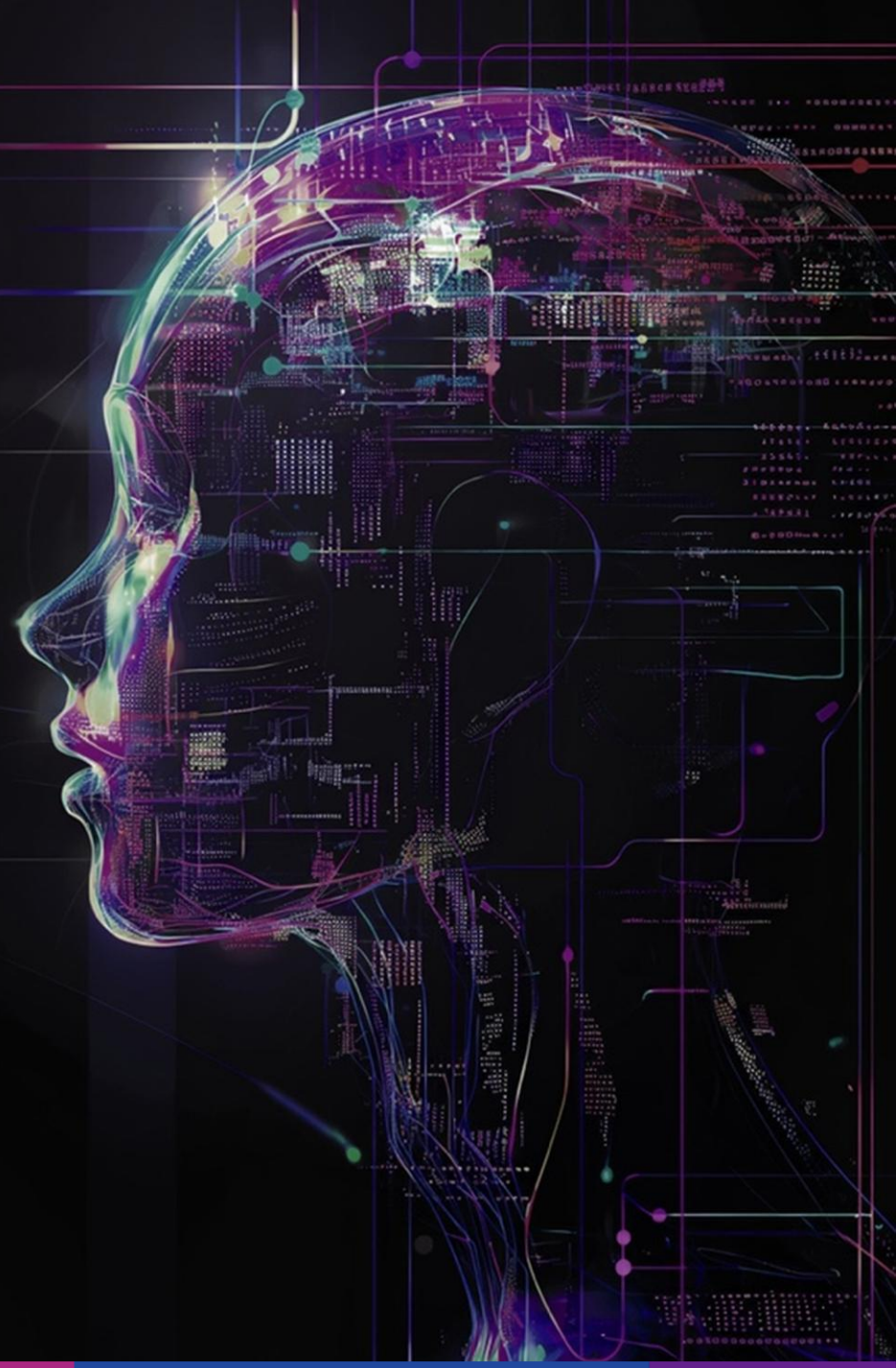
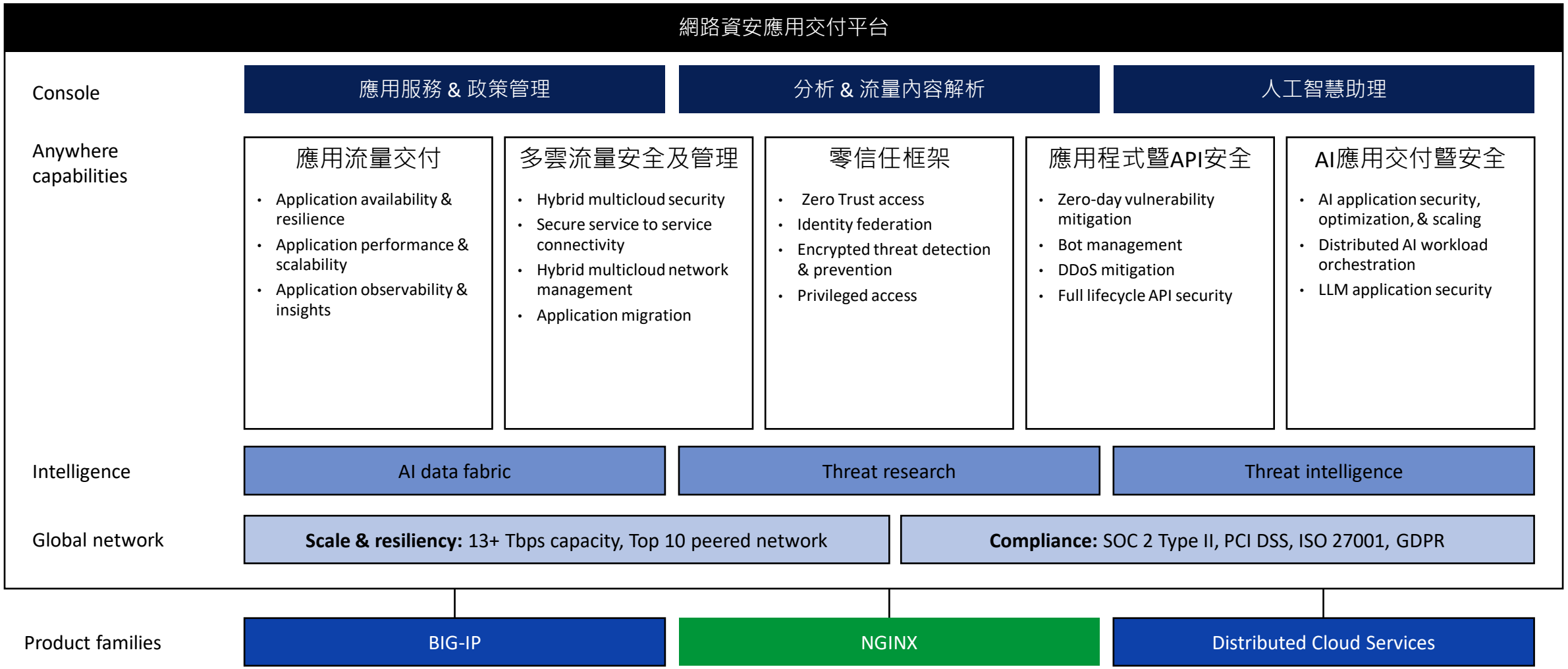
應對 AI 安全與效能挑戰

