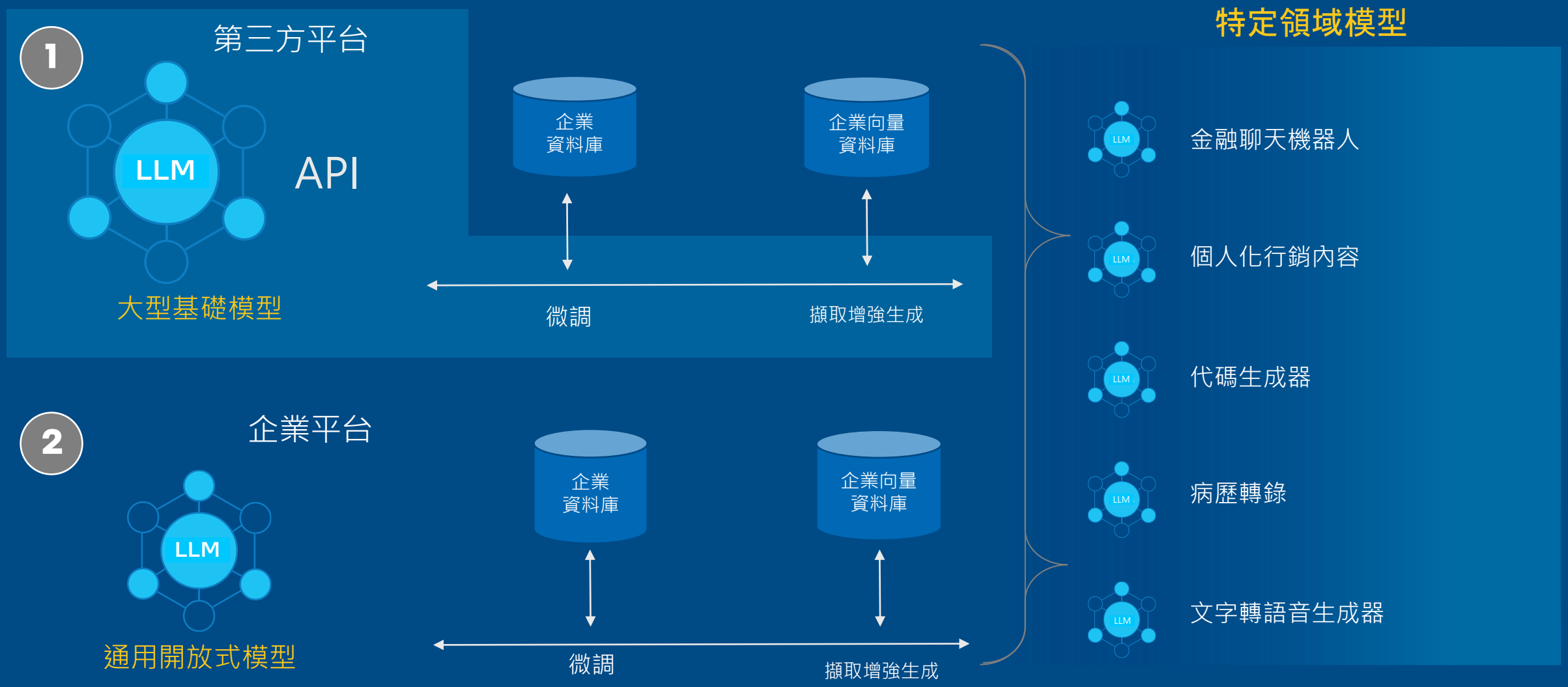
證明小型模型是可行的選項，相較於 ChatGPT 等大型模型，表現仍然優異



根據 GPT-4 在 Vicuna 評估集的評估，Orca 的表現優於多種基礎模型，包括 OpenAI ChatGPT

資料來源：Microsoft Research（2023）。Orca：從 GPT-4 的複雜解釋軌跡中漸進式學習

建構特定領域模型


1
第三方平台
LLM
API
大型基礎模型
企業
資料庫
企業向量
資料庫
微調
擷取增強生成
2
企業平台
LLM
通用開放式模型
企業
資料庫
企業向量
資料庫
微調
擷取增強生成
特定領域模型
LLM
金融聊天機器人
LLM
個人化行銷內容
LLM
代碼生成器
LLM
病歷轉錄
LLM
文字轉語音生成器

特定領域模型為企業帶來許多優勢

小型的專屬模型可以提供同等或更高的效能，透過減少時間和成本投資來提高投資報酬率

更準確的輸出

使用您的企業資料，獲得更具體的領域準確性

成本較低

對預先訓練的模型進行微調，和/或使用 RAG，並推斷小型模型

在您所選擇的平台上隨處部署

在本機執行推斷；邊緣、用戶端與內部部署

安全且具有隱私

符合資料安全與監管需求

負責任的 AI

透過微調和 RAG 讓模型能夠引用資料來源

未來趨勢

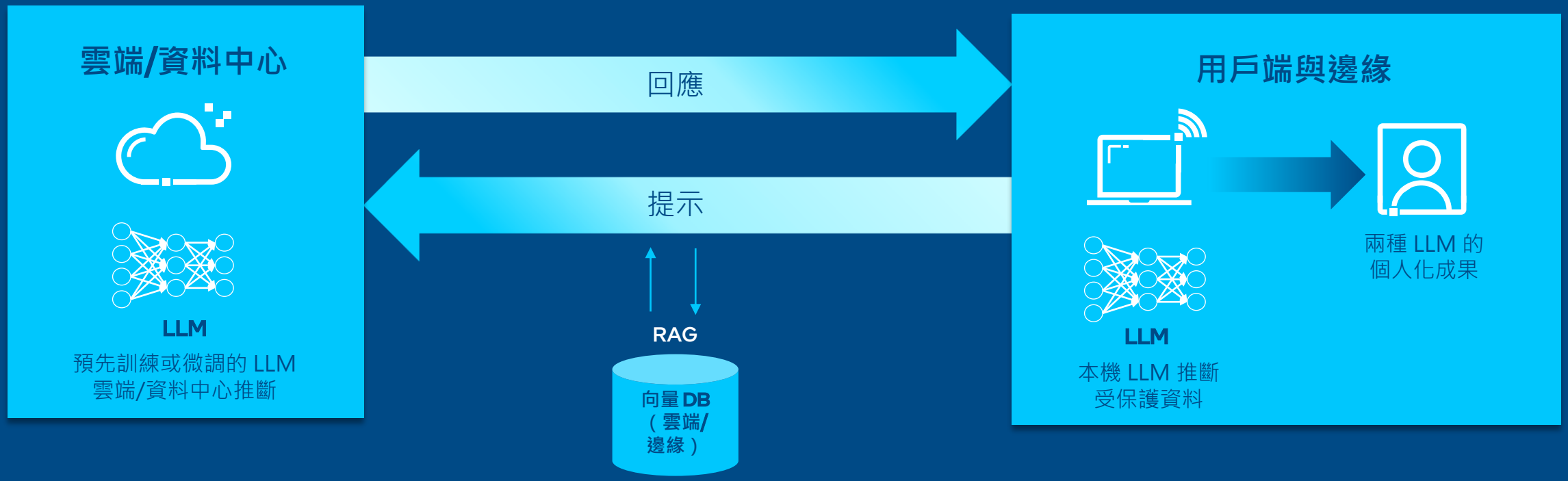
會有少量巨型模型和大量小型、更靈活的 AI 模型嵌入無數個應用程式[1]

