Red Hat AI

打造 AI Ready 智慧校園

以 AI、容器與虛擬化建構
新世代校園數位平台



Ato Lin
Sr. Solution Architect
Red Hat

Red Hat

AI 素養，已是大學的正式課題

教學端的 AI 政策已明確，資訊處需要能承接的平台



未來人才 AI 能力圖像

教育部 2026/05/22 公布，
作為各級學校 AI 教育的政策藍圖，
定義四大 AI 能力維度



人工智慧基本法

2026/01/14 公布施行，
國科會為中央主管機關

資料來源：教育部新聞稿（2026/05/22）；全國法規資料庫《人工智慧基本法》

Red Hat

資料不出校園，AI 才能放心用

公立學校得準用《行政院及所屬機關（構）使用生成式 AI 參考指引》

不得輸入的資訊

不得向生成式 AI 提供應保密、個人及未經同意公開之資訊
（第 4 點）

地端部署的但書

封閉式地端部署模型，確認環境安全後，得依機密等級分級使用
（第 4 點）

適用範圍

公立學校得準用；其他機關得參照訂定規範
（第 9、10 點）

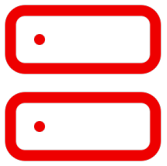
公有雲 AI 服務	校內地端 AI
公開課程資訊	學籍與成績資料
招生與活動文宣	輔導相關紀錄
非機密文書草稿	未公開研究資料

資料來源：行政院 2023/10/03 函頒（院授科會前字第1120059686號）；草案於 2023/08/31 行政院院會報告

Red Hat

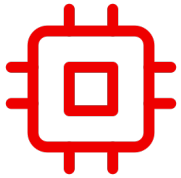
資訊處的三個現實挑戰

問題不在缺工具，而在平台太分散



虛擬化平台的下一步

續約成本與授權模式變動，
促使學校重新評估底層平台



GPU 算力孤島

各系所各自採購 GPU，
使用率不均，難以共用與計費



人力有限、平台太多

VM、容器、AI 各用一套工具，
維運人力難以負荷

Red Hat

一個平台，
承載三代工作負載

從既有 VM、容器化應用，到 AI 模型服務，
用同一套方式管理。

Red Hat

AI Ready 校園平台：分層架構

從「各系自建小教室」走向「一棟共用的教學大樓」

校園服務	校務系統	教學環境	研究平台	行政 AI 助理
工作負載	虛擬機器（VM）	容器化應用	AI 模型推論	
統一平台層	**統一排程 · 權限 · 監控 · 治理**			
既有基礎設施	伺服器	SAN / NAS 儲存	GPU	網路