F5 AI Gateway 解決方案

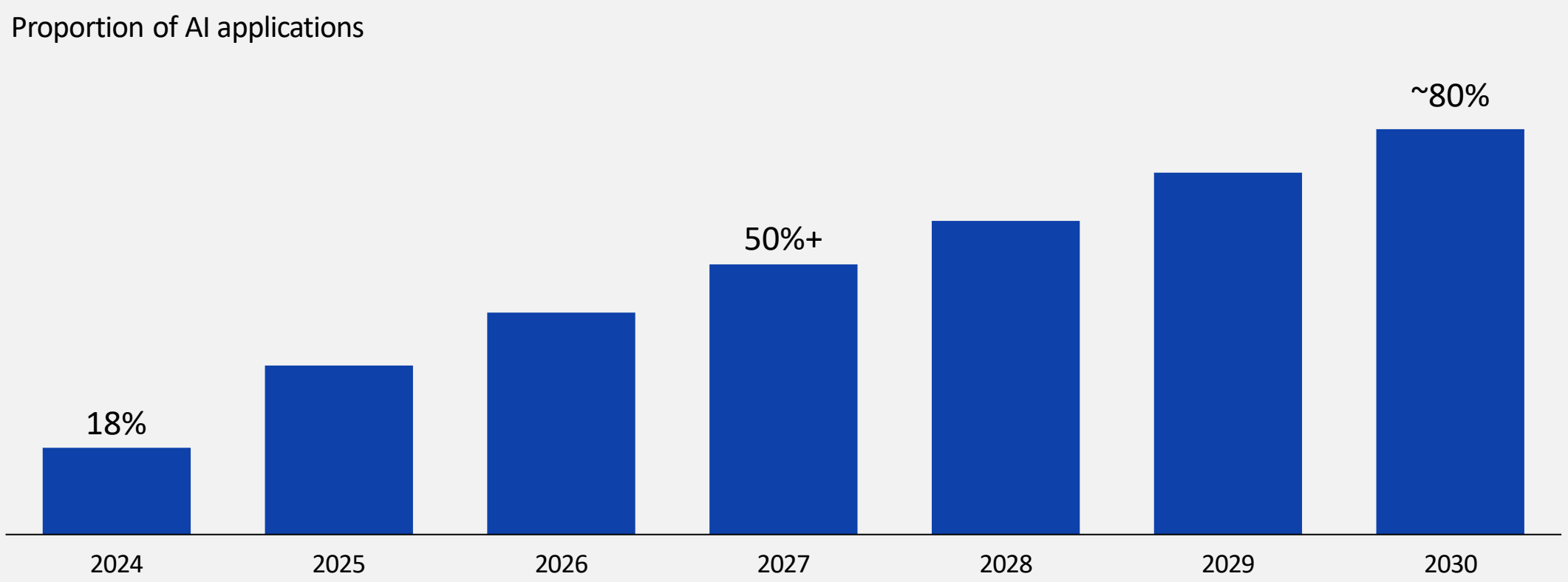


F5 ADC 3.0 網路資安應用交付平台

ADC 3.0



未來幾乎每個應用都會成為 AI 應用

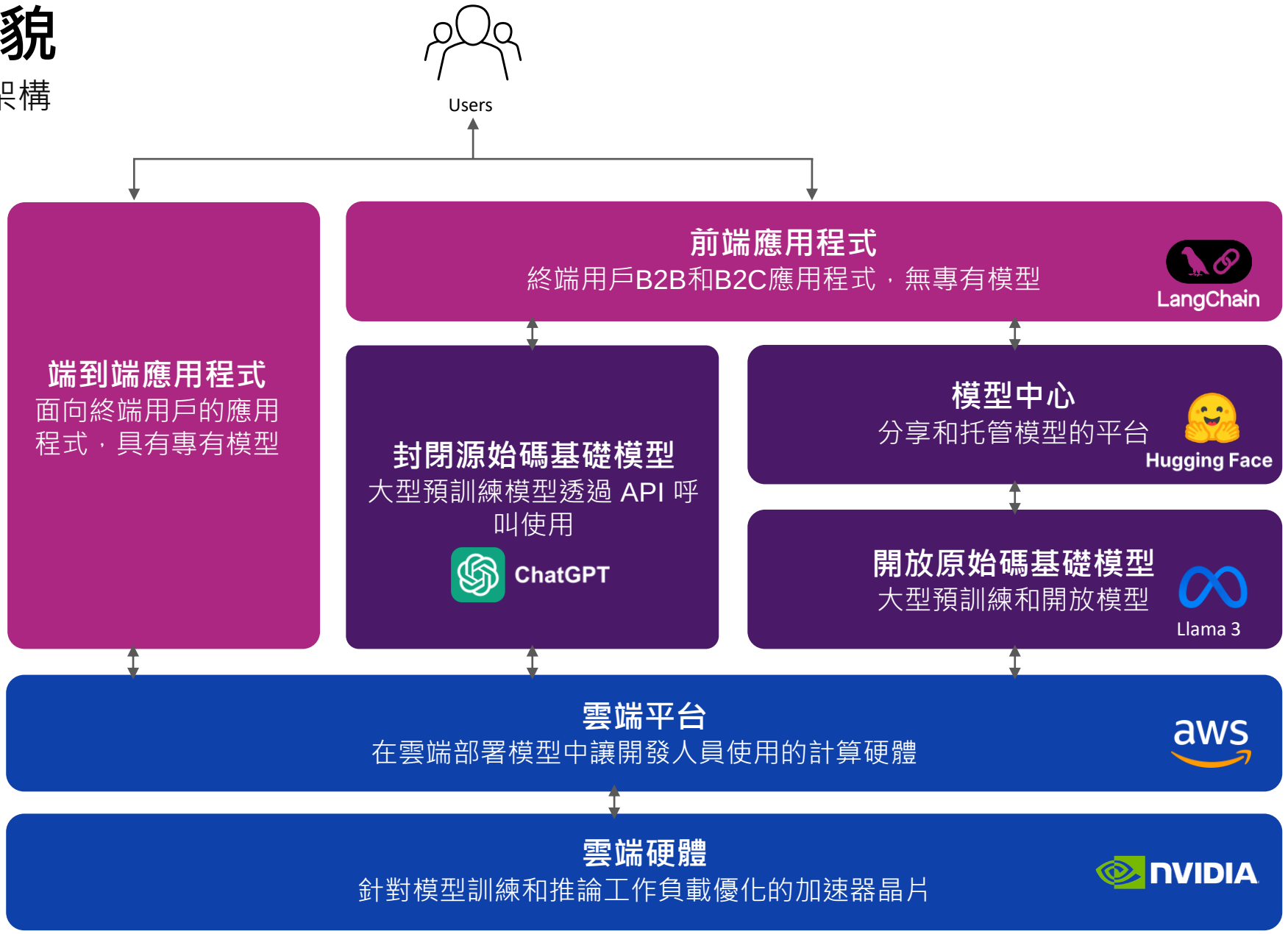


Source: IDC, Gartner, and F5 Corporate Strategy

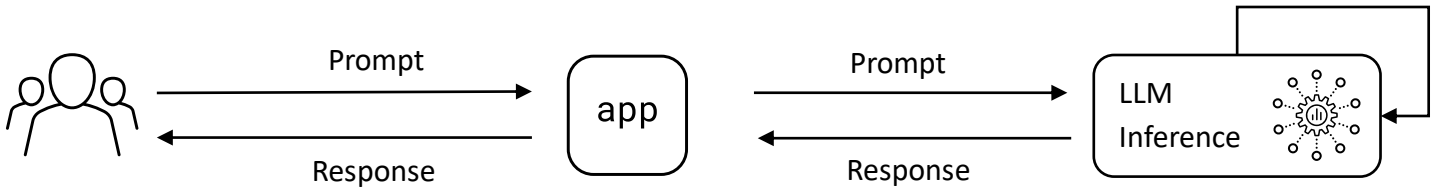
AI 應用架構全貌

應用、模型與雲端基礎架構

- 應用程式
- 模型
- 基礎設施



AI 應用在企業落地的基本架構



app

前端應用
(Application Front-end)