[1]資料來源：適者生存：小型生成式 AI 模型是具有成本效益的大規模 AI 的未來

無縫雲端到邊緣 AI 平台

Intel AI portfolio

在雲端進行訓練與推斷。使用 RAG 提高領域準確性。



生成式 AI：一年的生產

特定領域但高度智慧模型的使用正在增加

2022 實驗

2023 試行方案

2024 生產

巨大的模型奠定了基礎

- 對於通用案例非常有效
- 訓練和部署費用高昂
- 建立於大型公共資料集
- 易於使用

小型特定領域模型

- 使用您的私人資料，獲得特定業務成果
- 在您擁有的硬體上部署
- 提高效率、準確性、安全性與可追蹤性
- 建構時間

閱讀部落格

適者生存：小型生成式 AI 模型是具有成本效益的大規模 AI 的未來

Intel 的特定領域模型方法

特定領域模型

優點

+ 模型縮小 10-100 倍，同時維持/提高準確性
+ 通用運算經濟實惠
+ 正確性；來源歸屬；可解釋性
+ 利用私人/企業資料
+ 持續更新資訊

挑戰

- 縮小任務範圍
- 需要少量微調和索引

INTEL 目標

落實最具成本效益且無所不在的方法，利用產業框架、預先訓練的模型和 Intel AI SW 與工具，在 Intel 硬體上微調及部署 10,000 個模型。

閱讀更多

一觸可及的
生成式 AI

電子書 ▪ 資訊圖

企業 AI：協助克服入門障礙

需求	與 Intel® 合作如何提供協助
上市速度	使用 Intel 和 Hugging Face、Gaudi 開發者中心和 5 個參考套件，在生成式 AI 上旗開得勝
使用者體驗（準確性/延遲）	在 Intel® Gaudi® 加速器上對大於 100 億參數的模型進行推斷，在搭載 Intel® AMX 的 Intel® Xeon® 處理器上對小於 200 億參數的小型模型進行推斷，為使用者提供即時體驗[1]
運算可用性	Intel® Xeon® CPU + 加速器為全球 GPU 短缺的問題提供了符合成本效益的替代方案。Intel® Gaudi® 2 現在可透過 SuperMicro 使用，且 Intel® Gaudi® 3 的可用性更高。
熟悉的技術	小型模型的推斷幾乎可以在任何硬體上完成，包括可能已融入您運算環境且無所不在的解決方案[2]
大規模操作	Intel® Gaudi® 2 透過將 24 個 100 GbE 連接埠整合至每個加速器，提供近乎線性的可擴充性。Intel® Xeon® 已融入您的資料中心、現場；雲端到邊緣。65% 的資料中心推斷都在 Intel® Xeon® 上運行[3]
高成本效益	在實際工作應用程式中，Intel® 提供更優異的效能、更低的定價，以及更平衡的 AI 推斷平台，藉此顛覆產業，落實 AI 普及化。查看 NVIDIA 展示 Intel® Gaudi 2 每美元效能比 H100 高出四倍

[1]資料來源：落實生成式 AI 的四大障礙

[2]資料來源：適者生存：小型生成式 AI 模型是具有成本效益的大規模 AI 的未來

[3]根據截至 2022 年 12 月執行 AI 推斷工作負載的全球資料中心伺服器安裝基數的 Intel® 市場模型。

簡化生成式 AI 訓練與部署的軟體資源

開放原始碼模型

- BLOOM — 176B
- BioGPT — 1.5B 領域
- Stable Diffusion — 影像
- Llama2 GPT-J MPT Falcon — 7-65B LLM
- Stanford Alpaca — 微調功能 7B LLM
- 知識 庫

開放式軟體

- OpenVINO
- Intel® Extension for PyTorch (IPEX)
- Intel® Extension for Transformers (ITREX)
- Intel® Extension for DeepSpeed (IDEX)
- haystack
- fastRAG

GenAI 平台

- intel. GAUDI
- intel. DATA CENTER GPU MAX SERIES
- intel. XEON
- intel. CORE ULTRA

閱讀更多

利用無所不在的硬體和開放式軟體解鎖生成式 AI

發揮最大價值

Intel 的開放式 AI 方法為什麼適合您的 AI 業務需求

避免綁定供應商
開放原始碼標準型軟體

MODIN
SciPy
dmlc XGBoost
PyTorch
pandas
NumPy
scikit learn
TensorFlow
Spark
Numba
SigOpt AutoML
oneAPI

利用 Intel 的硬體產品組合
針對 AI 使用案例進行最佳化

intel XEON
intel CORE ULTRA
intel ARC GRAPHICS
intel DATA CENTER GPU
intel GAUDI
intel ETHERNET

利用軟體最佳化的硬體和未來 AI 的開放式標準，
從用戶端和邊緣到資料中心和雲端，開創嶄新的機會

Intel® AI 軟體產品組合

工程化資料 → 建立模型 → 最佳化與部署

MODIN · SciPy · pandas · NumPy · Spark · Numba

dmlc XGBoost · scikit learn · SigOpt AutoML · PyTorch · TensorFlow · intel Neural Compressor · ONNX RUNTIME · DirectML · OpenVINO Write Once Deploy Anywhere

大規模資料分析† | 機器與深度學習架構、最佳化和部署工具†

oneAPI

Intel® oneAPI Deep Neural Network Library | Intel® oneAPI Collective Communications Library | Intel® oneAPI Math Kernel Library | Intel®oneAPI Data Analytics Library

CPU、GPU 和其他加速器開放的跨架構程式設計模型

雲端與企業 | 用戶端與工作站 | 邊緣

intel AI ANALYTICS TOOLKIT
加速端對端資料科學與 AI

Intel® Tiber™ AI 雲端 與 Intel® Developer Catalog
試用最新的 Intel 工具與硬體，並使用最佳化的 AI 模型

cnvrg.io
全堆疊機器學習作業系統

Intel® Geti
註釋/訓練/最佳化平台

Hugging Face
Intel 最佳化與微調配方、最佳化推斷模型和模型服務

注意：堆疊每一層的元件都根據預期的 AI 使用模型，針對其他層的目標元件進行最佳化，而最右欄的解決方案並未使用每個元件
† 這份清單包含針對 Intel 硬體最佳化的熱門開源架構

