ARX OS - 金融專用規格書（臺灣）

- 類別：套裝軟體（人工智慧與數據應用）
- 品名：ARDGE - ARX OS - 金融專用規格書

---

1. Integrated with APMIC Financial LLM 行業專業模型整合

1.1 透過預載內建 APMIC 金融專業大語言模型，金融從業人員能夠以自然語言即時查詢法規與市場知識、自動摘要冗長的研究報告與合約文件並輔助投研判讀，進而縮短研究工時、降低人工判讀誤差並加速決策產出。

---

2. Multi-Compatible API Endpoints 統整管理 API 端點服務

2.1 透過將各類模型統一發布為 OpenAI 與 Anthropic 相容的標準 API 端點（Endpoint），開發團隊能夠沿用既有 SDK 與工具直接串接內部系統與應用，進而免除客製整合成本、縮短導入時程，並保留日後在不同模型間無痛切換的彈性，並降低對單一模型供應商的依賴、避免供應商鎖定（vendor lock-in）。

---

3. AI Engine Platform 內建多引擎模型中台

3.1 透過支援多種推論引擎（Inference Engine / Runtime）、地端部署與 GPU 資源配置的模型中台，使用者能夠在同一平台上靈活部署與調度不同類型的模型，進而免除自建異質推論環境的整合負擔、提升 GPU 使用效率，並降低地端 AI 基礎設施的整體維運成本。

3.2 透過 GPU 共享（GPU Sharing）機制，於單一 GPU 上同時運行多個模型，並支援單一模型多次部署機制，使用者能夠充分利用閒置算力，進而提升 GPU 整體使用效率、降低單位推論成本並減少硬體採購需求。

3.3 透過模型資源預測（Model Resource Prediction），事先估算各模型的 GPU 記憶體用量，使用者能夠在部署前掌握資源需求，進而確保正式環境穩定上線、避免資源不足導致服務中斷，並為容量規劃提供依據。

3.4 透過集中式的 API Key 生命週期管理、RPM（Requests Per Minute）流量控管與使用量統計儀表板，管理者能夠依團隊或應用核發、撤銷並設定有效期限之金鑰，進而防止金鑰濫用與成本失控，確保服務穩定與資源公平分配。

---

4. Model Permission Control 模型權限控管

4.1 透過模型層級（Model-level）的細緻權限控管，使用者能夠依部門、角色與應用分別授權可調用的模型實例服務，進而落實最小權限原則、滿足金融業內控與法遵要求，並降低模型誤用與資料越權存取的風險。

---

5. Model Usage Dashboard 模型使用量儀表板

5.1 透過依模型統計請求數、Token 用量、回應時間、GPU 使用率、GPU 記憶體使用率與 GPU 記憶體分配，並支援多時間區間（即時 / 日 / 3 日 / 週）檢視的使用量儀表板，使用者能夠清楚掌握各模型的使用趨勢與效能表現，進而支援成本歸屬與容量規劃、及早發現效能劣化，並為模型汰換與資源投資決策提供數據依據。

---

6. Project-based LLM Service Isolation 專案級別模型隔離

6.1 透過依部門或專案獨立啟動模型實例（支援一對一或一對多等彈性部署）的隔離機制，使用者能夠為不同團隊配置專屬的模型服務與知識空間，進而落實資料分艙（data segregation）管理、降低跨部門資料混用與外洩風險，並符合金融業資料治理與客戶資料保護規範。

---

7. One-click Model Deployment 一鍵模型部署

7.1 透過一鍵導入 APMIC 金融模型、企業自有模型或核備開源模型（與 Hugging Face 整合，支援直接下載最新開源模型），並自動配置推論引擎與 GPU 資源的部署機制，使用者能夠在數分鐘內完成模型上線，進而大幅縮短導入時程、降低對專業 MLOps 人力的依賴，並加速 AI 應用在金融場景的落地。

---

8. Model Lifecycle Management 模型生命週期管理

8.1 透過涵蓋模型上架、版本控管、更新與多版平測的全生命週期模型管理功能，使用者能夠有序管控每一個模型版本，並在必要時手動切換回先前的穩定版本，進而建立可追溯的模型治理(Model Governance)流程、降低版本變更風險，並符合金融監理對模型管理與可稽核性的要求。