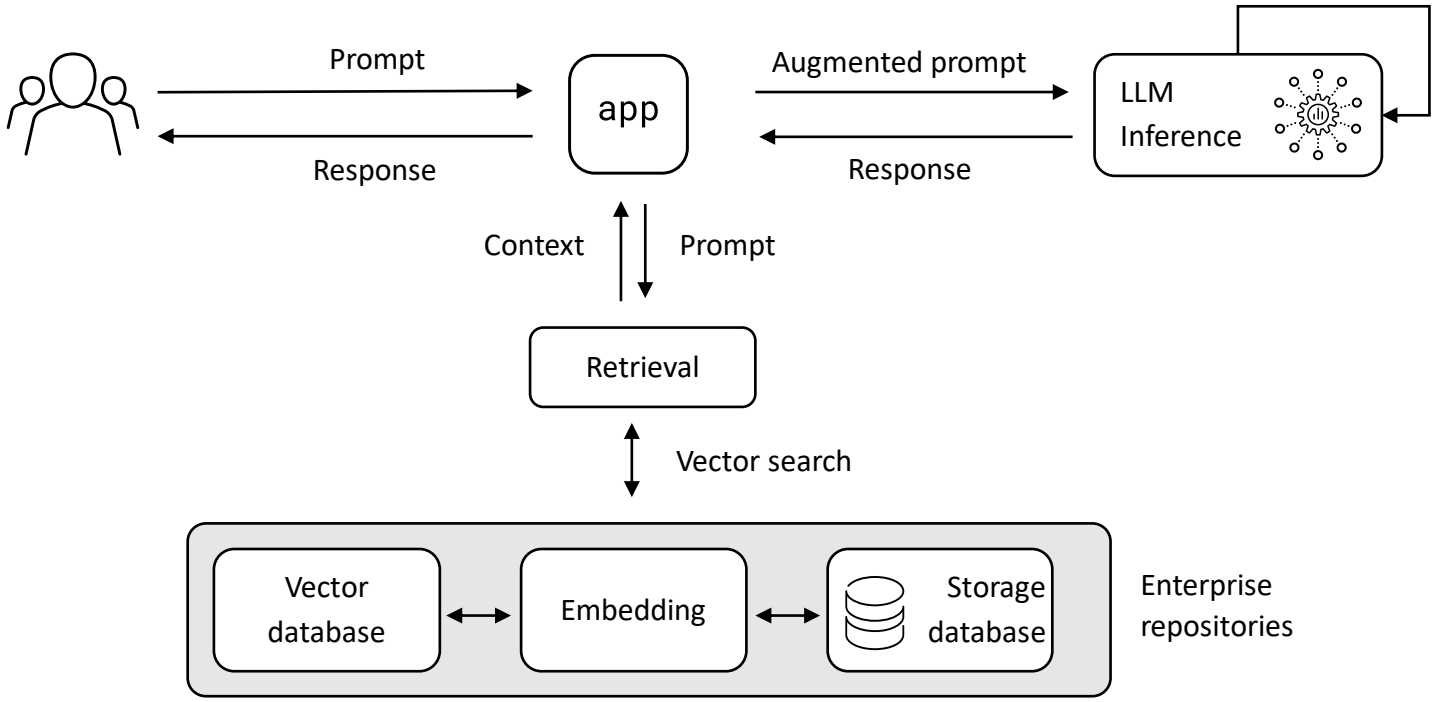
- 可部署於私有資料中心或雲端環境
- 可搭配既有工具與開發框架靈活整合

大型語言模型
(Large Language Model)

- 可透過 AI 工廠、自建私有資料中心、雲端平台或即服務 (as-a-service) 方式提供
- 可選擇使用私有 LLM 或公開模型 (如 OpenAI、Claude 等)
- 商用版本或開源模型皆可彈性應用

ChatGPT Gemini Llama 3 MISTRAL AI_

RAG：整合企業知識庫的 AI 應用架構



Retrieval-Augmented Generation (RAG)

LLM 客製化
(LLM Customisation)

- 可預訓練 (Pre-training)
- 中心微調 (Fine-tuning)
- 檢索增強生成 (Retrieval-Augmented Generation, RAG)

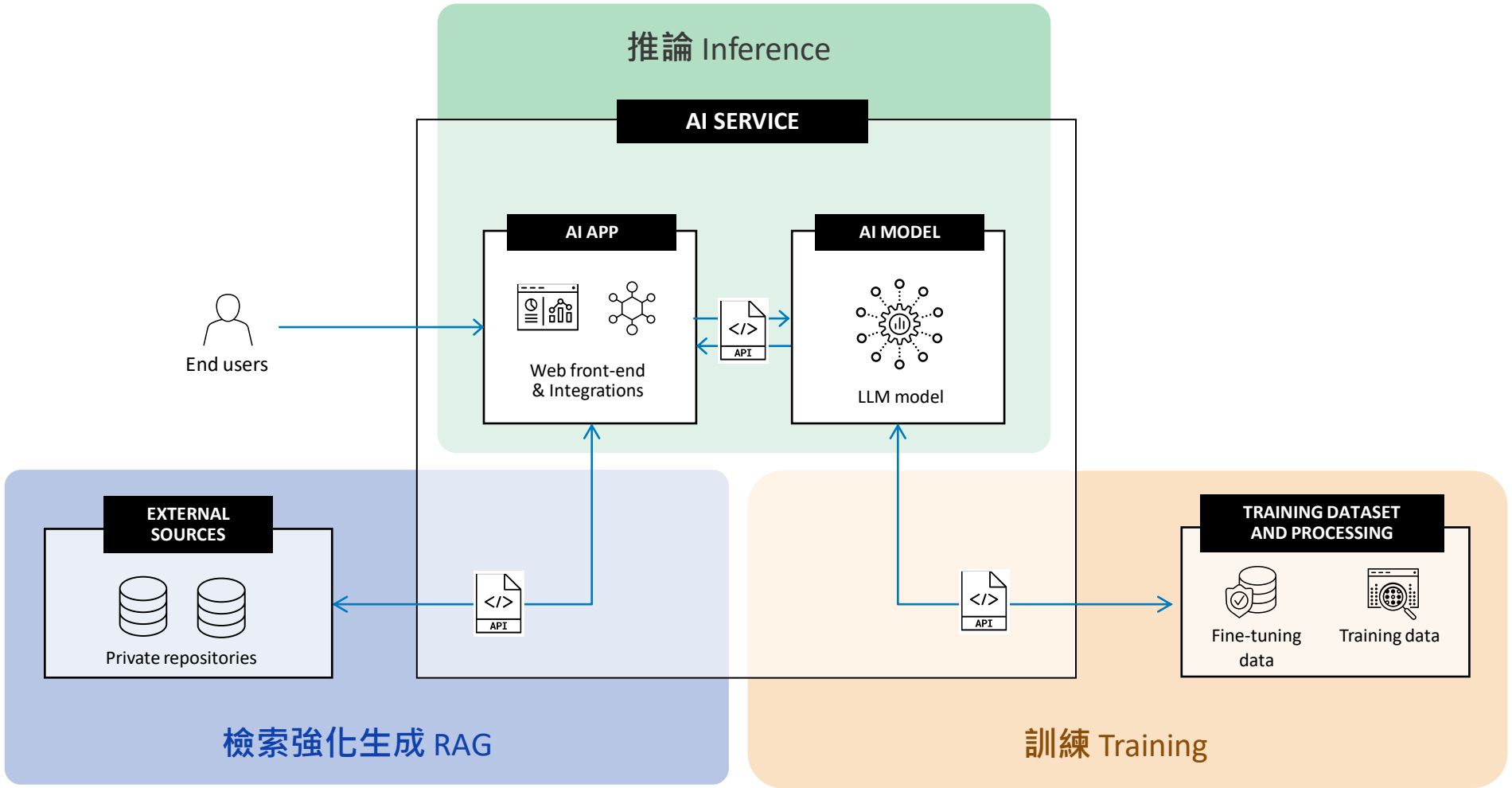
Retrieval

檢索與嵌入
(Retrieval and Embedding)