簡化企業生成式 AI 的採用，縮短可靠、可信任的解決方案生產時間

OPEA：

簡化企業生成式 AI 的採用，縮短可靠、可信任的解決方案生產時間

Open Platform for Enterprise AI

OPEA 合作夥伴

anyscale

CLOUDERA

DATASTAX

kx

DOMINO

Hugging Face

Red Hat

drant

MariaDB

MINIO

sas

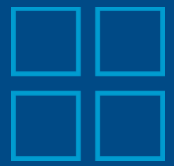
Yellowbrick

vmware

OPEA 價值

- 協助企業使用生成式 AI（LLM、RAG）更快、更輕鬆地釋放手中資料的價值
- 降低分散式生態系統的複雜性，協助解決方案擴大生產規模
- 與 Linux Foundation 合作，激發產業領導者之間的協作與貢獻






效率
利用您選擇的現有基礎架構、AI 加速器或其他硬體。

流暢
與企業軟體整合，具有跨系統和網路的異質支援和穩定性。

開放
匯聚一流的創新技術，不受專有供應商的限制。

無所不在
專為雲端、資料中心、邊緣和電腦打造的靈活架構，在任何地方都可執行。

受信任
採用安全的企業級管道和工具，實現責任、透明度和可追溯性。

可擴充
提供機會入充滿活力的合作夥伴生態系統，協助建置及擴充解決方案。

與 Hugging Face 合作發展生成式 AI

Hugging Face

為了促進生成式 AI 和語言 AI 訓練及創新，Intel 與 Hugging Face 合作，利用這個熱門平台分享 AI 模型與資料集。最引人矚目的是，Hugging Face 是以專為 NLP 打造的 Transformer 資料庫聞名遐邇。

intel XEON

Intel® 與 Hugging Face 合作打造尖端的軟硬體加速，利用 Transformer 模型訓練、微調及預測。

硬體加速由 Intel® Xeon® 可擴充處理器驅動，軟體加速則是由我們最佳化的 AI 軟體工具、架構和資料庫產品組合支援。

intel GAUDI

此外，Intel® Gaudi® 深度學習加速器透過 Optimum Habana Library 搭配 Hugging Face 開放原始碼軟體，可讓開發者
輕鬆使用 Hugging Face 社群最佳化的成千上萬種模型。

Hugging Face 也針對 Intel® Gaudi® 2 在生成式 AI 模型的
效能發布了多項評估：Stable Diffusion、T5-3B、BLOOMZ 176B 和 7B，以及全新的 BridgeTower 模型。

Intel®、Articul8 和 BCG 攜手合作，提供企業級的安全生成式 AI

由 Intel AI 超級電腦支援的開創性解決方案利用自訂資料集釋放商業價值，同時維持高水準的安全性和資料隱私

Articul8* 提供一站式式 GenAI 軟體平台，帶來速度、安全性與成本效益，協助大型企業客戶操作及擴充 AI。該平台是在 Intel® 硬體架構上推出並最佳化，包括 Intel® Xeon® 可擴充處理器和 Intel® Gaudi® 加速器，但會支援各種混合式基礎架構替代方案。

intel GAUDI　intel XEON

波士頓諮詢公司（BCG）早期部署這項技術後，該團隊將平台擴展至需要高度安全性和專業領域知識的產業部門的企業客戶，包括金融服務、航太、半導體和電信。

閱讀更多

Articul8 公告

Articul8 網站

適用於企業的負責任 AI

挑戰：

生成式 AI 模型從網際網路上的大量資料學習，這些資料可能含有社會存在的偏見，也可能會在無意中套用這些偏見。可以操縱 LLM 來產生或傳播錯誤資訊、網路釣魚電子郵件或社交工程攻擊。

LLM 可能常會有「幻覺」並產生不準確的資訊，這在醫療保健等產業尤其嚴重，模型可能會影響診斷和治療決策，並可能對患者造成傷害。

進一步瞭解

將生成式 AI 的風險降至最低

解決方案：

從事 AI 技術的公司和個人需要確保他們的軟體是按照倫理的 AI 原則開發及部署

開放原始碼 Intel®Explainable AI Tools 允許使用者執行事後模型蒸餾和視覺化，以檢查 TensorFlow* 和 PyTorch* 模型的預測行為

LLM 一般會在大型公共資料集上進行訓練，然後對潛在的敏感資料（例如金融和醫療保健）進行微調

Intel 的 Open Federated Learning（OpenFL）等技術結合機密運算，讓 LLM 可以對敏感資料安全地進行微調，進而提高模型的通用性，同時減少幻覺和偏見