資料來源：《人工智慧基本法》（114 年 12 月 23 日三讀通過、115 年 1 月 14 日公布施行）；國家科學及技術委員會公開資訊

Red Hat

AI Ready 平台的四個設計原則

選平台之前，先思考這四件事



開放標準

以 Kubernetes 等
開放標準為基礎，
避免廠商綁定



資料主權

敏感資料留在校內，
同時保留混合雲的彈性



統一維運

VM、容器與 AI
用同一套工具管理



治理內建

權限、用量與安全護欄
從一開始就設計

Red Hat

從虛擬化到 AI，一步一步來

Red Hat OpenShift ・
OpenShift Virtualization ・
Red Hat AI

Red Hat

虛擬化的三條路徑

不必一次全搬：依系統特性分流

1 Refactor

VMware vSphere® → Cloud Native Application

從 vSphere 虛擬主機到雲端原生（容器、無伺服器）應用程式架構的新變化

2 Replatform

VMware vSphere® → Alternative Virtualization

在保持虛擬化技術的同時，虛擬機器進程被遷移到比VMware「更便宜的位置」。

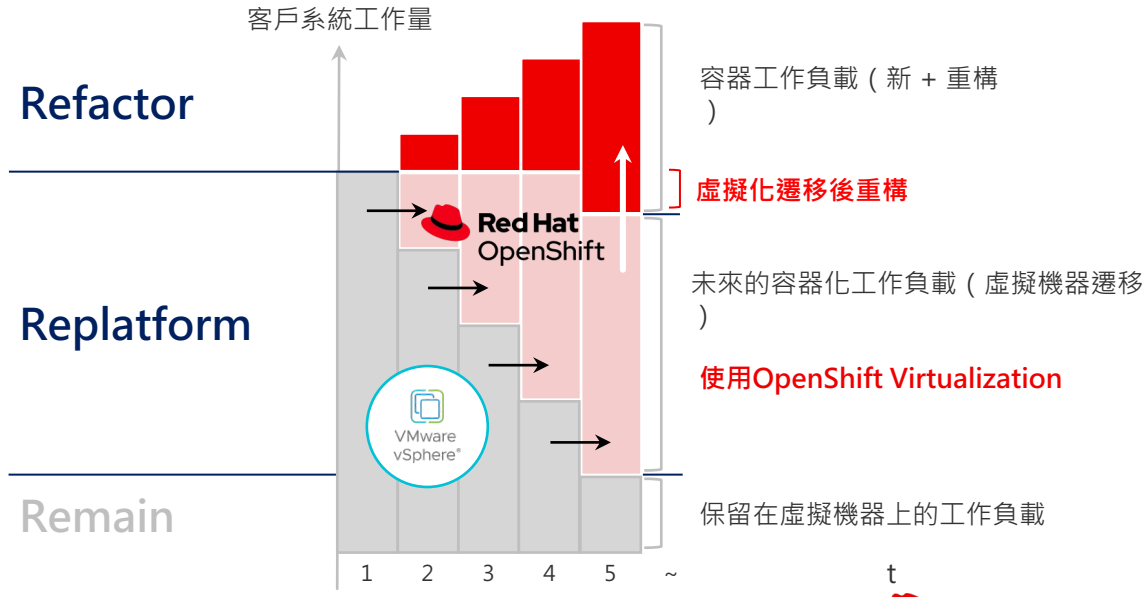
3 Remain

VMware vSphere® → VMware vSphere®

向 VMware 支付 vSphere (ESXi) 維護費用並繼續照常運行虛擬機

Red Hat

混合平台重構(Refactor)與遷移虛擬機(Replatform)



加速容器採用，同時減少對 VMware 的依賴

Red Hat

第一步：虛擬化現代化

Red Hat OpenShift Virtualization

成熟的 KVM 技術堆疊

KVM + QEMU + libvirt，KVM
自 2007 年即為 Linux 核心的一部分

MTV 遷移工具

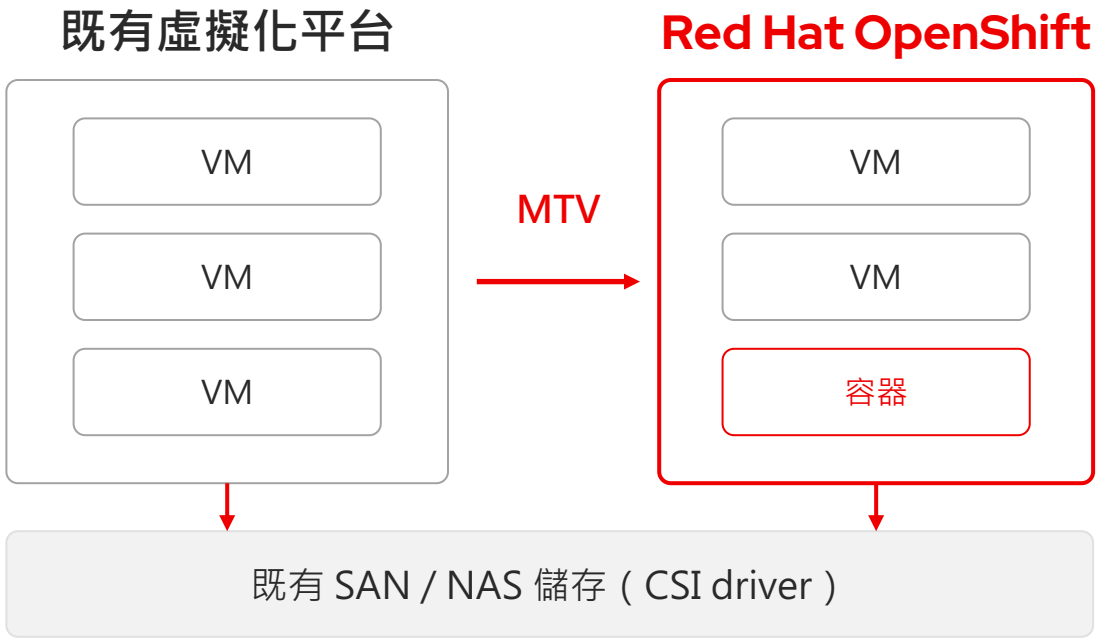
支援 vSphere、RHV、OpenStack、KVM、OVA 等來源，
包含在 OpenShift 訂閱中