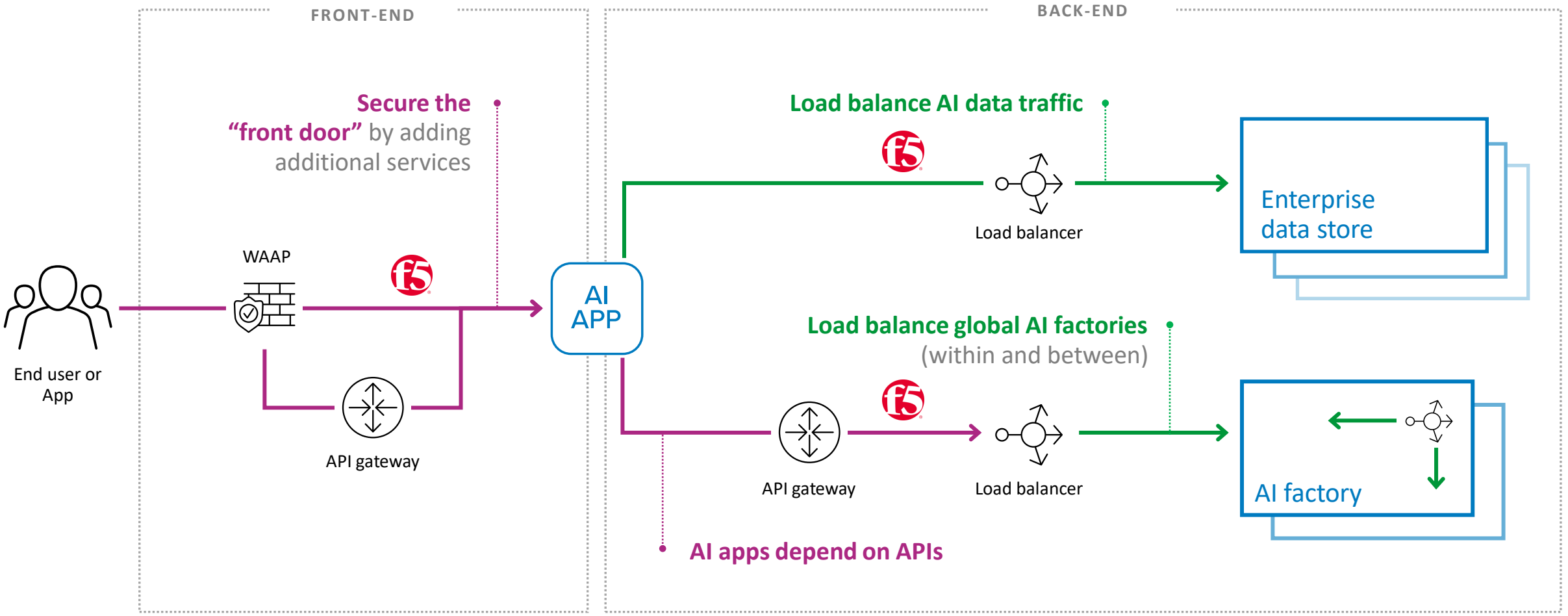
- 可部署於私有資料中心或雲端平台，依資料敏感度彈性規劃
- 可使用主流開發框架，如 LangChain、Haystack、LlamaIndex 等進行檢索流程建構與語意嵌入處理

LangChain Haystack by deepset LlamaIndex

AI 應用架構三核心：訓練、推論與檢索強化生成 (RAG)



ADC 3.0 將保護和交付 AI 應用



防護和運維 AI 應用的挑戰

安全Security

- 來自使用者輸入的敏感數據
- 來自使用者的惡意輸入（代碼）
- 敏感數據輸出 – PII
- 數據竊盜
- AI 應用和 AI /API 蔓延擴大了攻擊面

運維Operations

- 影子 AI 擴散
- 缺乏成本控制 + 昂貴的基礎設施
- 多模型配置和複雜性（路由、緩存、轉換）
- LLM 生命週期和治理

可觀察性Observability

- 合規性需要端到端的可觀察性、API 和有效負載
- API 發現 + 可觀測性
- 合規性需要身份感知的 LLM 可觀察性
- 與 SIEM 工具整合

AI Gateway：
保護 AI 的第一道防線

F5 AI Gateway專注於 AI 應用的 3 個重點

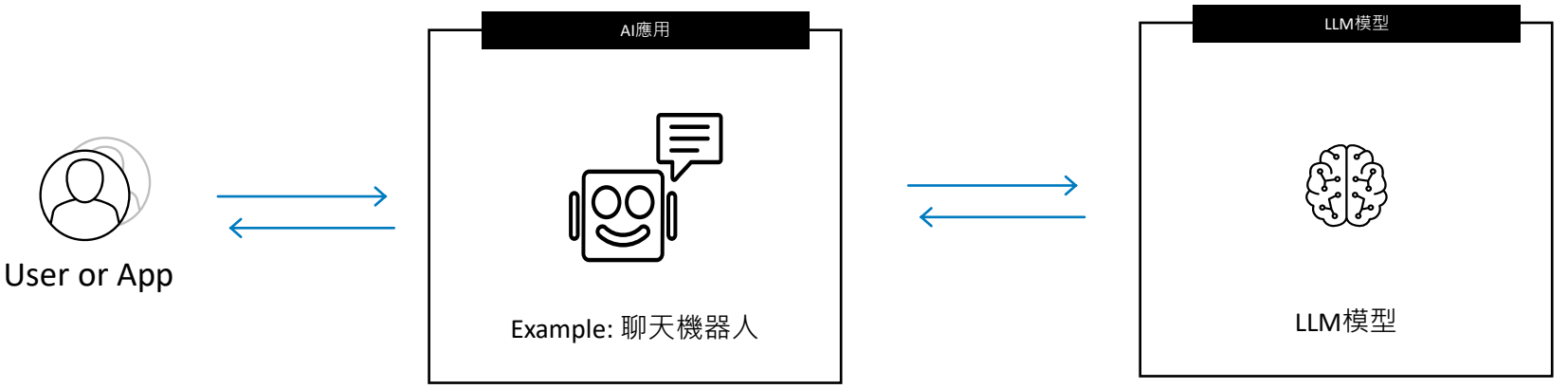
AI模型是必須保護的攻擊面

必須監控和控制生成式 AI 輸出

Layer 8(AI)流量管理對於高效的 AI 應用至關重要

一個基本的生成式AI應用

單一LLM直接應用



- 大模型推理能力的直接輸出
- 應用負責代理使用者輸入與模型輸出

生成式AI應用需考慮的四個重點

安全防護

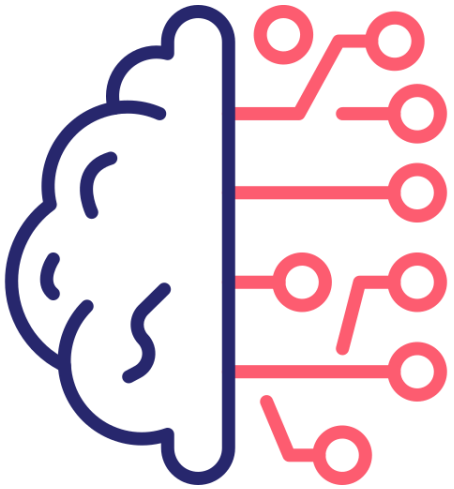
一個基本的生成式AI應用

需在應用中增加提示詞注入的防護模組，需識別提示詞語義，以及需要額外的特定專有模型支援

OWASP Top 10 LLM 安全威脅

INPUT

01 提示詞注入 Prompt Injection
03 供應鏈漏洞 Supply Chain Vulnerabilities
04 資料和模型中毒 Data and Model Poisoning
07 系統提示洩露 System Prompt Leakage
08 向量和嵌入漏洞 Vector & Embedding Weaknesses



