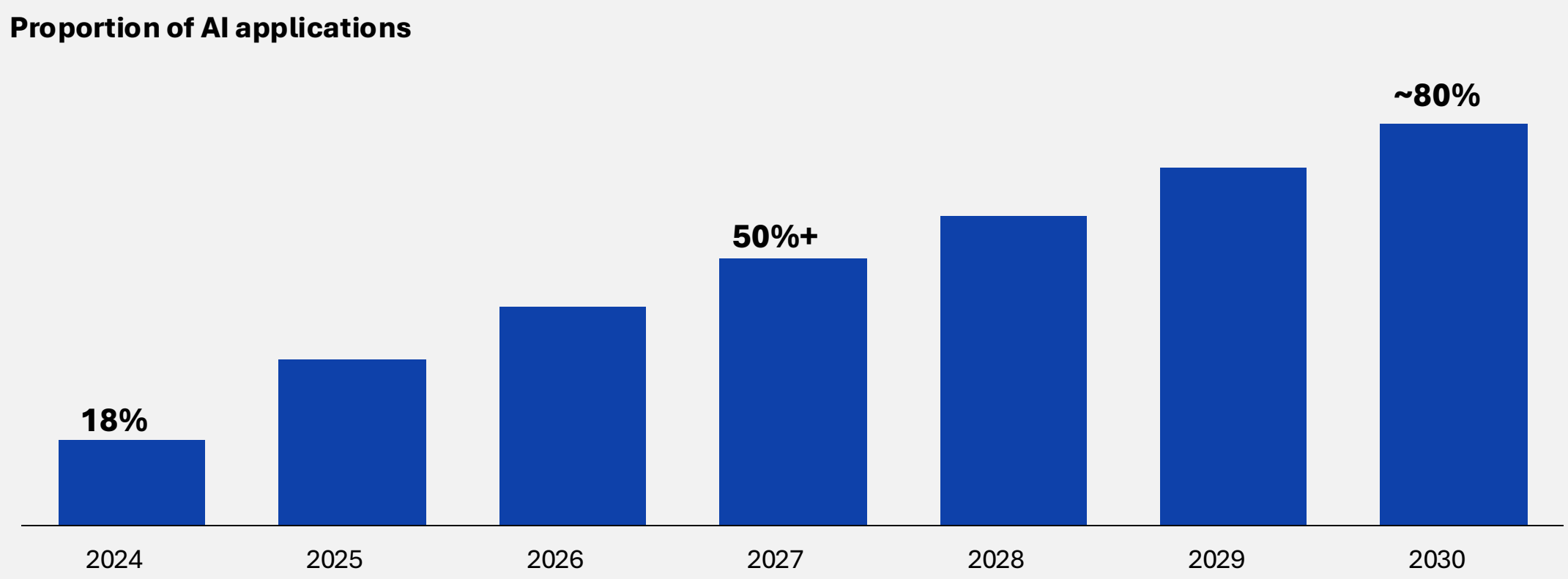
構建未來AI應用

從模型到基礎設施的全方位防護策略

“AI workloads are the most modern of modern apps”

AI應用是最現代的現代應用

未來幾乎每個應用都會成為 AI 應用



Source: IDC, Gartner, and F5 Corporate Strategy

GenAI 應用將通過四個屬性來區分：多模態、分散式、以數據為中心和 API 密集

1

GenAI應用體驗將是 multi-modal
多模態

2

GenAI應用將是 highly decomposed

3

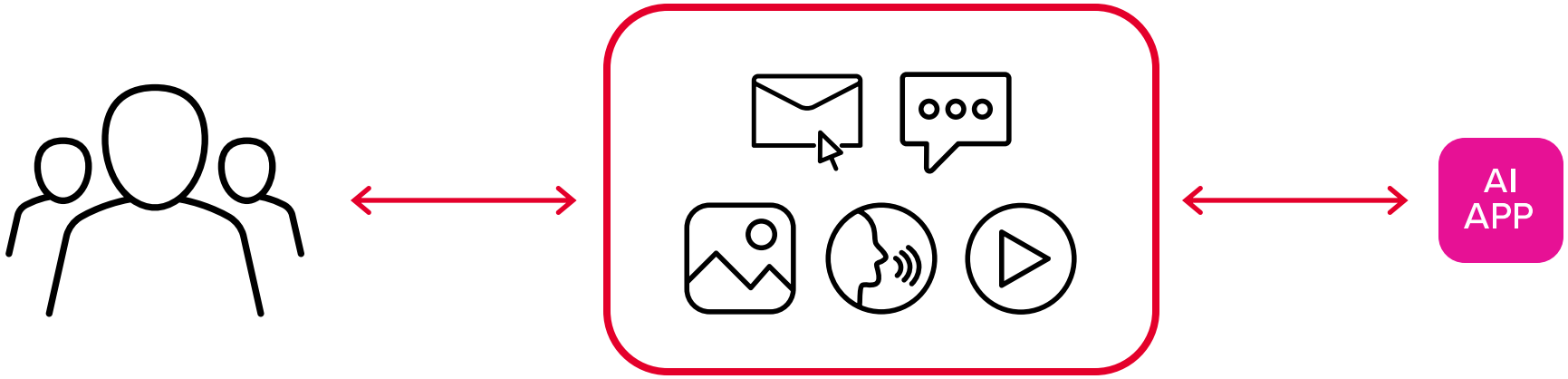
“Data gravity“
將顯著影響應用和模型的位置

4

GenAI應用將特別
倚賴於 API

GenAI 應用體驗將是跨視覺、文字、圖像、音訊和影片的多模態形式

1



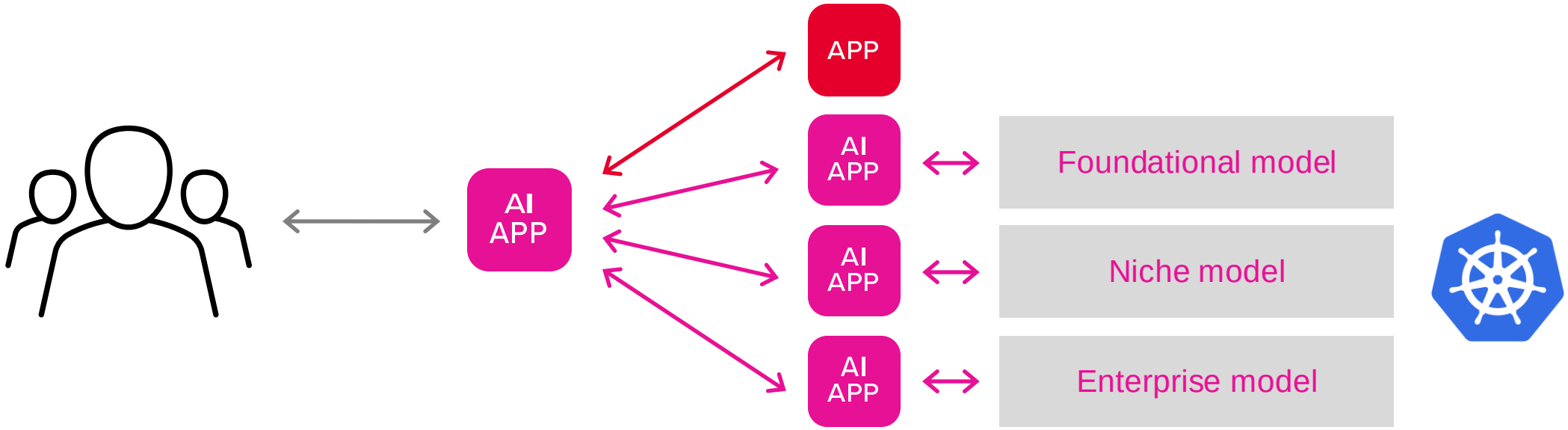
用戶

視覺、文本、圖像
、音訊或視訊

Multi-modal指可以跨不同類型的數據模態（如文本、圖像、音訊和視頻）處理和生成內容的系統或模型。這意味著多模態 GenAI 可以理解和生成多種形式的內容，從而實現更豐富、更通用的交互和輸出。