沿用既有儲存

既有 SAN / NAS 透過 CSI driver 接入，保護既有投資

VM 與容器同平台

遷移後可在同一平台逐步現代化應用



Red Hat

第二步：校務與教學環境容器化

Red Hat OpenShift



校務系統

選課、註冊尖峰自動擴充，
離峰自動縮回資源



教學實驗環境

CUDA、PyTorch 等預建環境自助開啟，可指派 GPU，閒置自動停止



研究平台

以 CI/CD 與 GitOps 管理環境，
研究成果可重現

Red Hat

第三步：Red Hat AI Enterprise

整合 OpenShift + OpenShift AI 的全堆疊 AI 平台

最佳化 AI 使用成本

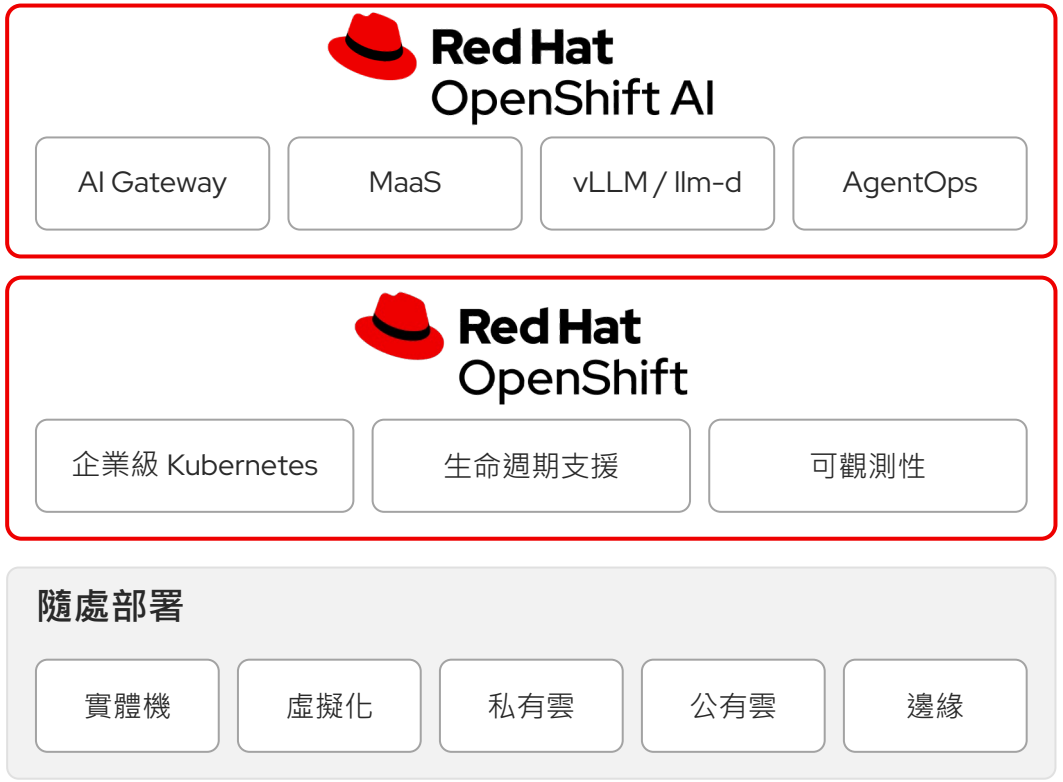
依用途靈活搭配雲端 API 與校內自建模型

一致的開發與維運

容器化、標準化的 AI 應用開發與維運環境

開放的技術選擇

自由選擇模型、加速器與執行環境，不綁定特定廠商



Red Hat

校內 AI 總機：Models-as-a-Service

一個 AI Gateway，集中管理校內自建模型與雲端模型

師生 | AI 代理 | 校務應用

↓

AI Gateway