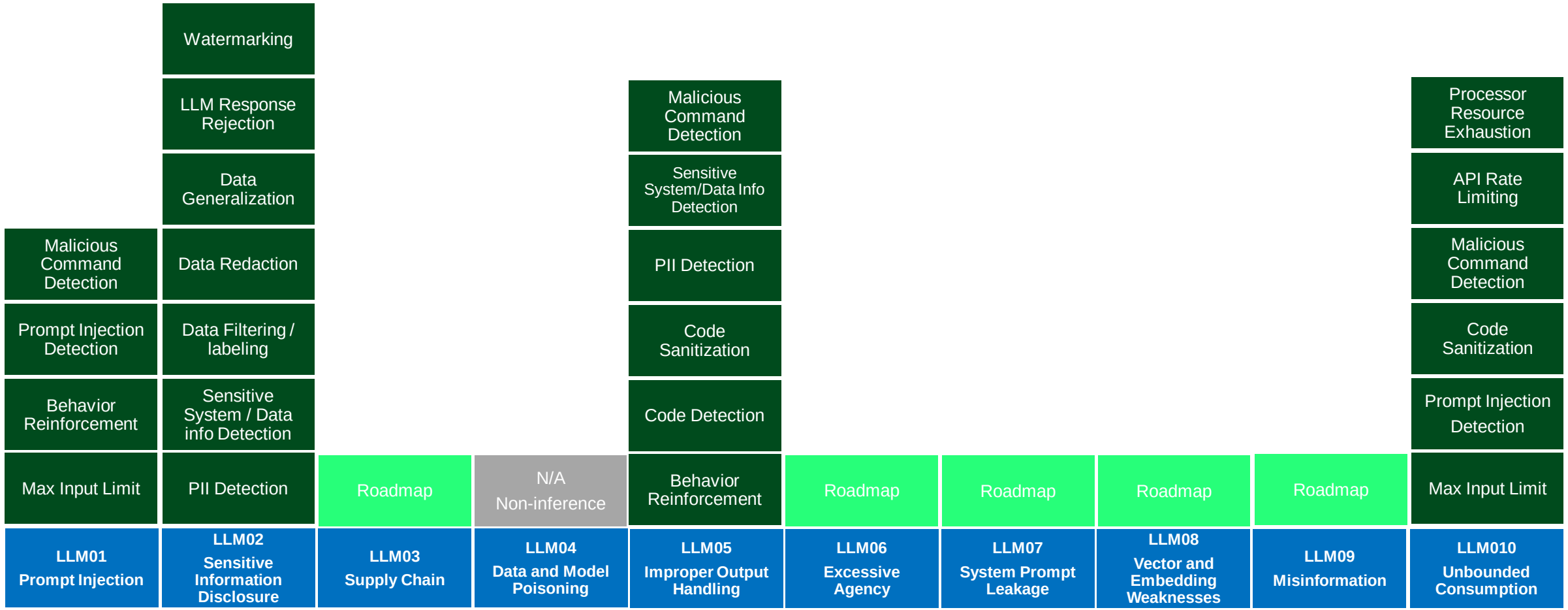
OUTPUT

02 敏感資訊洩露 Sensitive Information Disclosure
03 不當的輸出處理 Improper Output Handling
06 過度代理 Excessive Agency
09 錯誤資訊 Misinformation
10 無限制的資源消耗 Unbounded Consumption

- **輸入模型的威脅**：針對模型本身的攻擊。
- **模型輸出的威脅**：模型生成的內容對外部使用者或系統造成的潛在風險，包括錯誤、洩露和濫用。

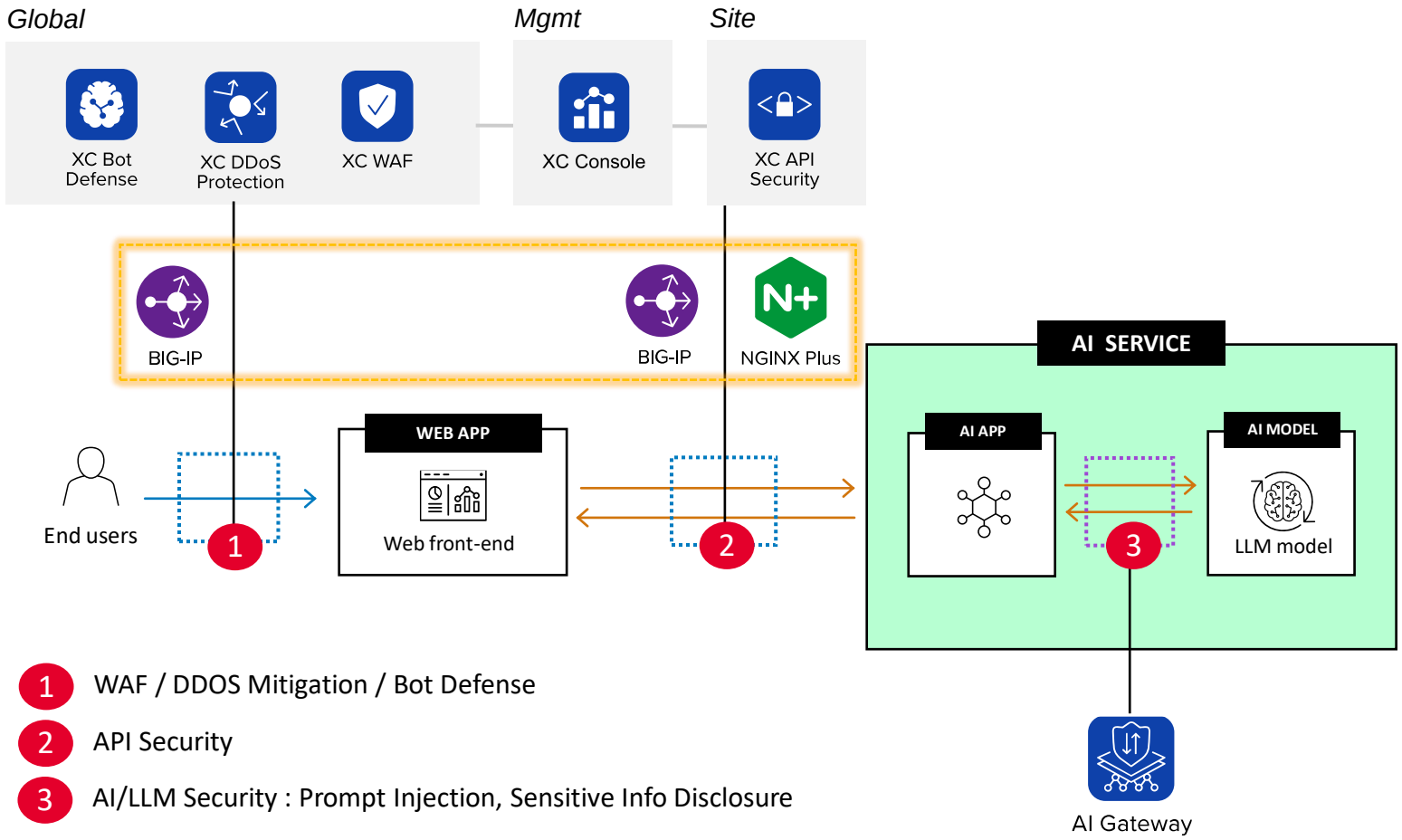
F5 AI Gateway Processors & 2025 OWASP Top 10 LLM

Base package from F5



F5 AI 安全架構：全面守護 API 與 AI 應用

從流量防護、API 閘道，到語意層的 AI 防線，F5 建構橫跨應用與模型的整合式安全架構



1. WAF / DDOS Mitigation / Bot Defense
2. API Security
3. AI/LLM Security : Prompt Injection, Sensitive Info Disclosure

要保護 AI 應用，需要涵蓋以下防護要素

- Web Application Firewall (WAF)
- DDoS Mitigation
- Bot Defense
- API Security
- AI/LLM Security

F5 提供的安全解決方案

- WAAP Solutions
- API Security
- New : F5 AI Gateway (Security)