---

9. Sensitive Data Stays On-prem 敏感資料留存地端

9.1 透過將模型、文件、向量資料庫、Log 與金融資料完整保留於內網、自有機房或 Air-gapped 環境的地端架構，使用者能夠在敏感資料不外送雲端，且不依

賴公有雲端介接的前提下運行 AI 服務，進而滿足金融業資料落地與個資保護法規、降低資料外洩風險，並確保客戶資料主權完全掌握在機構自身。

---

10. GUI-based Operation Console 圖形化操作主控台

10.1 透過瀏覽器即可遠端存取的圖形化操作介面，使用者能夠集中管理模型、API 端點、API Key、App、GPU 資源、系統資源、程序監控、存取日誌、事件日誌、儲存掛載管理等狀態，進而免除繁複的指令列操作、降低維運門檻，並讓非技術背景的管理人員也能在授權範圍內安全地操作平台。

10.2 系統組態設定，透過集中式的管理控制台，提供系統資訊、一般資訊、使用者與群組、網路、安全性、終端機、更新（支援 OTA 模式遠端更新）、重置系統、通知設定與帳號等分類頁面管理，提供管理人員簡易上手與快速查找相關的組態設定，進而降低系統設定的操作門檻與人為設定錯誤，縮短組態調整與交接時間，更可降低跨據點維運成本、加速安全修補與版本更新的落地。

---

11. Admin Dashboard & System Monitoring 管理儀表板與系統監控

11.1 透過即時監控 GPU、VRAM、GPU 溫度、CPU、RAM、網路、儲存空間、磁碟溫度、程序資源使用情形、錯誤與告警的管理儀表板，使用者能夠及早發現資源瓶頸與異常徵兆，進而提前調度資源、降低服務中斷風險，並確保金融業務系統維持高可用性與穩定服務水準。

11.2 提供單一頁面集中呈現系統各項服務（包含名稱、類別、狀態及連接埠），以協助系統管理人員，快速查詢相關應用程式的服務連接埠資訊，進而加速故障排除與問題定位，縮短服務中斷的排查時間。

---

12. Security Control 安全控管

12.1 透過黑白名單、自動封鎖嘗試未授權或異常存取登入的 IP 位址與 API 存取控管，並提供 API 文件與系統日誌的安全機制，使用者能夠有效阻擋未授權存取，並完整記錄且可追溯存取行為，進而順利通過金融業資安檢查與供應鏈風險審核，並加速內部稽核與導入核准流程。

12.2 可建立不同使用帳號，帳號可指定不同的角色與權限，區隔系統與應用程式的存取範圍，管理者能夠落實職責分離（Segregation of Duties）與最小權限，進而降低內部越權與誤操作風險，並符合金融業內控與資訊管理合規要求。

12.3 透過系統日誌功能，完整記錄所有系統變更與使用者操作活動，形成可追溯的稽核軌跡（Audit Log），進而支援法規遵循與內外部稽核舉證，並在資安事件發生時提供調查依據。

12.4 透過在 LLM 安全閘道層對進出對話流量即時識別並遮罩敏感資料（含個人可識別資訊 PII 與 API 金鑰、憑證等機敏字串，支援模型回應後還原，且還原對應資訊僅存於地端），使用者能夠不必於各用戶端安裝端點防護軟體即完成集中式內容防護；敏感資料僅於記憶體（in-memory）中處理、不寫入磁碟，處理完成後即釋放；系統日誌僅記錄遮罩或代碼化後之內容，IT 管理人員無法由日誌得見原始個資或敏感資料；服務並以容器化（Docker / Kubernetes）方式隔離部署，進而落實金融業對個資與機敏資料的管控要求、降低敏感資料外洩風險，並以可一鍵安裝之 App 形式提供此模組功能。

12.5 原廠依系統版本提供 SBOM（軟體物料清單）相關佐證文件，揭露軟體組成，以符合相關資訊管理合規要求，進而加速金融機構的供應鏈資安審核與導入核准，降低第三方元件的資安盲點。

---

13. Backup & Restore 備份與還原

13.1 透過支援 App、模型與向量資料庫並可跨機器（異機）還原的備份機制，使用者能夠在硬體故障或災難情境下快速重建服務，進而縮短復原時間（RTO）、確保業務持續運作，並符合金融業營運持續與災難復原（BCP / DR）的規範要求。

---

14. Agent / MCP / CLI Integration 多重整合介面

14.1 透過支援 Agent、MCP、CLI 與 API 等多元整合介面，使用者能夠將平台串接至既有內部系統、知識庫、資料庫與金融工作流程，進而讓機構自身或第三方廠商快速建構自動化應用、擴充平台價值，並保留未來延伸與生態整合的彈性。