驗證授權 · 訂閱額度 · 速率限制 · 政策套用

↓

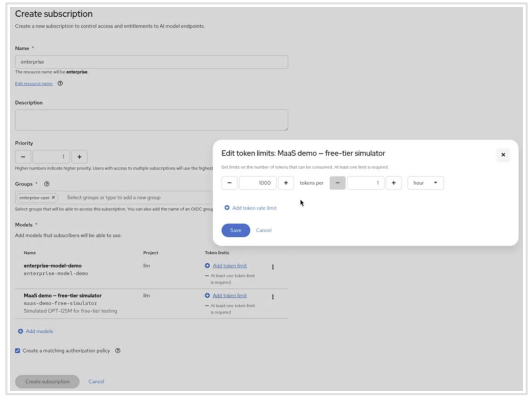
校內自建模型

vLLM ｜ 成本與隱私

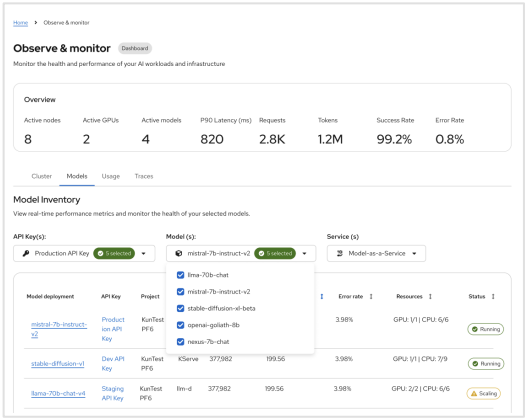
雲端 API 模型

外部模型轉接

價值：套用校內 AI 使用政策，防範 Shadow AI



額度設定：依訂閱設定 token 上限



觀測與監控：token 用量

資料來源：《人工智慧基本法》（114 年 12 月 23 日三讀通過、115 年 1 月 14 日公布施行）；國家科學及技術委員會公開資訊

Red Hat

GPU-as-a-Service：讓 GPU 不再閒置

以 Kueue 排程，把分散的 GPU 變成全校共用資源池

Kueue 公平排程

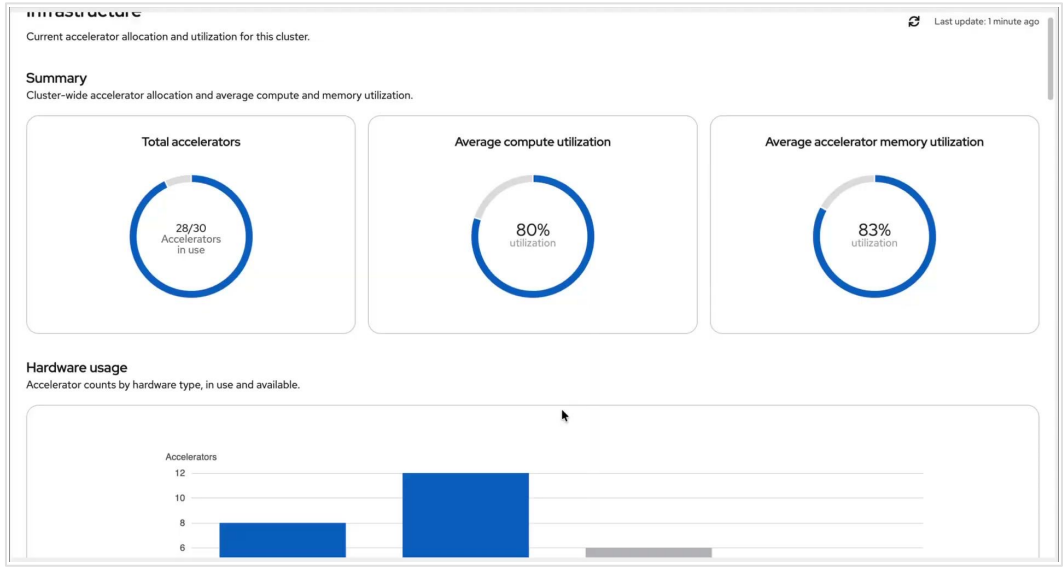
任務排隊取用 GPU，完成後自動釋放給下一個任務

資源池分級

依課程、研究計畫、行政、AI 服務設定配額與優先順序

GPU 使用視覺化

從叢集、硬體類型、佇列與借用狀況檢視使用率



GPU as a Service View (OpenShift AI 3.5)