GenAI 應用將使用多個元件構建，包括多個 AI 模型

2

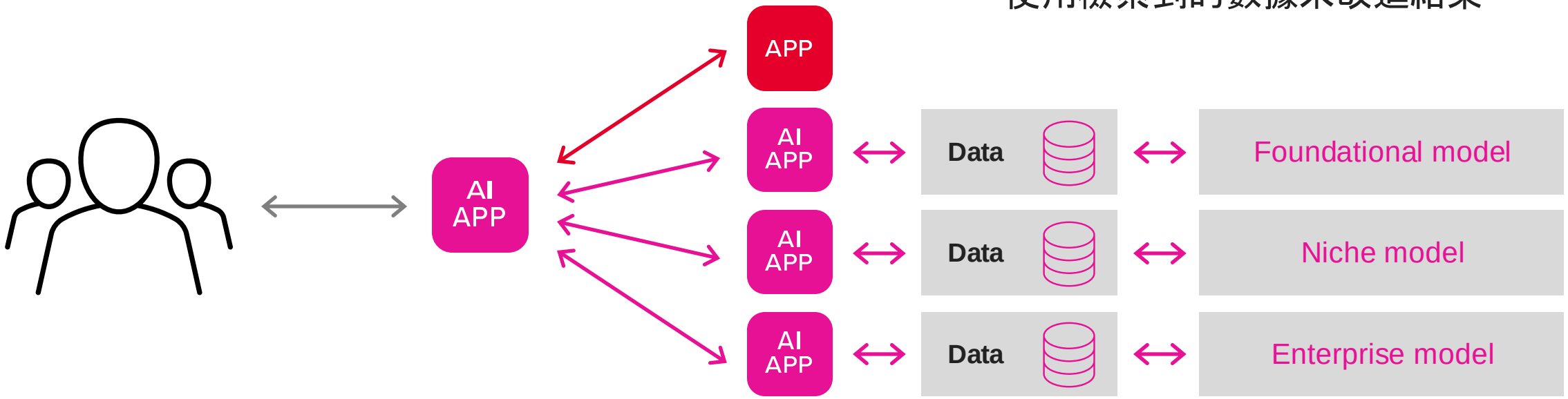


Decomposed指將複雜的任務或模型分解為更小、更易於管理的元件。這種方法簡化了問題解決過程，並且可以通過允許單獨處理或優化每個元件來提高效率和性能。

這些模型中的每一個通常都依賴於數據

3

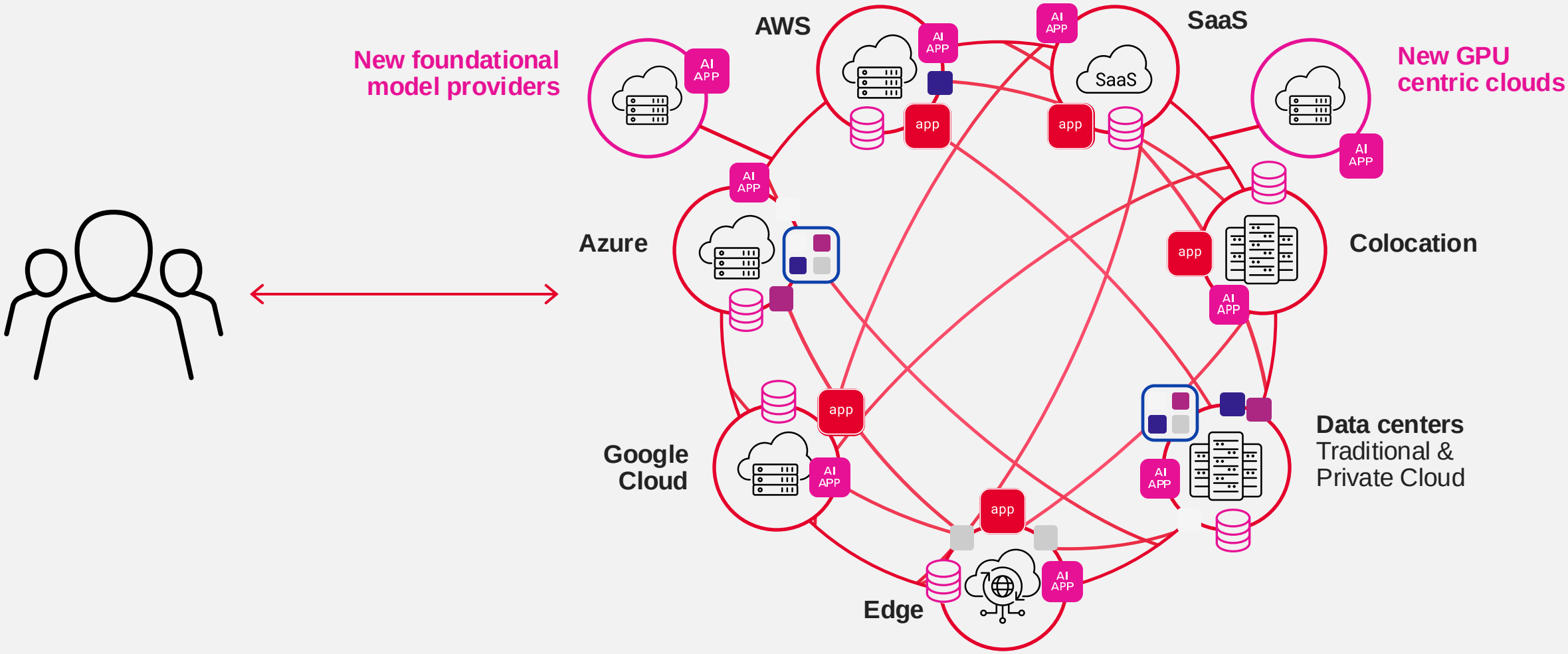
檢索增強生成 （RAG） 等模式
使用檢索到的數據來改進結果



Data Gravity指對於訓練和提煉人工智慧模型至關重要的大數據集吸引相關數據、服務和計算資源的趨勢。隨著數據量的增長，將 AI 應用程式和處理工具放置在這些數據附近以減少延遲、增強性能並提高數據處理效率變得更加高效。這一概念強調了數據局部性在GenAI系統的部署和運行中的重要性。

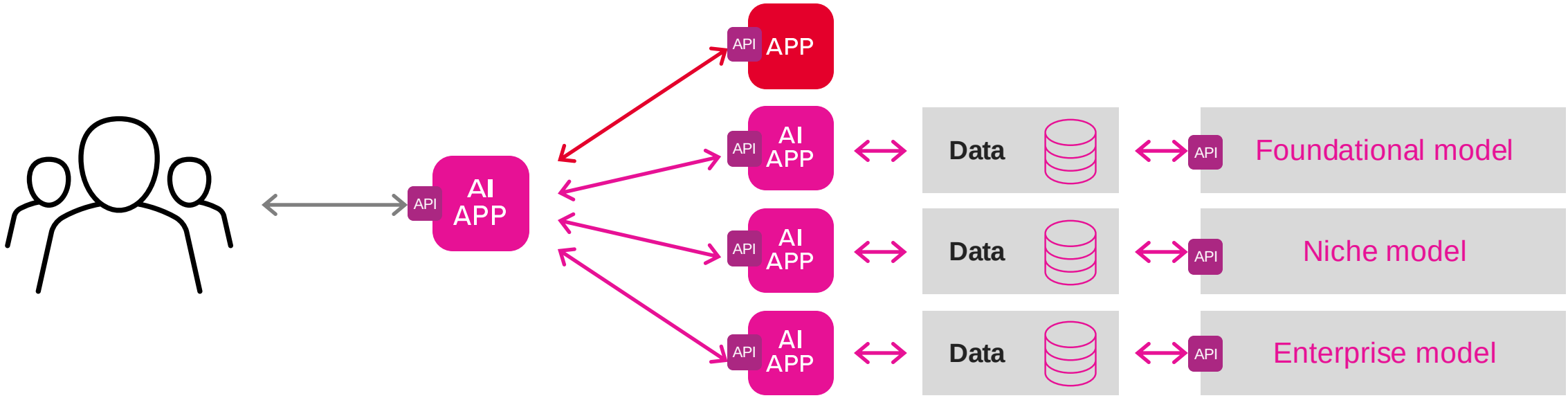
GPU資源，數據分散和新模型
將創造全新雲互連架構需求

3



AI 應用將更加依賴 API

4



API Reliance: GenAI依靠API實現無縫整合，訪問外部數據和服務，增強功能，並有效地促進不同系統和平臺之間的通信。

AI 驅動的應用與眾不同

多模

數據

安全

AI 應用帶來的變化

傳統 App	VS	AI App
人	訪問模式	人與機器
資料庫	數據形態	資料庫、文件
靜態、集中	資源配置	動態、分散式
單向	存取控制	雙向
字元	資料類型	多模態