14.2 提供特定功能或內建應用程式之 AI Skills，並可整合目前市場主流之代理型程式設計工具（Agentic Coding Tool），提供開發人員介接整合，且無需額外訂閱費用，進而降低開發整合門檻與工具授權成本，加速自動化與 AI 應用開發並優化總體擁有成本（TCO）。

---

15. Data Integration & App Market 資料整合與 App 商店

15.1 透過將既有 NAS 等遠端儲存掛載（Remote Storage Mount）至平台，使用者能夠把靜態檔案轉化為可對話的知識資產，進而讓既有資料立即可用於 GenAI 應用、免除重複搬移與轉檔。

15.2 透過 App 商店（App Market）一鍵安裝生成式 AI 應用，使用者能夠快速部署常用 GenAI 應用，進而降低導入與開發門檻、縮短應用上線時程。

15.3 透過內建 RAG、OCR 與音視訊轉譯等資料前處理工具，使用者能夠將文件、影像與音視訊內容轉為可檢索、可問答的知識，進而提升知識問答的準確度與涵蓋範圍，並加速非結構化資料的活化利用。

ARDGE - Datacenter AI Appliance 資料中心 AI 一體機

- 類別：搭配 ARX OS 與 GPU 算力之一體機
- 品名：ARDGE ARX-300
- 目標客群：私有 AI 地端需求、大型企業、AI 資料中心、正式環境

參考之建議硬體規格：

品項	規格說明
AI Server / AI 伺服主機	NVIDIA Blackwell (GB300) 含以上
Operating System / 作業系統	ARX OS
CPU / 中央處理器	NVIDIA Grace 72-Core (ARM Neoverse V2) CPU 含以上

品項	規格說明
GPU / 圖形處理器	NVIDIA Blackwell Ultra (B300) GPU 含以上
Memory / 記憶體	Total Coherent Memory: 748GB 含以上 - System Memory: 496GB LPDDR5X - GPU Memory: 252GB HBM3e
Storage / 磁碟儲存	2 x 1TB M.2 NVMe PCIe 5.0 SSD 含以上 2 x 2TB M.2 NVMe PCIe 5.0 SSD 含以上
LAN / 網路介面	1 x 10 Gigabit Ethernet 1 x NVIDIA ConnectX-8 SmartNIC

- 附註：專案需求規劃，請先聯繫洽詢。

ARDGE - Enterprise AI Appliance 企業級 AI 一體機

- 類別：搭配 ARX OS 與 GPU 算力之一體機
- 品名：ARDGE ARX-200
- 目標客群：私有 AI 地端需求、中大型企業、正式環境

參考之建議硬體規格：

品項	規格說明
AI Server / AI 伺服主機	Rack Server 含以上
Operating System / 作業系統	ARX OS
CPU / 中央處理器	Intel Xeon CPU 含以上
GPU / 圖形處理器	2 x NVIDIA RTX PRO 6000 Blackwell 含以上
Memory / 記憶體	256 GB DDR5 含以上
Storage / 磁碟儲存	2 x 1TB M.2 NVMe PCIe 5.0 SSD 含以上 2 x 2TB M.2 NVMe PCIe 5.0 SSD 含以上
LAN / 網路介面	2 x 10 Gigabit Ethernet

- 附註：專案需求規劃，請先聯繫洽詢。

ARDGE - Edge AI Appliance 基礎版 AI 一體機

- 類別：搭配 ARX OS 與 GPU 算力之一體機
- 品名：ARDGE ARX-100
- 目標客群：私有 AI 地端需求、中小企業、開發環境

參考之建議硬體規格：

品項	規格說明
AI Server / AI 伺服主機	NVIDIA Blackwell (GB10) 含以上
Operating System / 作業系統	ARX OS
CPU / 中央處理器	ARM v9.2-A CPU 含以上
GPU / 圖形處理器	NVIDIA Blackwell GPU 含以上
Memory / 記憶體	128 GB LPDDR5x 含以上
Storage / 磁碟儲存	1TB M.2 NVMe PCIe 4.0 SSD 含以上

品項	規格說明
LAN / 網路介面	1 x 10 Gigabit Ethernet 1 x NVIDIA ConnectX-7 SmartNIC

- 附註：庫存有限，價格波動，請先聯繫洽詢。

- 最新產品消息：
  https://ardge.com/#products
- 採購下單、夥伴關係或業務聯絡：
  usasales@ardge.com

- 文件發佈版本：2026.07.15-1