Red Hat

壓縮模型：同樣能力，更少 GPU

Red Hat AI 提供經驗證的量化模型與 LLM Compressor 工具

-50%

權重記憶體

70B 模型由 FP16 改為 FP8，權重約由 140 GB 降至 70 GB



99%

準確度保留

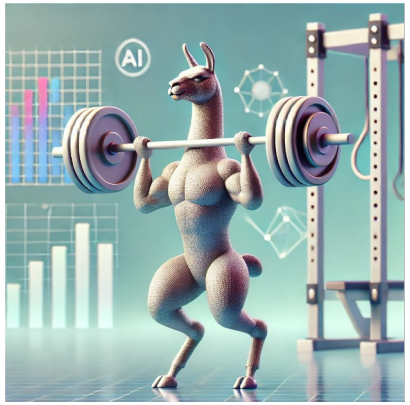
Llama 3 70B FP8 在 OpenLLM 基準相對原版的表現



2×

推論延遲改善

Llama 3 70B 於 2 張 H100 上，token 間延遲最高改善 2 倍



Red Hat

AI 治理：把護欄內建在平台

NeMo Guardrails 與 EvalHub：對應「地端分級使用」的管理需求



輸入護欄

阻擋個資輸入、
越獄攻擊與不當內容



輸出護欄

過濾個資外洩、
與事實不符及歧視性內容



上線前評測

EvalHub 以 Garak 模擬攻擊，
涵蓋 OWASP LLM Top 10

Red Hat

運行Agent，安全也很需要注意

在 OpenShift 上建構安全的 AI 代理執行環境

Gateway / Policy



- 以政策定義 AI 代理可執行的程序，以及可存取的檔案系統、網路與推論端點
- 建立存取 Sandbox 環境的 Gateway

Sandbox



- Kubernetes SIG Apps 正在開發的 Agent Sandbox 專案
- 透過 Kubernetes API，以 Custom Resource 形式建立 Stateful 的 Sandbox 環境
- 可設定 gVisor 或 Kata Containers 作為 Runtime Class

Runtime 隔離



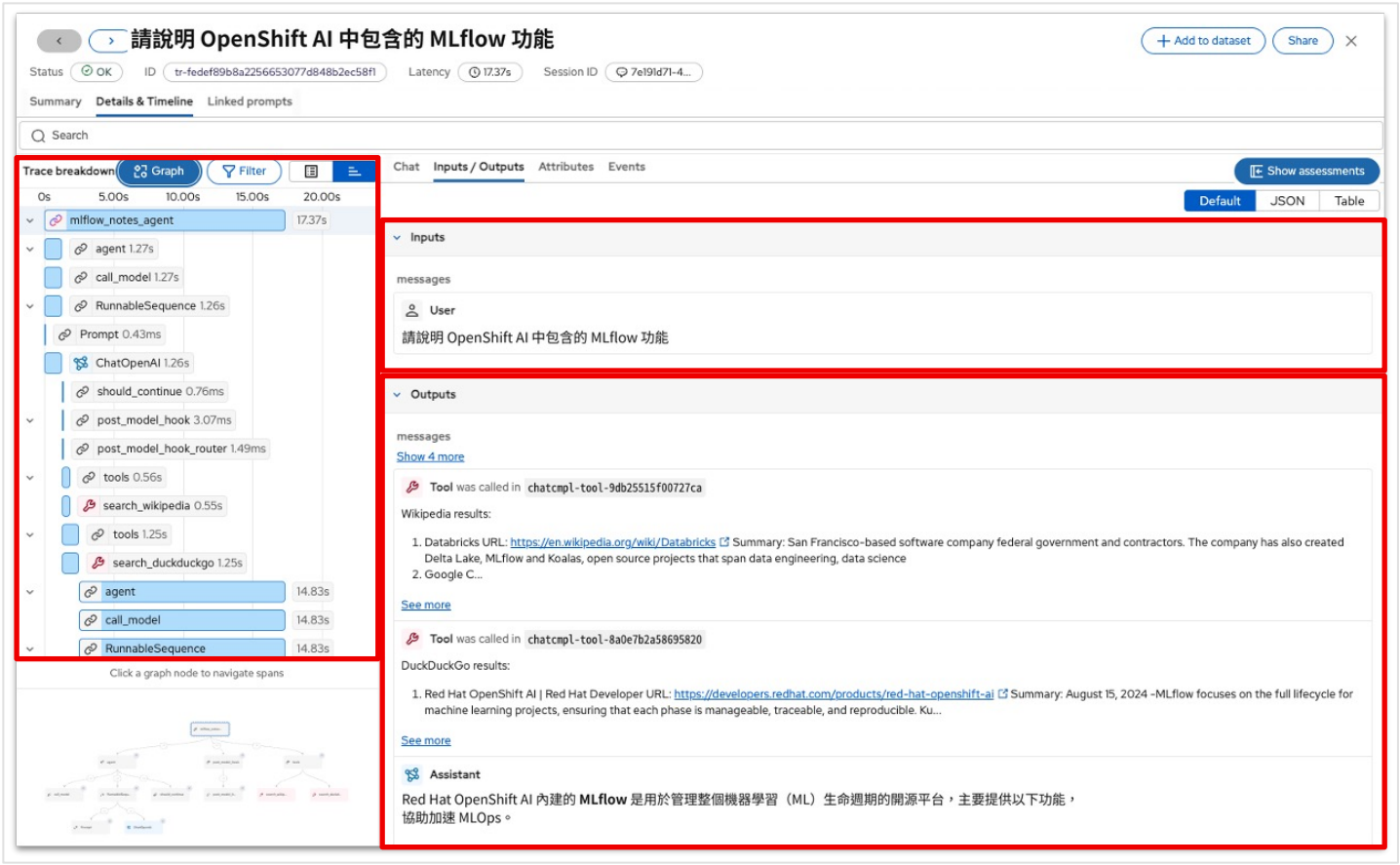
- 在 Kubernetes 上啟動 micro VM，並於其中執行 AI 代理容器
- 核心（Kernel）層級的安全隔離
- OpenShift 上的產品化版本：OpenShift sandboxed containers