適用於生成式 AI 的 Intel® 產品

讓 AI
無所不在

適用於 AI 的可擴充系統

訓練與微調

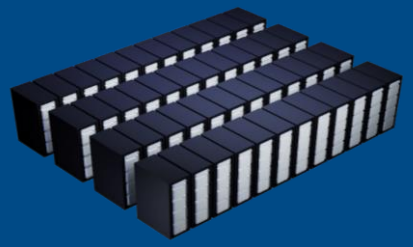
訓練 | 峰值推斷 | 主流推斷/微調 | 基準線推斷 | 端點推斷

推斷與部署

雲端
資料中心

邊緣

用戶端

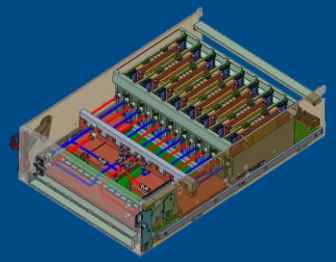


叢集與資料中心規模

intel. Gaudi

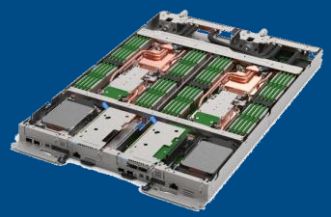
每個機架的
多節點部署

intel. Gaudi | intel. Xeon | intel. Xeon Max Series



多 GPU
或多插槽 CPU

intel. Xeon | intel. Xeon Max Series | intel. Data Center GPU Flex Series



單插槽 CPU 或
單一 GPU

intel. Xeon | intel. Data Center GPU Flex Series | intel Core Ultra

用戶端 CPU

intel Core Ultra | intel. Arc Graphics

intel. Ethernet

適用於 NLP/LLM 的 Intel® 產品

訓練推斷

GAUDI®2

Intel® Gaudi® 2 AI 加速器專為加速 LLM 和 NLP 等大型模型的訓練與推斷而設計。

利用 Intel® Gaudi® 2 加速生成式 AI 與大型語言模型

Intel® Gaudi® 2 為 AI 訓練提供領先的效能和最佳的成本節約

新聞稿

立即觀賞

Intel® 網路研討會錄音討論了 Intel® Gaudi® 2 AI 處理器在挖掘生成式 AI 與大型語言模型（LLM）潛力上的頂尖功能

Intel® Gaudi® 2 深度學習加速器在深度學習訓練與推斷方面表現優異，效能比 **NVIDIA A100 快上最多 2.4 倍**

新聞室 ▪ 技術文章

Intel® Gaudi® 2 仍然是 NV H100 GenAI 效能的唯一基準替代方案

[1]效能因使用、配置和其他因素而異；工作負載和配置詳細資料請見：intel.com/performanceindex 結果可能會有所差異。

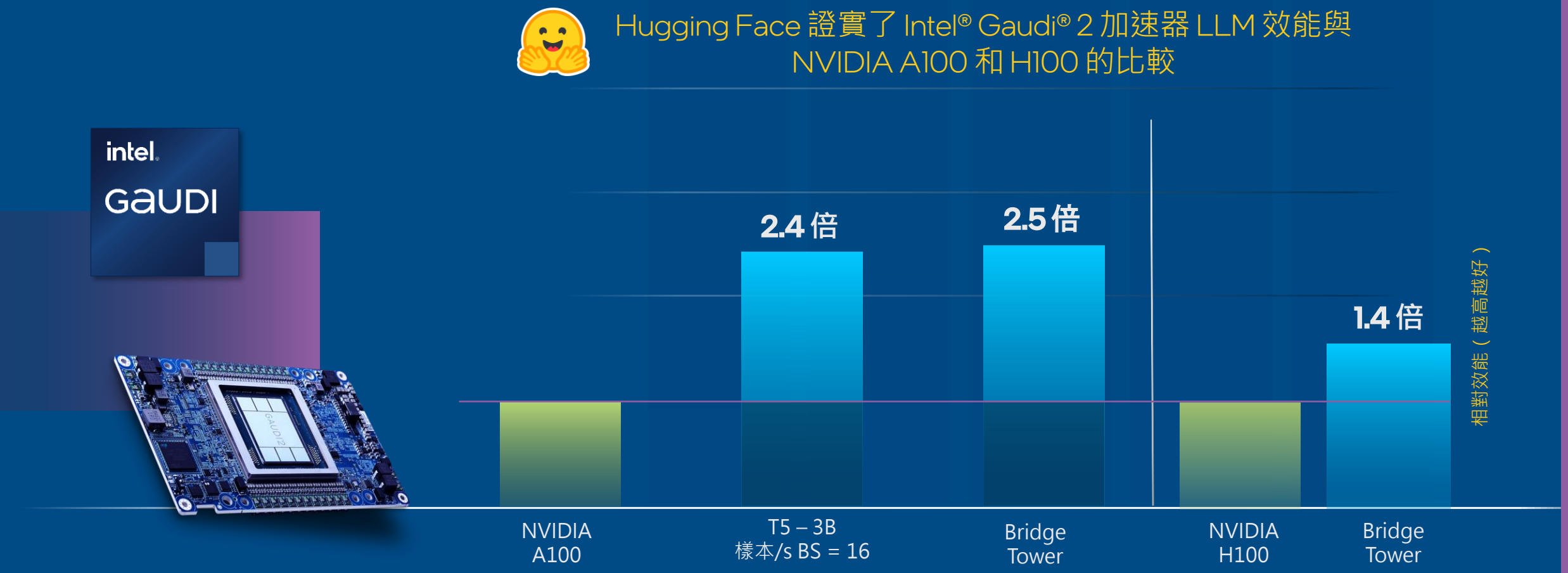
Gaudi2：非常適合高效訓練與基礎模型推斷

Gaudi2 的設計旨在實現深度學習效能、效率和可擴充性，
以滿足 LLM（GPT）和 GAI（Stable Diffusion）等大型基礎模型的需求

需求	Gaudi2
速度	訓練與推斷速度比 A100 **快 1.5-2 倍**
記憶體	每部 Gaudi2 裝置配備 **96 GB 晶片上高頻寬記憶體**，更易於在記憶體中容納大型基礎模型，並大規模訓練及部署
擴充性	透過**晶片上整合的 24 個 100GbE 連接埠**、伺服器中 8 卡之間的直接全面連線，以及伺服器內和跨伺服器開放式的 ROCEv2 通訊來提升效率
易於使用	使用 SynapseAI、PyTorch 和 DeepSpeed，**以最小的程式碼變更**來移轉或建置模型
電源效率	**相較於 A100，輸送量/瓦數提升至 1.8 倍**
成本效益	以專門打造的第 1 代 Gaudi 架構為基礎，其性價比在 Amazon 雲端的 A100 **提升高達 40%**

資料來源：https://habana.ai/habana-claims-validation/

對眾多 LLM 進行微調



請造訪 https://habana.ai/habana-claims-validation 瞭解工作負載和配置。結果可能會有所落差。

https://huggingface.co/blog/habana-gaudi-2-benchmark
https://huggingface.co/blog/bridgetower

GPT-J：Intel® Gaudi® 2 結果

GPT-J 的 Intel® Gaudi® 2 推斷效能結果為其競爭表現提供了強而有力的驗證

- GPT-J-99 和 GPT-J-99.9 上用於伺服器查詢和離線樣本的 Intel® Gaudi® 2 推斷效能分別為**每秒 78.58 和每秒 84.08**[1]
- 相較於 **NVIDIA 的 H100，Intel® Gaudi® 2 提供令人信服的效能**，相較於 Gaudi 2，H100 表現出 1.09 倍（伺服器）和 1.28 倍（離線）效能的微幅優勢[1]
- **Intel® Gaudi® 2 的效能比 NVIDIA 的 A100 高出了 2.4 倍（伺服器）和 2 倍（離線）**[1]
- Intel® Gaudi® 2 提交採用了 FP8，並且在這個全新的資料類型上達到 **99.9% 的準確性**[1]

閱讀更多

透過每六至八週發布的 Intel® Gaudi® 2 軟體更新，Intel® 預計會持續提供效能提升，並擴大 MLPerf 評測基準的模型覆蓋範圍

新聞室文章

MLCommons 公告

[1]效能因使用、配置和其他因素而異；工作負載和配置詳細資料請見：https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/結果可能會有所差異。