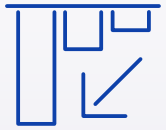
AI 對於傳統架構帶來的挑戰是什麼？

網路架構複雜

多雲自動化部署

部署問題

缺乏可視性



不同的部署工具

運維問題

數據洩漏、SSRF漏洞

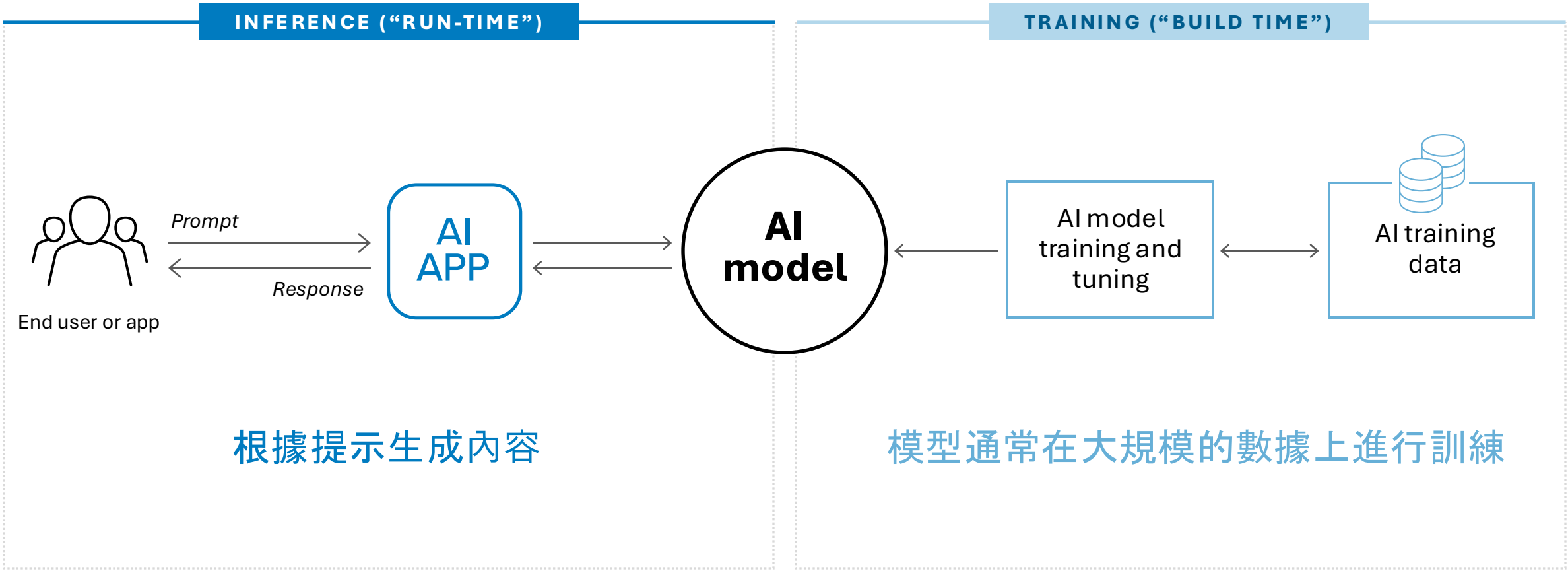
請求資料管理複雜

安全問題

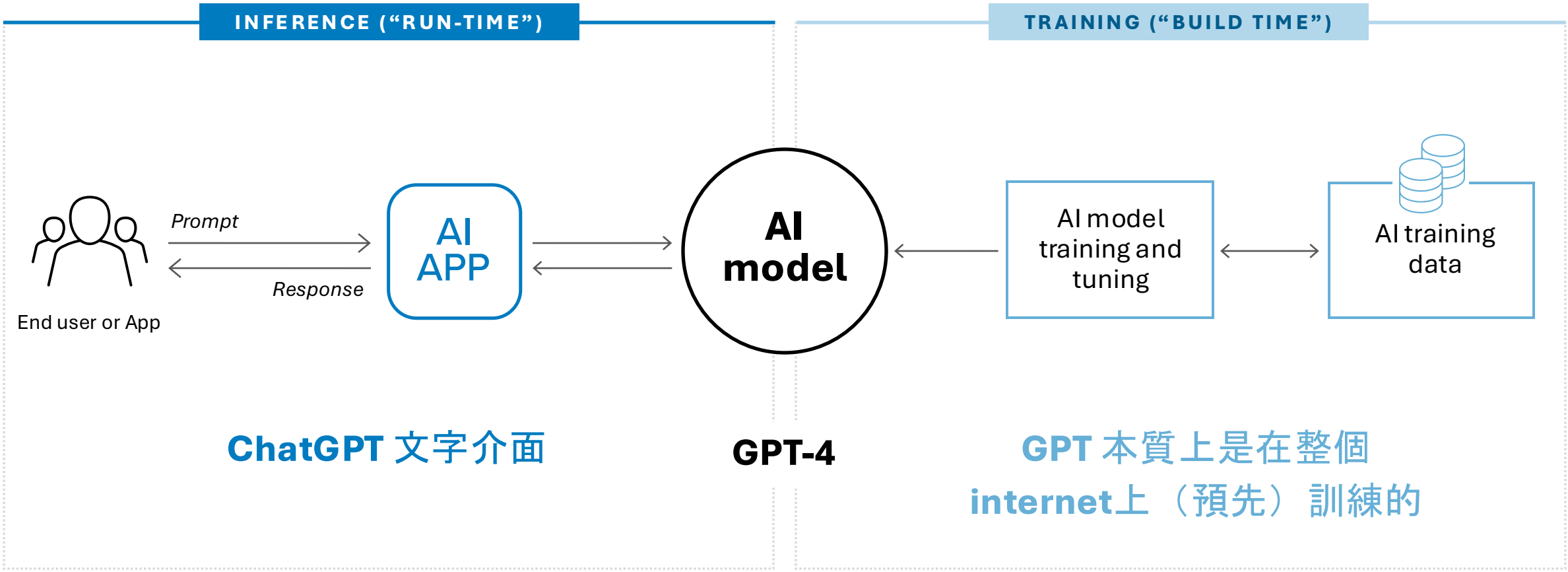