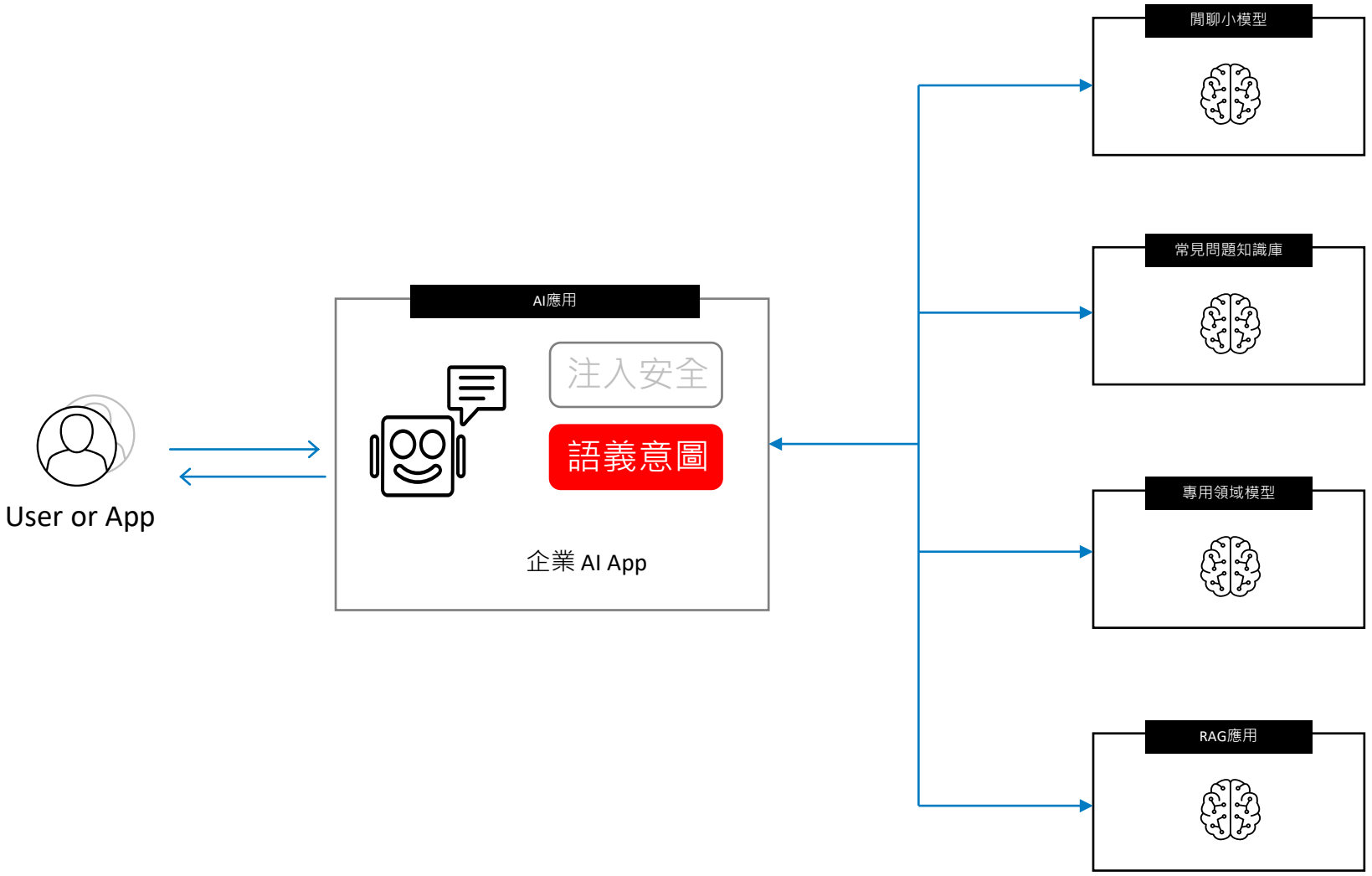
生成式AI應用需考慮的四個重點

安全防護

模型擴展

一個實際企業級AI應用不會只用一個模型

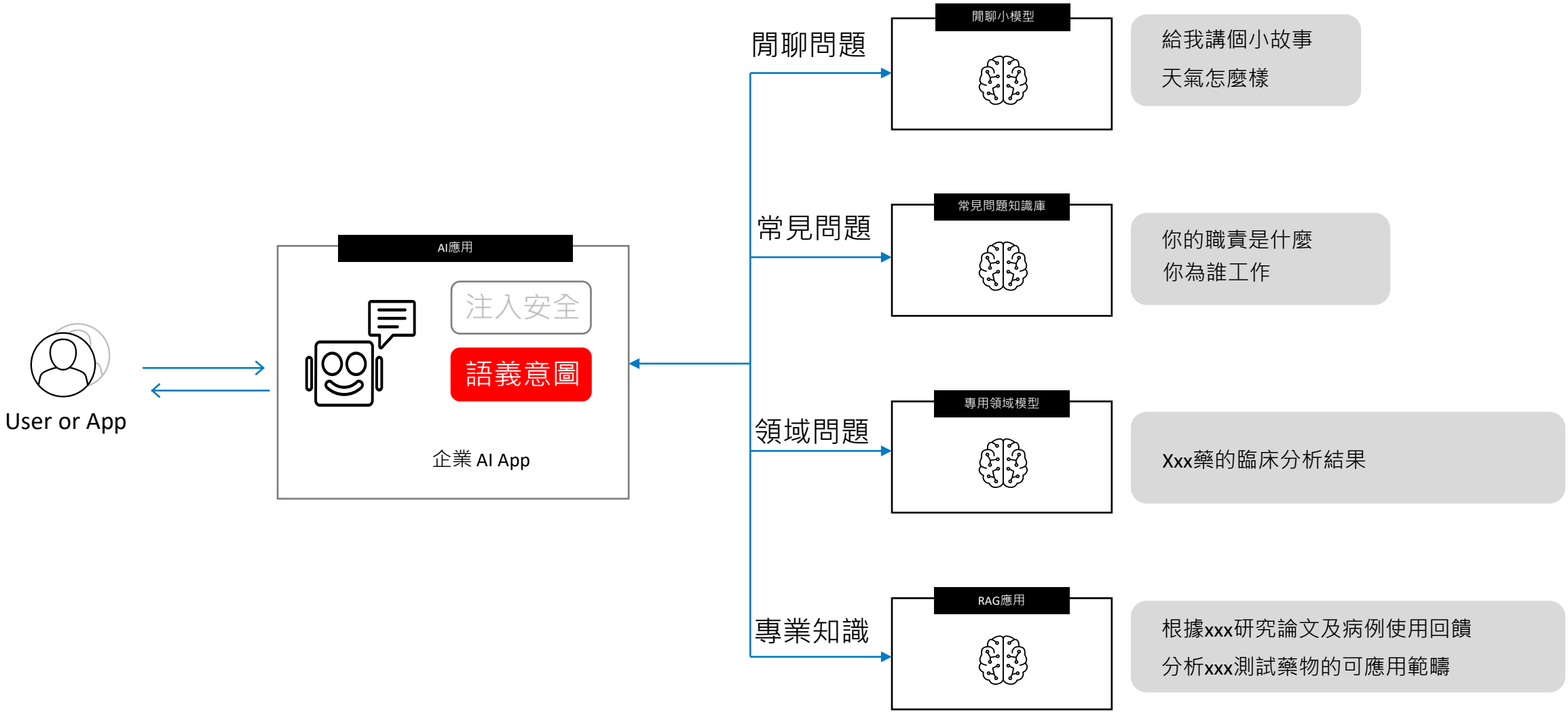
多LLM應用



- 不同的輸入由不同的模型或應用來提供服務
- 基於意圖來判斷用戶輸入
- 基於模型服務性能路由
- 基於模型服務價格來路由
- 解決使用者體驗、性能、可靠性、成本等問題