Intel® Gaudi® 2：評測基準結果

Supermicro 提供的
評測基準結果；
業界首款 Intel® Gaudi® 2
OEM

Gaudi 聲明驗證

利用 Intel® Gaudi® 2
AI 加速器的
LLM 訓練與推斷

評測基準

更快的訓練與推斷：
Intel® Gaudi® 2 與
NVIDIA A100 80 GB

評測基準

結果可能會有所落差。

Intel® Gaudi® 2：基礎模型訓練與推斷

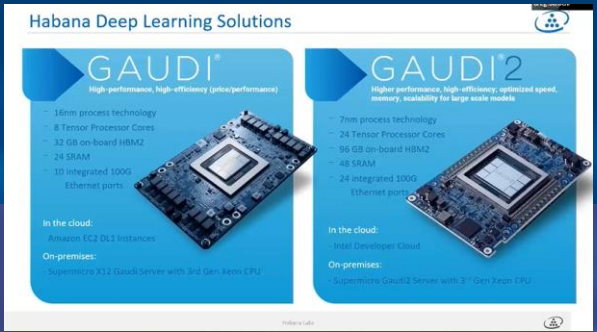
可用的 Gaudi 支援模型可從以下位置存取：

開發者目錄

GAUDI®2

Intel® Gaudi® 開發者訓練

[入門：Gaudi 的深度學習與推斷]



[將 Intel® Gaudi® 2 的能力發揮到極致：加速生成式 AI 與大型語言模型]

[利用 Intel® Gaudi® 處理器將模型效能發揮到極致：實現最佳結果的進階工具與策略]

Intel® Gaudi® 軟體（SynapseAI® 軟體套件）

簡化開發：以您想要的方式開發

目標：輕鬆將現有軟體移轉至 Intel® Gaudi® AI 加速器、保留軟體投資，並讓建構新模型變得容易：無論是對定義深度學習、生成式 AI 和大型語言模型的眾多且有增無減的模型進行訓練還是部署。為資料科學家、開發者、和 IT 與系統管理員提供以下支援：

- 開發者網站
- GitHub

Intel® Gaudi® AI 加速器

深度學習的軟體生態系統匯集了領先的軟體供應商、工具與程式碼，加速根據 PyTorch、TensorFlow、PyTorch Lightning 和 DeepSpeed 架構的最先進深度學習模型的開發

cnvrg.io

PyTorch Lightning

準備好使用 Intel® Gaudi® 軟體了嗎？

Intel® Gaudi® 2 AI 加速器現已推出！只在 Denvr Cloud

Intel® Gaudi® 2
軟體生態系統

Intel® Gaudi® 2
AI 加速器（7nm）

Intel® Gaudi® 2：非常適合生成式 AI 的需求

- 現在推出！Denvr Cloud 上的 Gaudi 2 叢集
- 測試最多 8 個 Gaudi 2 節點
- Intel 客戶的優先 VIP 定價
- Denvr Dataworks 高接觸式商業服務與支援
- 在 Denvr Cloud 上無縫移轉至 Gaudi 2 叢集
- Denvr Cloud 上的 Gaudi 3 叢集的獨家優先定位即將推出！

立即開始

即將推出

intel GAUDI

訓練推斷

Intel® Gaudi® 3

Intel® Gaudi® 3 加速器憑藉效能、可擴充性與效率，向更多客戶提供更多選擇，協助企業獲得洞察力、創新技術，並增加收入

即將推出 - Intel® Gaudi® 3 AI 加速器

為 GenAI 帶來效能、可擴充性與效率的選擇

intel. GAUDI

Intel® Gaudi® 3 將為期望大規模部署 GenAI 的全球企業帶來 AI 訓練與推斷的重大進展

新聞稿

Intel® Gaudi® 3 加速器效能相較於 NVIDIA H100

Intel® Gaudi® 3 預計可將具有 7B 和 13B 參數的 Llama2 模型和 GPT-3 175B 參數模型的訓練

時間平均提升 50%[3]

Intel® Gaudi® 3 預計可在以下方面超越 H100：

加速器推斷輸送量提升 50%[1]

推斷電源效率提高 40%[2]

跨 Llama 7B 和 70B 參數，以及 Falcon 180B 參數模型

閱讀更多

白皮書

Intel® Gaudi® 3 將從 2024 年第 2 季起向 OEM 提供，包括：

Lenovo DELL Hewlett Packard Enterprise SUPERMICRO

[1]NV H100 的比較根據為 https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8，報告的數字是針對每個 GPU。與 Intel® Gaudi® 3 對 LLAMA2-7B、LLAMA2-70B 與 Falcon 180B 的預測相比。結果可能會有所落差。

[2]NV H100 的比較根據為 https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8，報告的數字是針對每個 GPU。與 Intel® Gaudi® 3 對 LLAMA2-7B、LLAMA2-70B 與 Falcon 180B 的預測相比。NVIDIA 和 Gaudi 3 的電源效率是根據內部估計。結果可能會有所落差

[3]NV H100 的比較根據為：https://developer.nvidia.com/deep-learning-performance-training-inference/training，截至 2024 年 3 月 28 日，「大型語言模型」分頁相較於 LLAMA2-7B、LLAMA2-13B 和 GPT3-175B 的 Intel® Gaudi® 3 預測。結果可能會有所落差。

intel.

適用於 NLP/LLM 的 Intel® 產品

推斷

第 4 代和第 5 代 Intel® Xeon® 可擴充處理器利用 Intel® DL Boost、
Intel® AMX 和 Intel® AVX-512 加速 NLP。它專為高效能運算設計，可用於加速 NLP 工作負載。它們可以處理大量執行緒、大型記憶體容量和高記憶體頻寬，適合語言翻譯、文字摘要和文字轉語音等 NLP 工作負載。

第 5 代 Intel® Xeon®：專為 AI 設計的處理器

透過在每個核心加速 AI，第 5 代 Intel® Xeon® 處理器可在客戶必須添加獨立加速器之前，解決要求嚴苛的端對端 AI 工作負載

AI 推斷的效能更高

高達 42%

相較於前一代[1]

一般運算效能提升

平均 21%

相較於前一代[1]

更快的自然語言處理

高達 23%

相較於前一代[1]

Intel 執行副總裁暨資料中心與 AI 事業群總經理 Sandra Rivera

「我們專為 AI 設計的第 5 代 Intel® Xeon® 處理器為客戶在雲端、網路和邊緣使用案例中部署 AI 功能，提供更卓越的效能。經過與客戶、合作夥伴和開發者生態系統長期合作，我們在經驗證的基礎上推出第 5 代 Intel® Xeon®，以較低的總體擁有成本實現快速採用與擴展。」