ADC for AI and AI for ADC

大多數應用將成為 AI 應用，需要加強新功能

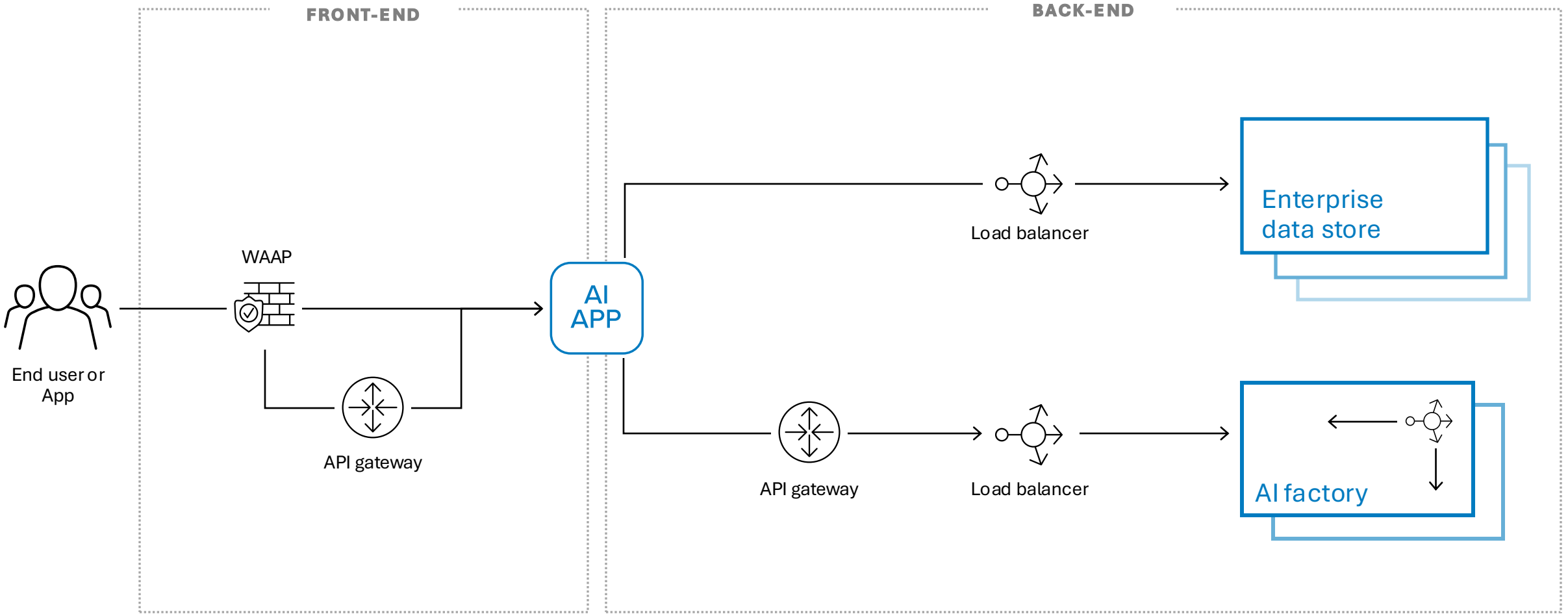
生成式 AI 涉及兩個不同的階段: training & inference



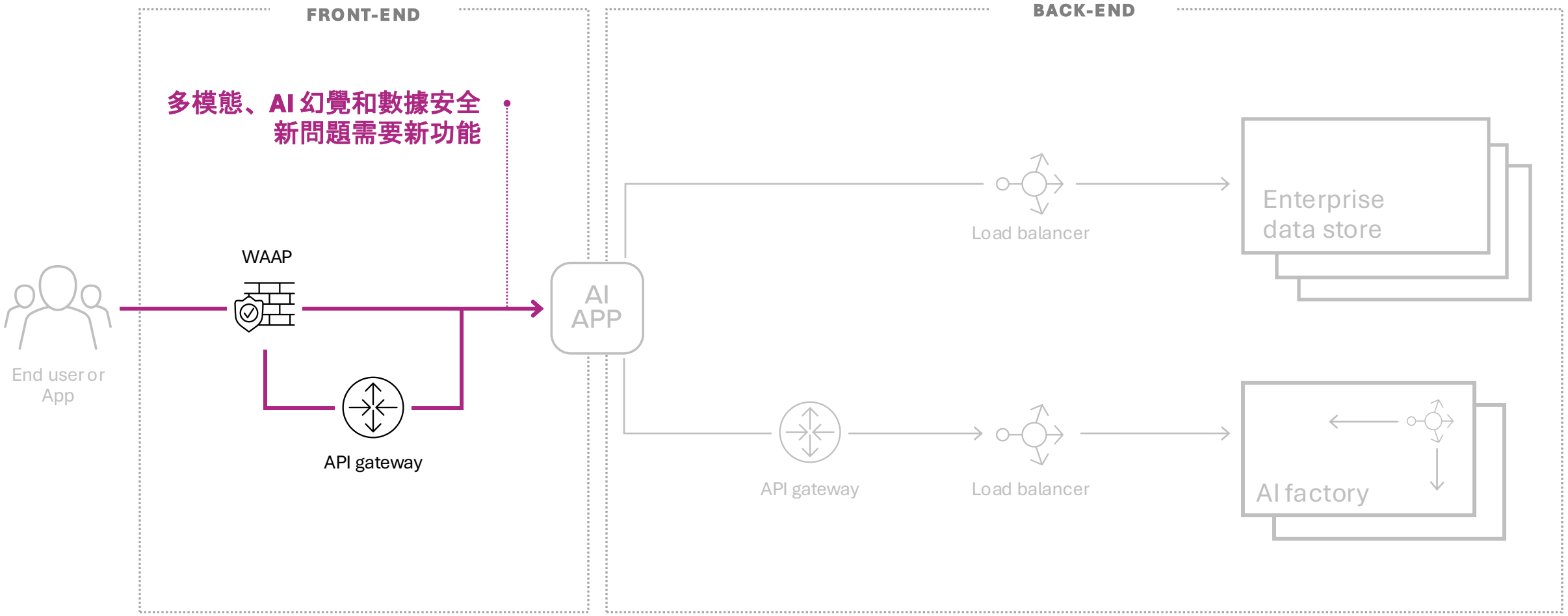
例如，用戶連接到 ChatGPT 以查詢在大型數據集上訓練的 GPT-4 基礎模型