Red Hat

MLflow 追蹤實例

一次提問的完整歷程：代理呼叫了哪些工具、模型回了什麼

應用程式
追蹤



使用者
輸入

輸出
（工具、模型）

Red Hat

案例：Boston University（美國）

同一套 OpenShift AI：2022 年公有雲起步，2024 年移至專屬叢集

指標	Before（2022 公有雲）	After（2024 專屬叢集）	變化
每位學生平台成本	約 US$150	→ **約 US$18**	**-88%** 每生成本
大量同時開啟 notebook	每個等待近 20 分鐘	→ **擴展性測試無延遲問題**	**消除** 啟動等待
課程容器映像檔	超過 10 GB	→ **每個基礎映像 < 2 GB**	**-80%+** 映像檔大小
使用規模	上線首月約 100 人	→ **2024 春 3 門課 610+ 人**	**6×** 學生人數

GPU 共享（2025）：以 Kueue 排隊共享 V100 / A100 / H100，任務完成自動釋放給下一位學生

資料來源：Red Hat Research Quarterly（2024/11、2025 秋季）；Red Hat Blog（2022/03/18）

Red Hat

三階段落地路線圖

從最有把握的地方開始

01 盤點與遷移

- → VM 盤點與相依性分析
- → MTV 試點遷移
- → 避開開學與選課尖峰

02 平台整合

- → 新系統採容器化開發
- → 教學環境自助化
- → 導入 GitOps 管理

03 AI 服務化

- → 上線校內 MaaS
- → 推出第一個 AI 助理
- → 建立治理與計量機制

資料來源：《人工智慧基本法》（114 年 12 月 23 日三讀通過、115 年 1 月 14 日公布施行）；國家科學及技術委員會公開資訊

Red Hat

The products that power Red Hat AI

Red Hat AI 能在採用 AI 的各個階段滿足其需求，提供模型、硬體與部署位置的選擇；無論是在地端、雲端或邊緣環境皆可部署。

Red Hat AI 的核心定位正是支援跨混合雲環境，以符合資料、合規與成本上的需求。

Red Hat AI Enterprise

For enterprise AI, deploy and scale efficiently, anywhere

Red Hat OpenShift AI

For the distributed deployments in OpenShift

Red Hat AI Inference

For optimized inference of large language models

Red Hat

Thank you

Red Hat is the world’ s leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

Red Hat