詳細資訊

網站

產品簡介

關於工作負載與配置，請造訪 intel.com/processorclaims：第 5 代 Intel® Xeon® 可擴充處理器。結果可能會有所落差。

Intel® Xeon®：
實際 AI 應用程式的 CPU 效能領導地位

在實際工作應用程式中，**Intel** 正透過以下方式為 **AI** 推斷提供更優異的效能、更低的價格和更平衡的平台，

進而顛覆產業並推動 **AI** 普及化：

- 較大的快取記憶體有助於資料局部性，且大型記憶體容量可解決較大的問題
- 更高的核心頻率、多個純量連接埠和亂序執行，有助於加速單執行緒或多執行緒但純量的運算
- Intel® Advanced Vector Extensions 512（Intel® AVX-512）協助非 DL 向量運算
- Intel® Advanced Matrix Extensions（Intel® AMX）是對 AI 加速的內建硬體支援

完整技術文章

資訊圖表

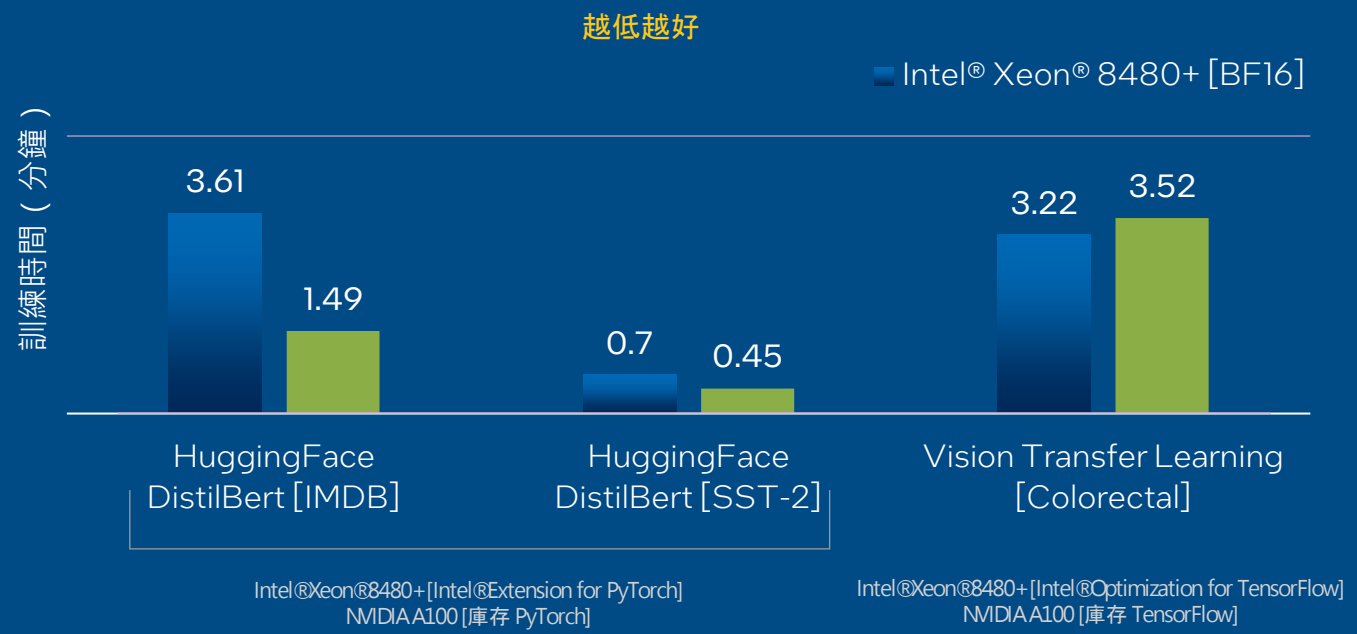
intel XEON

揭穿 GPU 的迷思：搭載內建加速器的 CPU 如何徹底改變 AI

利用 Intel® Xeon® 可擴充處理器即可在 4 分鐘內微調模型[1]

Hugging Face

微調訓練效能的時間 Intel® Xeon® Platinum 8480+ 處理器相較於 NVIDIA A100 GPU



亦請參閱：

更優異的效能：Numenta 搭載 Intel® CPU 與 NVIDIA GPU 的區別

[1]請參閱第 4 代 Intel Xeon 可擴充處理器的效能指數 [A221]。結果可能會有所落差。

第 4 代 Intel® Xeon® 處理器上的 LLMS

人工智慧（AI）聊天機器人技術是一種與客戶互動並改善客服的方式，在各個企業和組織中愈來愈受歡迎，但針對特定使用案例建構、最佳化及維護聊天機器人的成本高昂，對許多組織而言可能在財務上望之卻步

詳細資訊

第 4 代 Intel® Xeon® 可擴充處理器的 AI 微調指南

連結至指南 ›

第 4 代 Intel® Xeon® 處理器透過 **Intel® Advanced Matrix Extensions（AMX）** 提供改良的資料管理和高效運算，當與 Intel® Extension for PyTorch 提供的**自動混合精度（AMP）** 功能搭配時，這種技術堆疊對於轉移學習和從頭開始訓練小型/中型模型等工作負載變得相當具有競爭力

操作技術文章

搭載第 5 代和第 4 代 Intel® Xeon® 處理器的 Cisco UCS，適用於生成式 AI

小即是美：Q8-Chat LLM 是 Intel® Xeon® 處理器上的高效生成式 AI 體驗

LLM 需要大量的運算能力（通常由高階 GPU 提供）以夠快的速度預測搜尋或對話應用程式等低延遲使用案例。可惜的是，對於許多組織而言，相關的成本可能令人望之卻步，因此難以在應用程式中使用最先進的 LLM。

瞭解有助於減少 LLM 大小和推斷延遲的最佳化技術，協助它們在 Intel® CPU 上高效執行。

操作技術文章 ›

Hugging Face

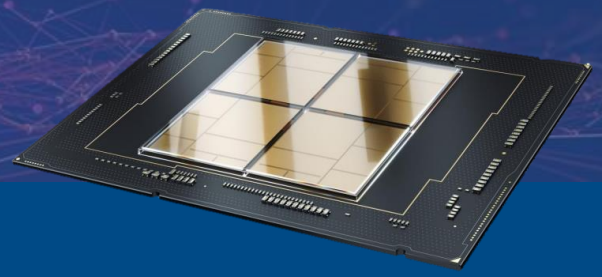
「更多公司會更注重於小型的特定模型，其在訓練及執行方面的成本更低。」

利用 Hugging Face 立即開始使用第 4 代 Intel® Xeon®

適用於 LLM 的 Intel® Xeon® 處理器

總結

- 非常適合推斷專業領域的 LLM
- 提供轉移學習的使用案例
- 利用開放原始碼 SW 在 Intel® Xeon® 上部署 LLM，輕鬆提供最佳效能



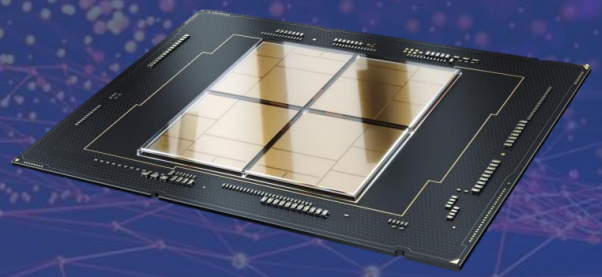
適用於 LLM 的 Intel® Xeon® 可擴充處理器

非常適合利用最熱門的 AI 架構和資料庫建構及部署通用 AI 工作負載

- 利用現有基礎架構進行特定領域的 LLM 推斷
- 提供轉移學習的使用案例
- 利用開放原始碼 SW 在 Intel® Xeon® 上部署 LLM，輕鬆提供最佳效能