AI 應用在推理期間的通用流量路徑架構



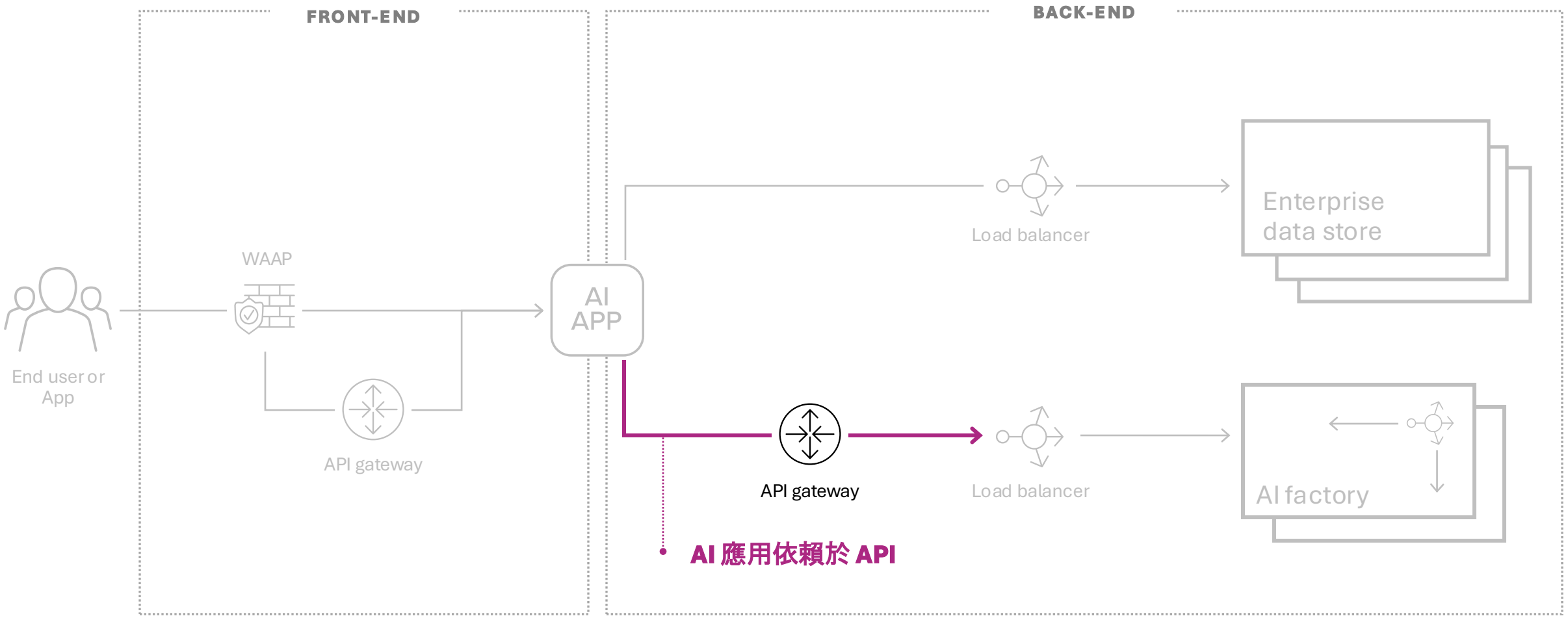
因應AI的入口安全防護



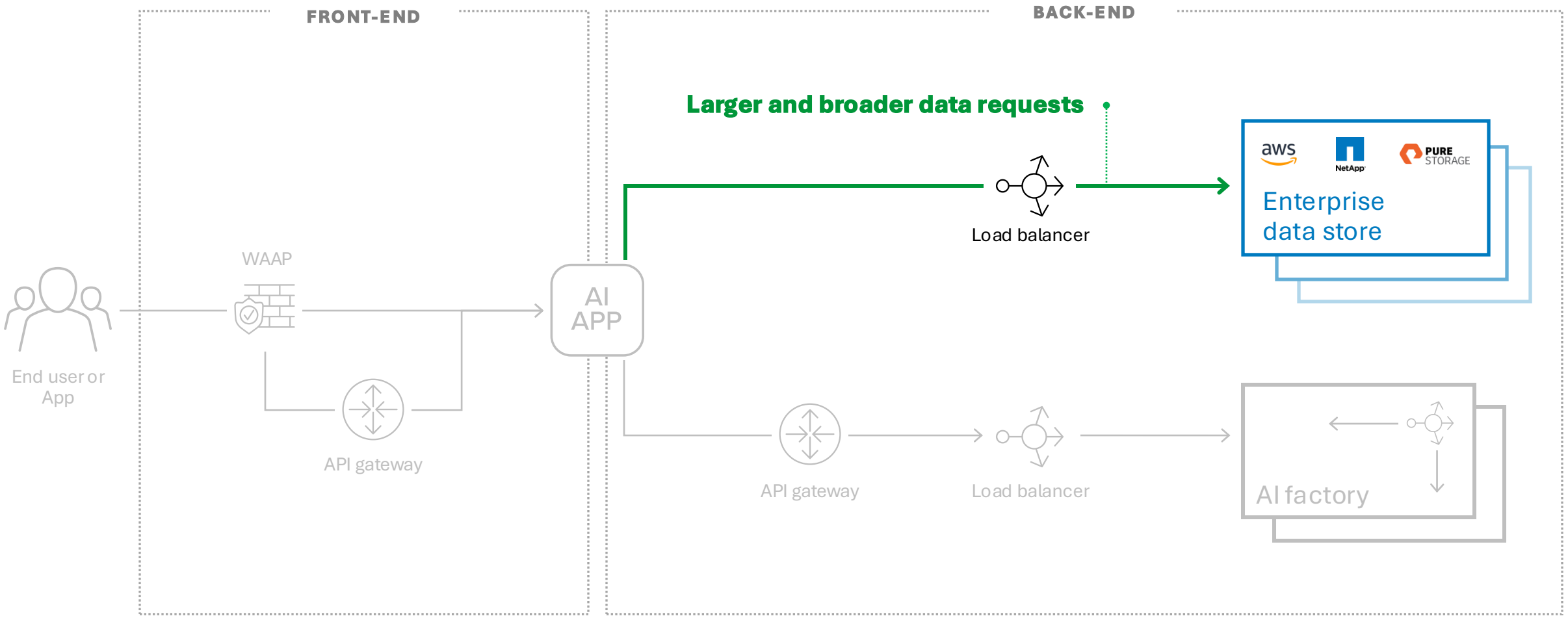
Changes to existing flow

Creation of new flow

使用新功能保護前端至後端的流量



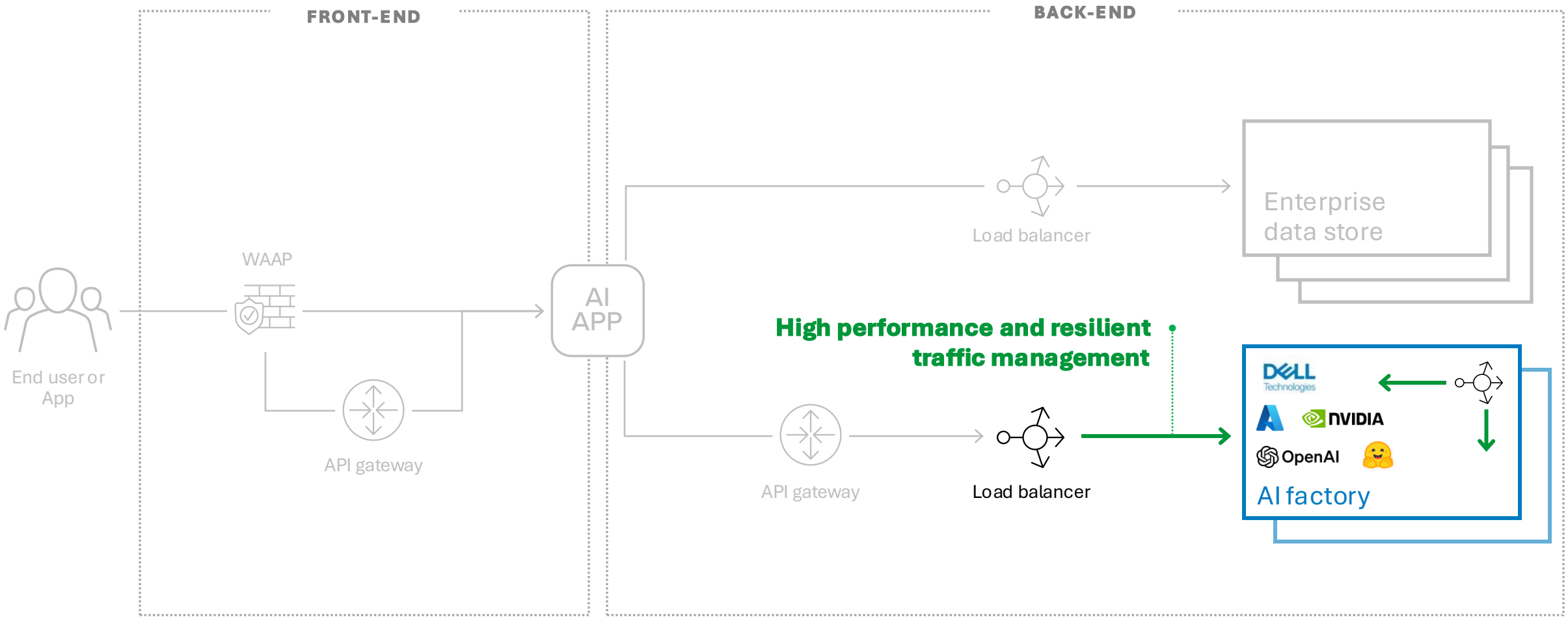
解決 AI 數據流量管理的新挑戰



Changes to existing flow

Creation of new flow

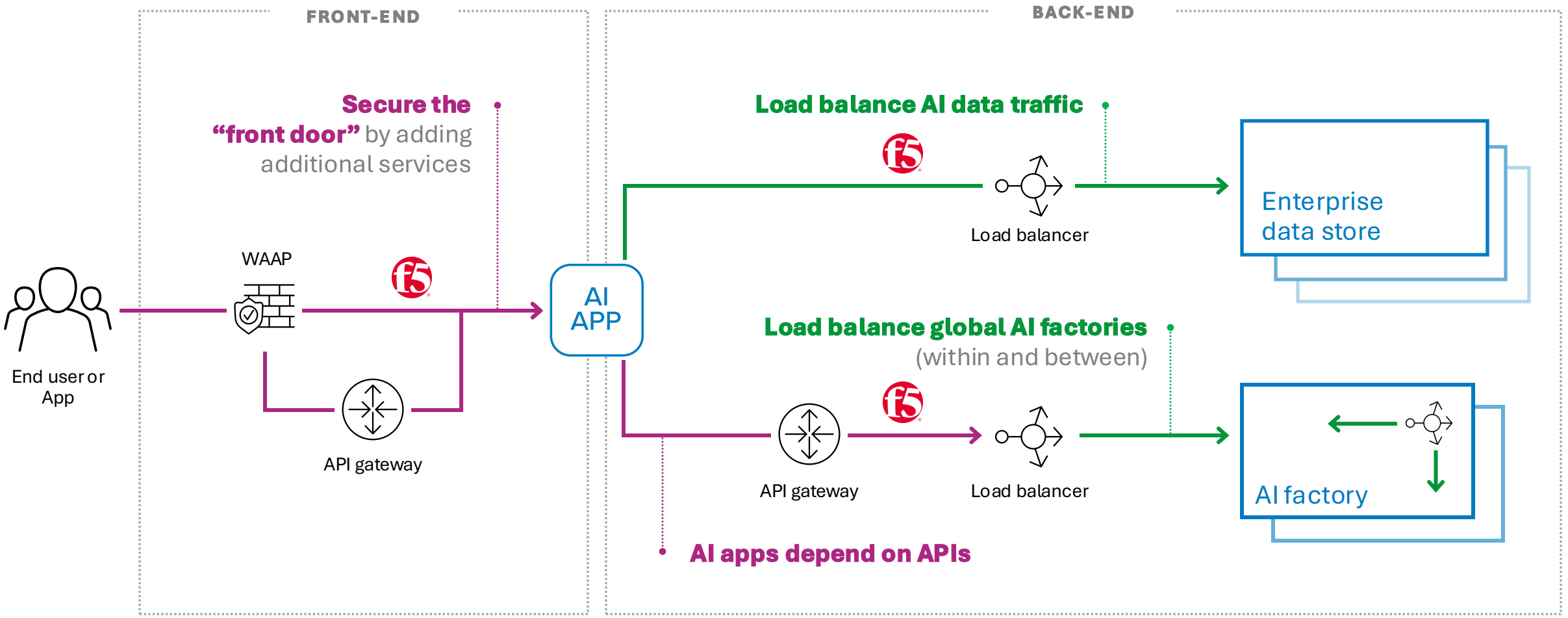