一個實際企業級AI應用不會只用一個模型

需基於意圖或語義將不同請求路由到不同模型中，並可以隨時擴展更多模型



生成式AI應用需考慮的四個重點

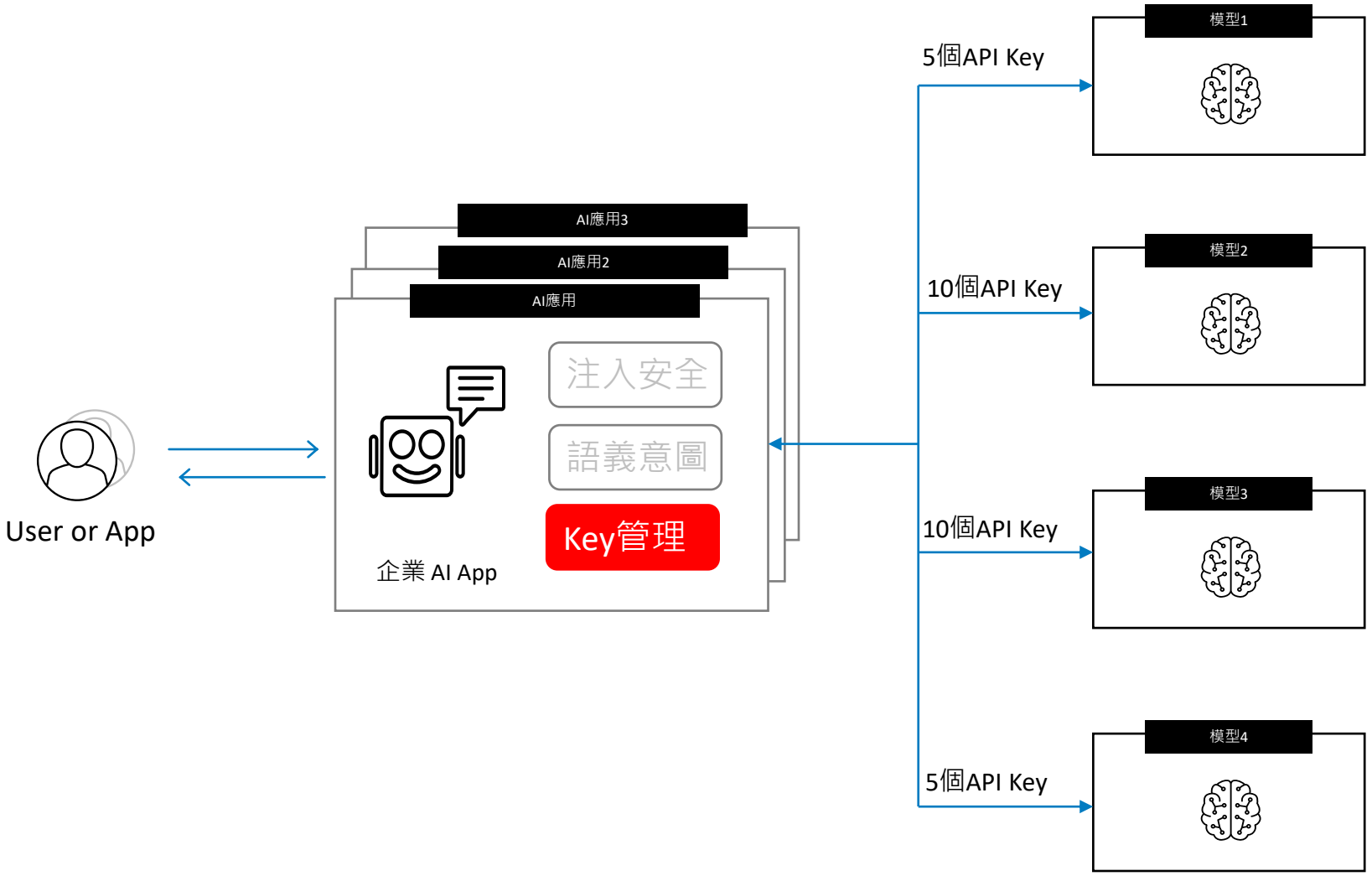
安全防護

模型擴展

易用管理

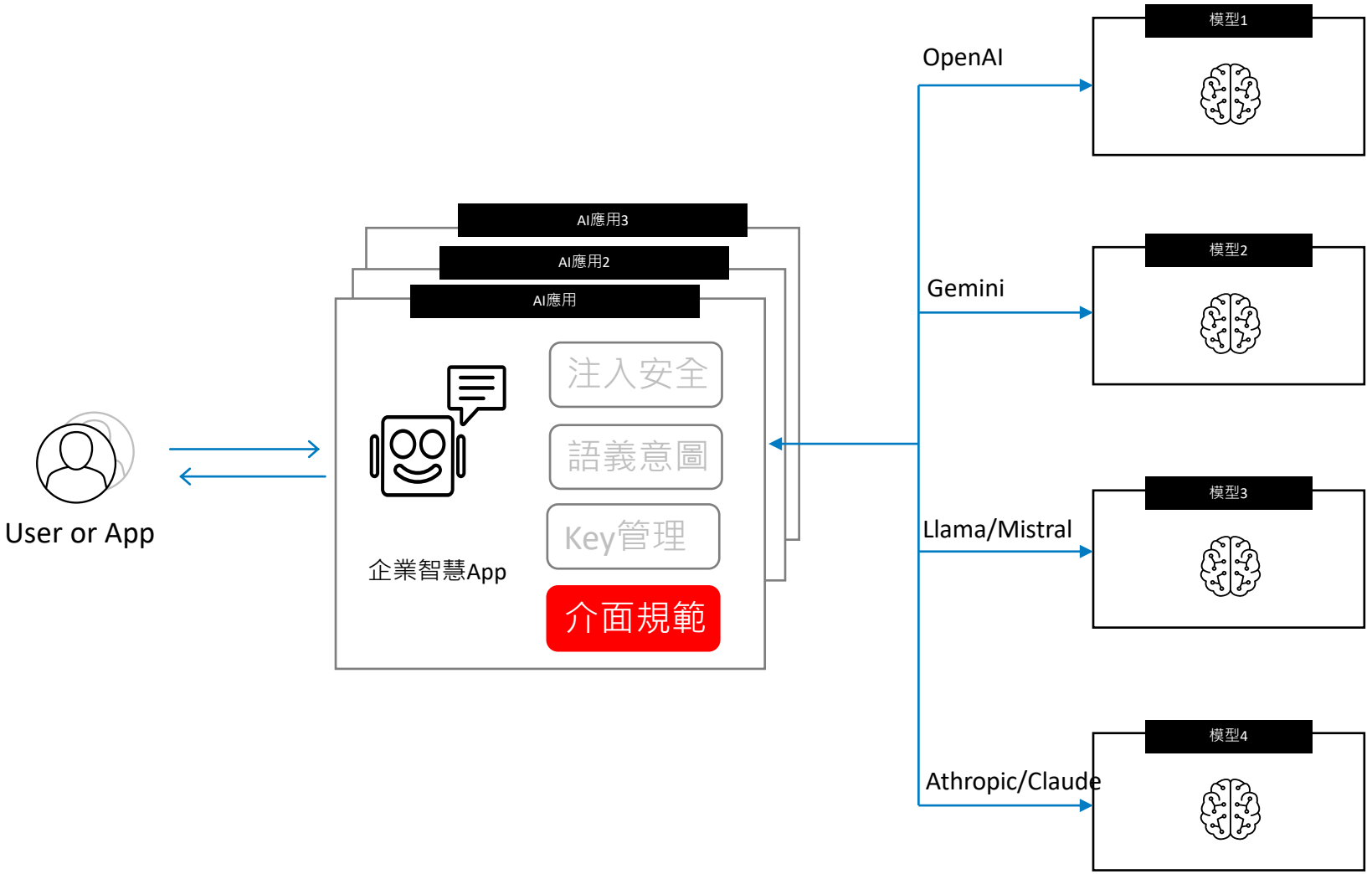
多模型的使用，需要考慮API key的易用性

多模型多API key，需建立API key統一管理機制



多模型的使用，需要考慮對接多模型介面的易用性

應用需要編寫多個API介面調用規範



- 模型的使用是基於API調用的
- 不同模型有著不同的API介面規範
- 應用需配合不同模型介面的調用
- 模型增加，介面調用配合也隨之調整
- 複雜度增加

生成式AI應用需考慮的四個重點

安全防護