Intel® Xeon®
實際 AI 應用的
CPU 效能領導地位

技術文章 ▪ 資訊圖表



GPT-J
第 4 代 Intel® Xeon® 處理器成果

2 離線模式的每秒段落[1]

新聞室文章

1 即時伺服器模式的每秒段落[1]

MLCommons 公告

揭穿 GPU 的迷思：搭載內建加速器的 CPU 如何徹底改變 AI

搭載第 4 代 Intel® Xeon® 和 Intel® AMX 的 Alibaba NLP 案例研究

閱讀更多

[1]效能因使用、配置和其他因素而異；工作負載和配置詳細資料請見：https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/結果可能會有所差異。

適用於 NLP/LLM 的 Intel® 產品

用戶端的小規模推斷

Intel® Core™ Ultra 引領 AI 電腦時代

Intel® Core™ Ultra 處理器專為頂級的超薄強大筆記型電腦最佳化，採用 3D 效能混合式架構、高階 AI 功能，並且內建 Intel® Arc™ GPU。Intel® Core™ Ultra 處理器採用全新的 Intel® 4 製程打造，針對外出時的遊戲體驗、內容創作和生產力，在效能與電源效率之間取得
最佳平衡。

使用案例：電腦上的 AI

創作者：
照片與影片搜尋及編輯

更快且更自然的濾鏡、
更高品質的預覽，以及更快的匯出時間和更快的自動搜尋。

協作/串流

新一代的視訊會議、串流和協作的新 AI 功能，維持電池續航力。

主流遊戲

遊戲內、3D 動畫的新 AI 功能，增加逼真度、可轉錄和翻譯聊天內容。

電腦上的 AI

「解鎖日常功用」

生產力

用於寫作、創作、編碼和離線功能（例如文字與語法預測）的 AI 助理。

輔助功能

AI 輔助的視聽功能滿足不同的使用者需求，在電腦上創作更輕鬆、生產力更高。

創作者：文字轉影像

只要輸入幾個描述性的詞語就能建立影像的新 AI 效果與功能（適用於行銷、廣告、設計）。

*Intel 2023 年 12 月 14 日，AI 無所不在活動新聞稿

適用於生成式 AI 的 Intel® Core™ Ultra

Intel 最節能的用戶端處理器引領 AI 電腦時代

效率與效能大幅提升

AI 效率

高達 **70%**

更快的生成式 AI 效能[2]

省電

高達 **25%**

減少耗電量[3]

閱讀更多　公告 • 產品簡介 • 網站

Intel® Core™ Ultra 採用 Intel 的首個用戶端晶片上的 AI 加速器（神經處理器或 NPU），比前一代**電源效率高出 2.5 倍**，實現全新節能的 AI 加速等級[1]

Intel® Core™ Ultra H 和 U 世代晶片均包含兩個適用於低強度工作負載的新型低功耗島（LP-E）核心，且 Intel AI NPU 內建有兩個神經運算引擎，旨在解決**生成式 AI 推斷**。

[1]以在 Intel® Core™ Ultra7 165H NPU 與 Intel® Core™ i7-1370P GPU 執行 int8 模型時 UL Procyon AI 評測基準上的每瓦效能測量。
[1、2、3]如欲瞭解工作負載與配置，請參閱 www.intel.com/PerformanceIndex。結果可能會有所落差。

加速 AI 創新

Intel® 正與領先的產業 ISV 合作，利用 AI 將您的體驗最佳化。

AI 電腦加速計畫的目標，是協助獨立硬體廠商（IHV）與獨立軟體廠商（ISV）取得各種 Intel® 資源，包括人工智慧（AI）工具鏈、訓練、協同設計、軟體最佳化、硬體、設計資源、技術專業知識、共同行銷，以及銷售機會。

進一步瞭解

利用 Intel® Tiber™ AI 雲端加速企業 AI 開發

在一系列最新的 Intel® 軟硬體上學習、設計原型、測試及執行應用程式與工作負載。

利用此開發環境的最新軟硬體創新**技術**，**加速及擴充 AI**。
獲得更多的運算能力，以及更多**微調軟體與生成式 AI** 的選擇。

利用 Intel 馬上開始

利用最新的 Intel 產品獲得實際操作體驗。利用 Intel 增強您的 AI 技能。

搶先取得技術

評估搶先版的 Intel 平台與相關的 Intel 最佳化軟體堆疊。

大規模部署 AI

利用 Intel 最新的機器學習工具組和 Intel® Tiber™ AI 雲端託管的程式庫加速 AI 部署。

閱讀技術文章 ›
開始使用 ›

行動方針

教育

瞭解 Intel®技術能如何應用於生成式 AI 與特定領域模型，以及 Intel®Xeon®和 Intel®Gaudi®產品線可協助您贏得更多業務的範疇

馬上開始

互動

開始使用

Intel® Tiber™ AI 雲端

利用此開發環境的最新軟硬體創新技術，加速及擴充 AI

&

使用 AI 參考工具組

聯絡人

請與您的 Intel® 代表聯絡以取得更多資訊

如何取得 Intel® 夥伴聯盟客戶支援

Intel Virtual Assistant

這個聊天機器人位於每個合作夥伴聯盟網頁右下角，為多數問題提供自助解答，或是即時支援代表的快速連結。

「取得協助」滿版廣告

提交線上支援要求。
這個連結位於合作夥伴聯盟網站多數網頁的頁尾。

夥伴聯盟「取得協助」頁面

取得協助頁面針對多數工具提供詳細的自助指南，詳述合作夥伴聯盟會員享有的福利。

AI 啟用區

數位優先的 AI 工作空間可策劃關鍵資源、工具與福利，激勵合作夥伴建立、行銷及銷售採用 Intel® 技術的解決方案

技術性協助

銷售與行銷支援

技術性協助

銷售與行銷支援

技術性協助

銷售與行銷支援

AI 參考工具組

利用這些參考工具組，組織可大幅縮短解決問題的時間並獲得實質性的效能提升

金融與保險

詐欺偵測

GitHub ▪ 部落格 ▪ 藍圖

醫療保健與生命科學

疾病防護

GitHub ▪ 部落格

製造業與公用事業

異常偵測

GitHub ▪ 部落格

車隊管理

可預測的維修

GitHub

程序自動化

文件自動化

GitHub ▪ 部落格 ▪ 藍圖

工作流程

- DL 轉移學習
- HF 微調與推斷最佳化
- DL-分散式壓縮
- 分散式分類 ML 工作流程
- 利用 Intel® 加速器的 DL 預先訓練
- 利用 DGL 與 PyG 的圖形分析與 GNN
- Big-DL 上的分散式訓練/推斷
- Ray 上的 LLM 預先訓練與微調