AI 工廠的新興流量



Changes to existing flow

Creation of new flow

ADC 3.0 將保護和交付 AI 應用

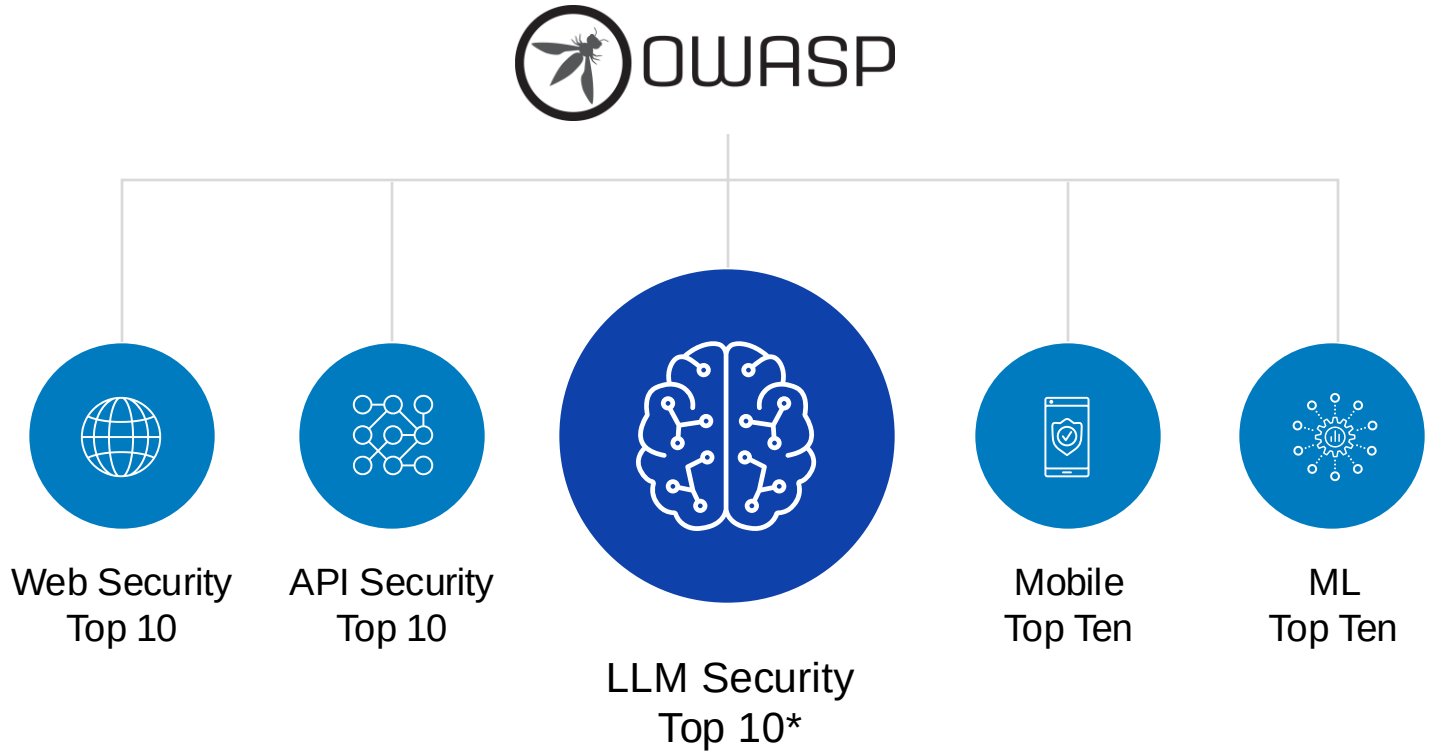


Generative AI LLM 與 API Security

> 無論你如何使用 Gen AI，到頭來，你都是在使用 SDK 透過 REST API 調用端點

Mete Atamel
Developer Advocate, Google Cloud

Google Cloud



面對數據丟失、治理、可觀察性、性能、成本管理和安全威脅等挑戰