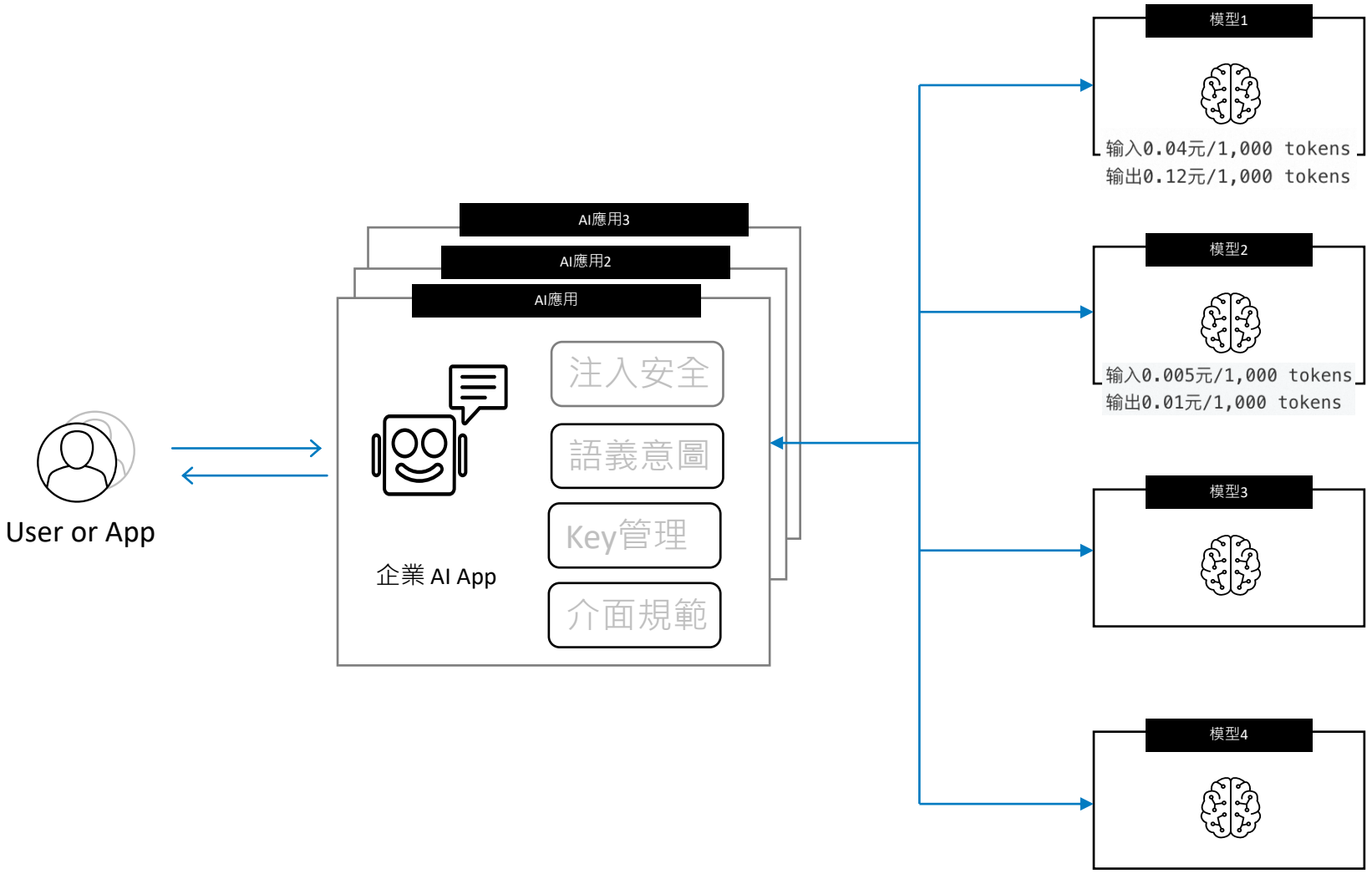
模型擴展

易用管理

成本控制

多模型的使用，需要考慮模型使用的成本控制

成本、配額、tokens限流



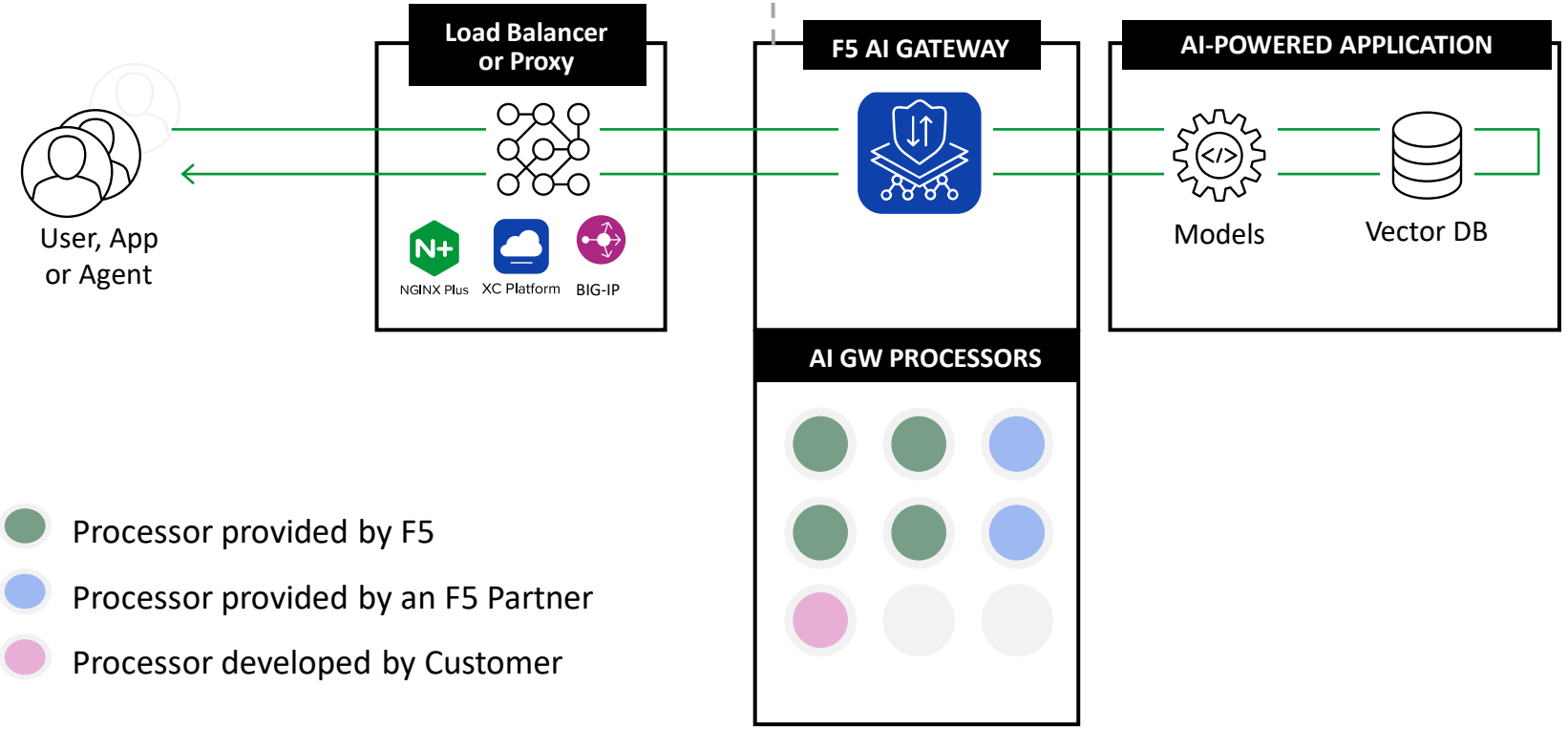
- 模型的使用是基於API調用的
- 外部或內部模型都會採用API Key
- 不同的模型有不同的價格
- API Key會有不同的RPS及配額限制

F5 AI Gateway

F5 AI Gateway

AI Gateway

- 支援Kubernetes
- *OpenAI API, Azure OpenAI, Anthropic, Ollama*
- Processor SDK (Python)



- 連線管理與失敗重試機制
- 安全防護與注入攻擊阻擋
- 成本觀測與資源使用管理
- 重複請求優化
- 模型輸出控管