工具

- Intel® Distribution for Python
- Intel® Optimized Modin
- Intel® Optimized XGBoost
- Intel® Extension for Scikit-Learn
- Intel® Optimized Tensorflow (ITEX)
- Intel® Optimized PyTorch（庫存與 IPEX）
- Intel® Neural Compressor
- SigOpt Python SDK 與 CLI
- CNVRG Python SDK 與 CLI
- Intel-optimized Horovod
- DeepSpeed

領域工具組

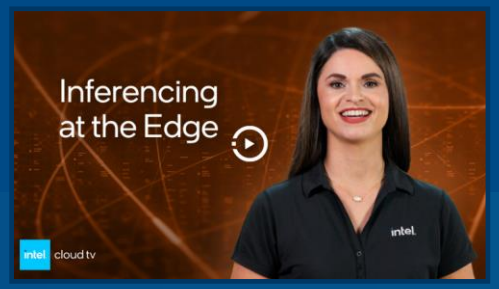
- 時間序列
- PPML
- 轉移學習
- Transformer/NLP

參考工具組以**容器**的形式提供，既可在**主要雲端**上使用，也可用於內部部署。
參考工具組以**工作流程**和**領域工具組**為基礎，可獨立利用，支援**多個產業的各種使用案例**。

Cloud TV

Intel® Cloud TV 探索雲端運算最新消息、趨勢與策略，助您馬到成功

利用 Intel® Gaudi® AI 加速器開創您的 GenAI 機會



利用邊緣資料推斷快速獲得深入解析

利用雲端 AI 創造競爭優勢

AI 推斷使用雲端技術

雲端 AI

邁上快速路徑到處擴充 AI

訓練

讓 AI 無所不在 -
生成式 AI 企業使用案例

生成式 AI 不僅適用於網際網路聊天機器人。無數企業都在考慮利用生成式 AI 和大型語言模型的威力來協助日常營運。本次會議將進一步探索企業生成式 AI 的使用案例，並提供多項考量因素，以便貴組織將其應用於日常營運。

註冊 ›

簡化資料生成與
大型語言模型的 AI

將 AI 整合至組織的工作負載或擴充現有的基礎架構，是一項技能繁重且運算密集的工作，需要開發基於大型資料集訓練的強大模型，以及強大的 GPU 才能充分執行。並非每個組織都有完成這項任務的必要資源。

本次會議聚焦於解決方案：來自 Accenture* 和 Intel® 的一系列開放原始碼 AI 參考工具組，旨在讓組織更易於取得 AI，並針對改善訓練和推斷時間進行最佳化。

其他訓練

技術性

素材類型	標題與連結
技能	雲端技能中的 AI
網路研討會	利用 Hugging Face 最佳化 Intel® 硬體的 AI
網路研討會	如何設定雲端型分散式訓練，以便微調 LLM
訓練課程	利用提示節約與情境學習改善 LLM
訓練課程	簡化資料生成與大型語言模型的 AI
訓練課程	自然語言處理
訓練課程	使用 TensorFlow* 應用深度學習
訓練課程	小巧靈活：使用企業 GenAI 的捷徑
訓練課程	GenAI 的新一波浪潮：特定領域的 LLM
指南	生成式 AI 的開發者入門指南：具體案例的應用方法
訓練課程	將 Intel® Xeon® 處理器的 AI 引進解決方案空間

其他訓練

非技術性

素材類型	標題與連結
影片系列	迎接生成式 AI
訓練課程	小巧靈活：使用企業 GenAI 的捷徑
訓練課程	GenAI 的新一波浪潮：特定領域的 LLM
訓練課程	應對 AI 無所不在
訓練課程	了解 AI 軟體與生態系統
訓練課程	參與 AI 生態系統：以軟體取勝、利用 SI 擴充並銷售解決方案
訓練課程	適用於現實世界的生成式 AI 與大型語言模型

其他資源

素材類型	標題與連結
網路研討會	生成式 AI 網路研討會系列
網路研討會	讓 GenAI 無所不在
Podcast	Copilot、ChatGPT、Stable Diffusion 與生成式 AI 將如何改變我們開發、工作與生活的方式
企業簡介	將人工智慧部署到各處
部落格系列	利用第 4 代 Intel Xeon 處理器進行生成式 AI 的微調與推斷
解決方案簡介	利用 Lenovo ThinkSystem SR650 V3 /第 4 代 Intel Xeon 處理器部署及擴充生成式 AI 推斷 全新的 Intel 與 VMware 技術為 Lenovo ThinkAgile VX V3 系統注入強勁動能
技術文章	利用 Intel® AI 軟硬體最佳化加速 Llama 2
研究 PR	10% 的組織在 2023 年將 GenAI 解決方案投入生產
爐邊談話影片	因應生成式 AI 的運算與永續發展挑戰
Podcast	Hugging Face 與 Intel - 推動實用、更快速、普及化且符合倫理的 AI 解決方案
Twitter / X 對話	大型語言模型如何推動 AI 開發
Supermicro 評測基準	Habana 聲明驗證
Hugging Face 評測基準	效能標竿
訓練/網路研討會	雲端解決方案架構師（CSA）技術講座：AI 與 Habana
白皮書	企業 AI 與開發者白皮書息息相關
資訊圖表	CPU 是企業 AI 的關鍵

其他資源

素材類型	標題與連結
解決方案簡介	使用 Intel Enterprise AI 和 Red Hat® OpenShift® AI 簡化 AI 採用與部署流程
指南	AI 指南
參考工具組	AI 非結構化文字資料生成
白皮書	Zoho 正在最佳化及加速影片 AI 工作負載
白皮書	Seekr 開發值得信賴的 AI 篩選系統
解決方案簡介	教育領域中的安全機制：AI 與機密運算有助於實現安全的遠端考試
案例研究與影片	Nature Fresh Farms 從種子到商店皆利用 AI
案例研究	QMed Asia 推動早期癌症偵測率
案例研究與影片	MetaApp 改良基於 AI 的推薦系統
解決方案簡介	最佳化自動光學檢測（AOI）的 AI 模型訓練與改良
部落格	LLM 的提示驅動效率


intel