F5 AI Gateway是全新 API 安全解決方案，可保護、加速和觀察 AI 驅動的應用。

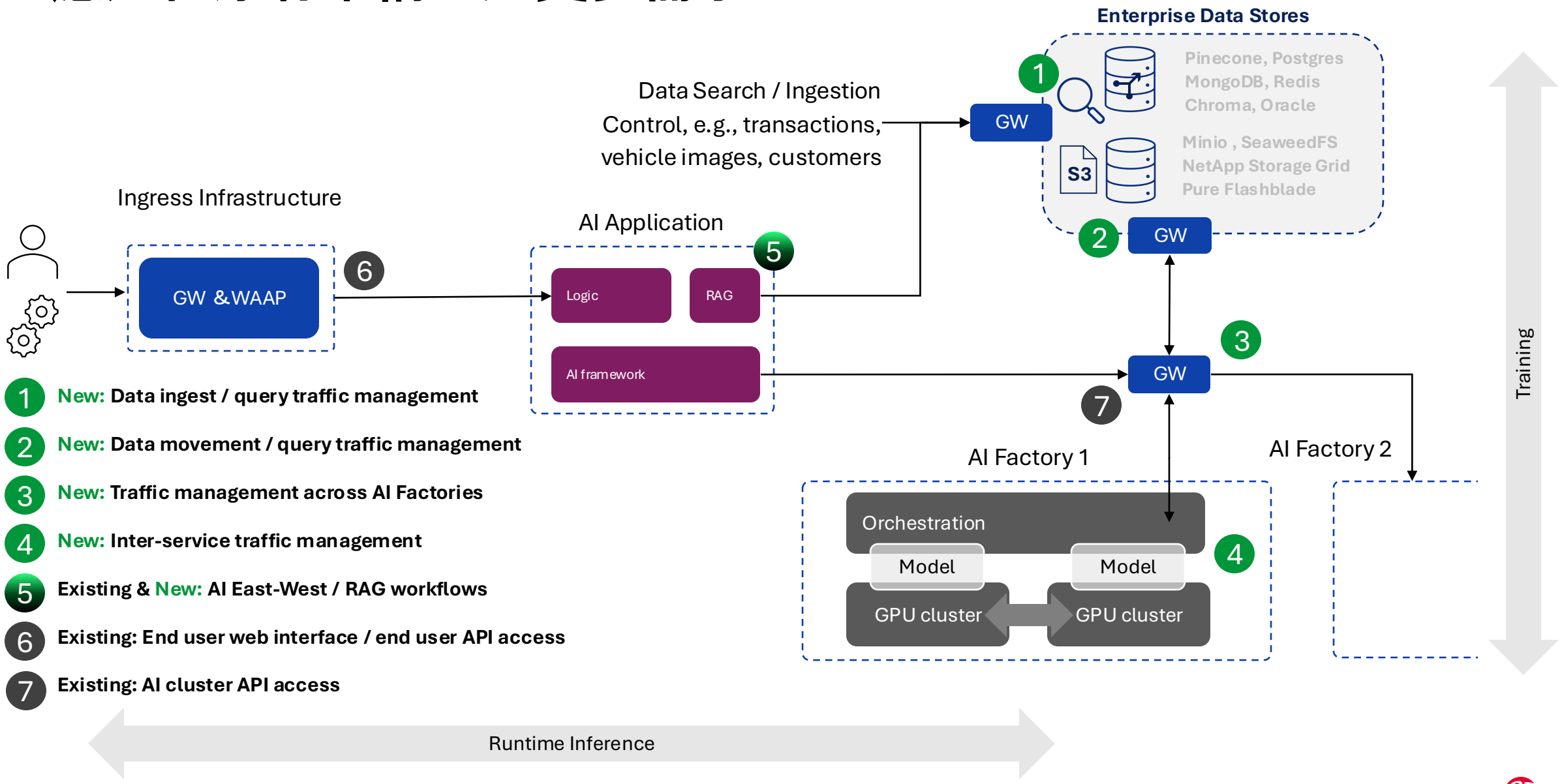
與 AI 初創公司和 API 閘道不同，F5 AI 閘道提供高度可擴展的平臺和開放的保護生態系統，可輕鬆與任何現有的 F5 平臺整合。

AI閘道並非API閘道的簡單變換

- 認證授權
- 限流限速
- L7路由
- OWASP API top 10

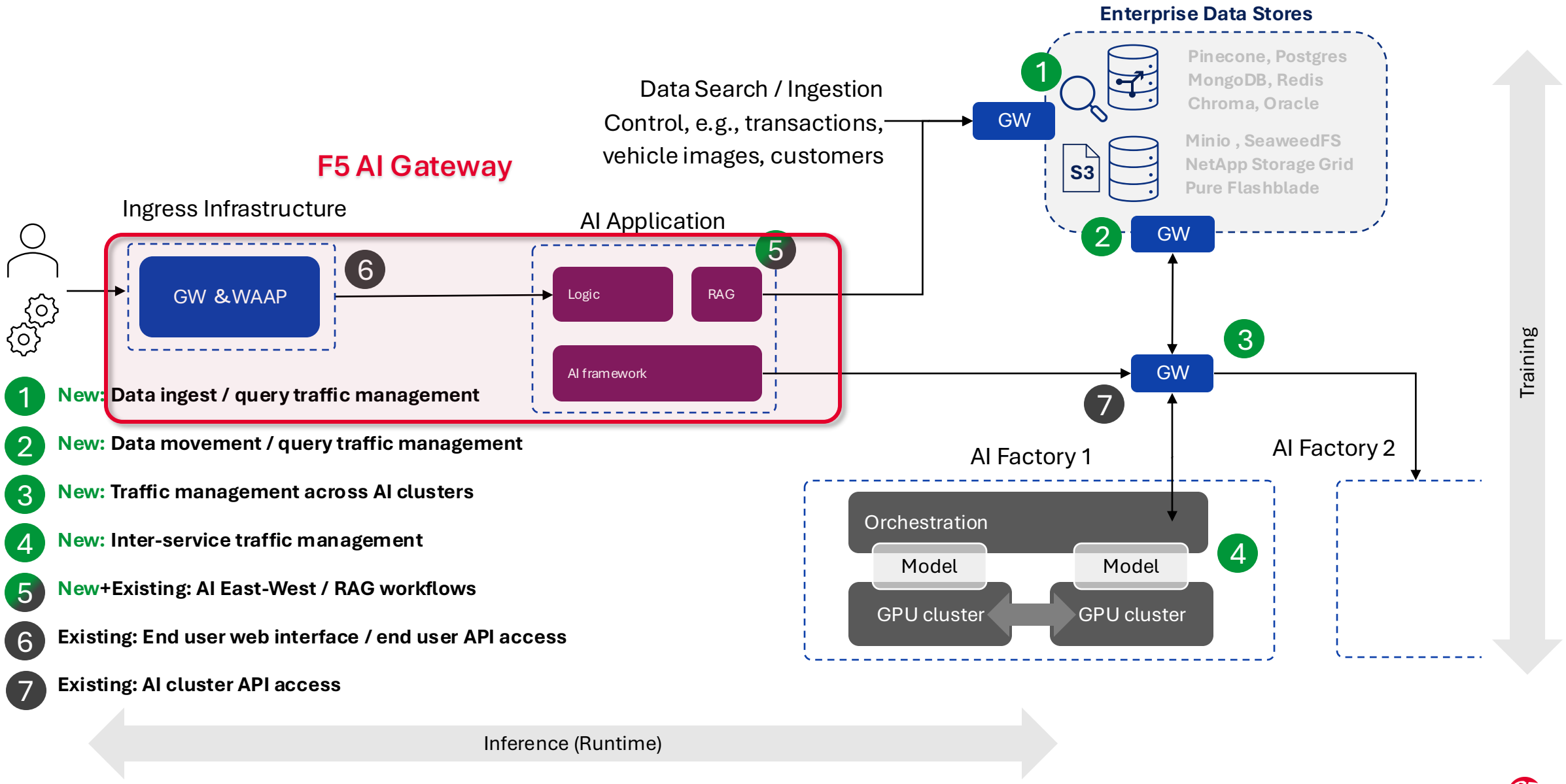
- 提示詞
- Tokens
- 回應控制
- 多模型支援
- OWASP AI top 10

AI 應用在原有架構加入更多需求



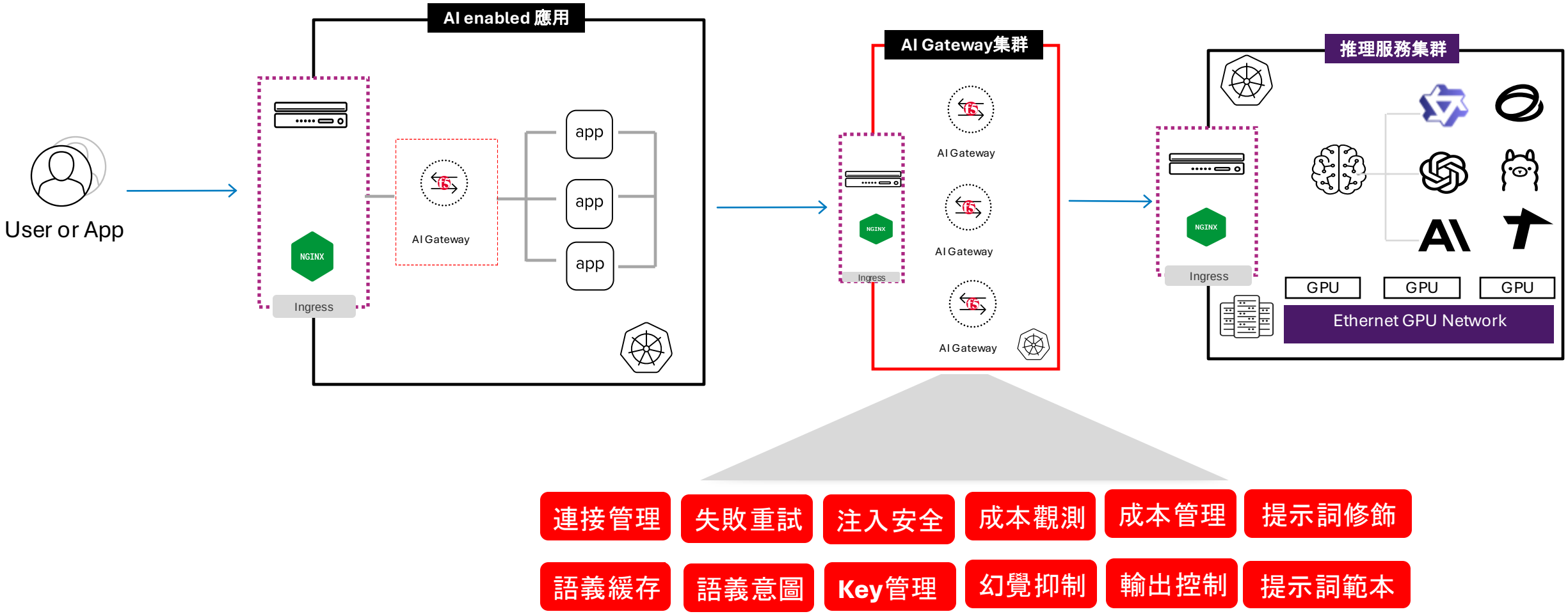
1. **New:** Data ingest / query traffic management
2. **New:** Data movement / query traffic management
3. **New:** Traffic management across AI Factories
4. **New:** Inter-service traffic management
5. **Existing & New:** AI East-West / RAG workflows
6. **Existing:** End user web interface / end user API access
7. **Existing:** AI cluster API access

AI 應用在原有架構加入更多需求



AI Gateway可以有效幫助降低AI應用開發複雜性

AI Gateway幫助解決AI模型的安全、擴展、成本、易用性管理以及語義緩存、幻覺等諸多問題



連接管理 | 失敗重試 | 注入安全 | 成本觀測 | 成本管理 | 提示詞修飾

語義緩存 | 語義意圖 | Key管理 | 幻覺抑制 | 輸出控制 | 提示詞範本

AI閘道 幫助解決AI應用四大挑戰

安全防護

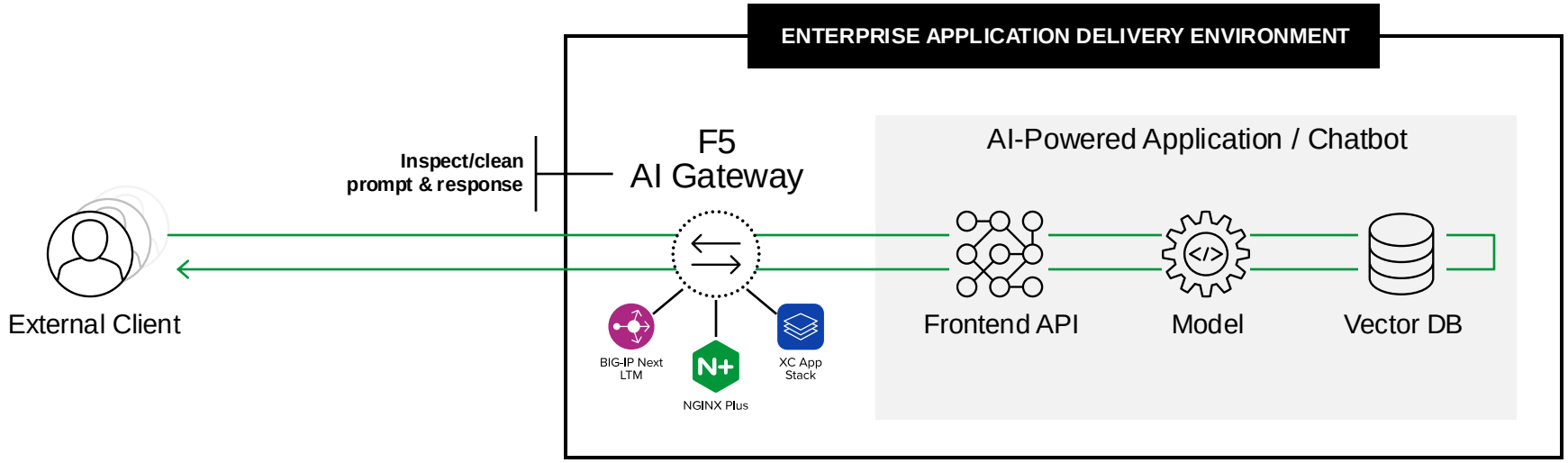
模型擴展

易用管理

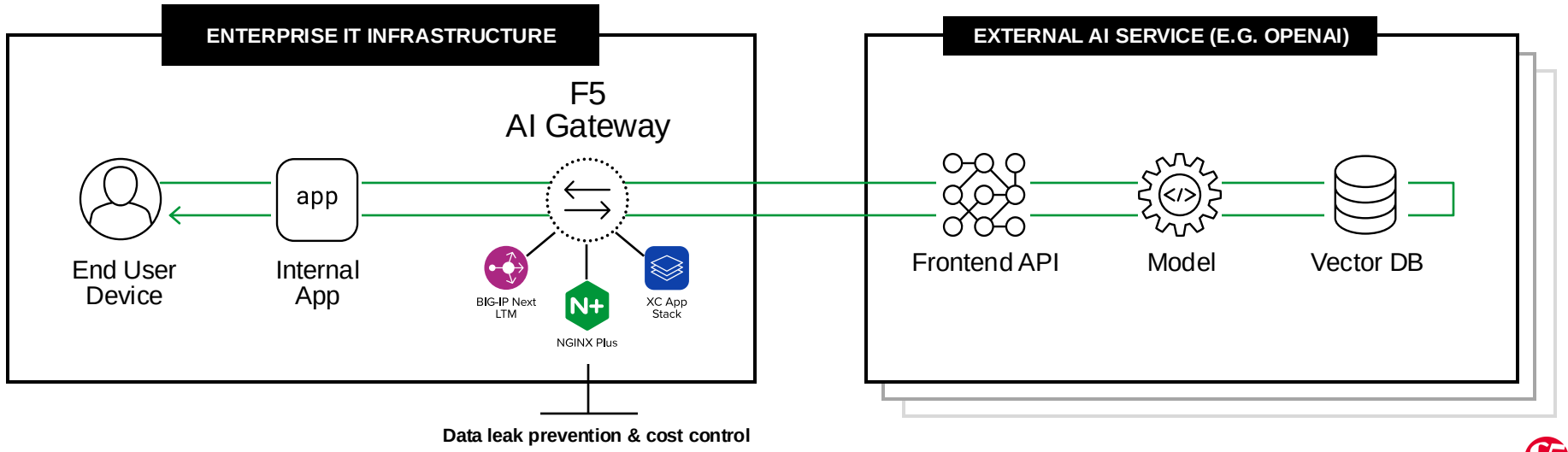
成本控制

構建 AI 驅動的應用的企業

路由、緩存、可觀察性和威脅防護適用於託管和發佈 AI 驅動的應用的企業



安全高效的使用外部 AI 服務



總結

#1

每個應用都將被AI-powered

#2

AI將使用API 構建新型應用

#3

AI 將讓模型和數據與運算無處不在

API安全與多雲網路防護

高效能Kubernetes 網路