ARX OS - Financial Industry Edition (Taiwan)

- Category: Packaged Software (Artificial Intelligence and Data Applications)
- Product Name: ARDGE - ARX OS - Financial Industry Edition

---

1. Integrated with APMIC Financial LLM

1.1 With the APMIC financial-domain large language model preloaded and built in, finance professionals can query regulatory and market knowledge in natural language in real time, automatically summarise lengthy research reports and contract documents, and support investment research interpretation, thereby shortening research hours, reducing human interpretation errors and accelerating decision-making.

---

2. Multi-Compatible API Endpoints

2.1 By publishing models of every kind as standard OpenAI- and Anthropic-compatible API endpoints, development teams can connect internal systems and applications directly with their existing SDKs and tools, thereby eliminating custom integration costs, shortening deployment timelines and retaining the flexibility to switch between models painlessly later, while reducing dependence on any single model vendor and avoiding vendor lock-in.

3. AI Engine Platform

3.1 With a model platform supporting multiple inference engines / runtimes, on-premises deployment and GPU resource allocation, users can flexibly deploy and orchestrate different types of models on a single platform, thereby removing the integration burden of building heterogeneous inference environments, improving GPU utilisation and lowering the overall operating cost of on-premises AI infrastructure.

3.2 Through GPU Sharing, which runs multiple models concurrently on a single GPU and supports deploying one model multiple times, users can make full use of idle compute, thereby improving overall GPU utilisation, lowering the cost per inference and reducing hardware procurement needs.

3.3 Through Model Resource Prediction, which estimates each model's GPU memory consumption in advance, users can understand resource requirements before deployment, thereby ensuring a stable production rollout, avoiding service interruptions caused by insufficient resources and providing a basis for capacity planning.

3.4 Through centralised API key lifecycle management, RPM (Requests Per Minute) rate control and a usage statistics dashboard, administrators can issue, revoke and set expiry dates for keys by team or application, thereby preventing key abuse and runaway costs, and ensuring stable service and fair resource allocation.

---

4. Model Permission Control

4.1 Through fine-grained model-level permission control, users can authorise access to model instance services separately by department, role and application, thereby enforcing the principle of least privilege, meeting financial-industry internal control and compliance requirements, and reducing the risk of model misuse and unauthorised data access.

---

5. Model Usage Dashboard

5.1 Through a usage dashboard reporting request counts, token usage, response time, GPU utilisation, GPU memory utilisation and GPU memory allocation per model, viewable across multiple time ranges (real-time / daily / 3-day / weekly), users can clearly track the usage trends and performance of each model, thereby supporting cost attribution and capacity planning, detecting performance degradation early, and providing an evidence base for model replacement and resource investment decisions.

---

6. Project-based LLM Service Isolation

6.1 Through an isolation mechanism that launches model instances independently by department or project (supporting flexible one-to-one or one-to-many deployment), users can provision dedicated model services and knowledge spaces for different teams, thereby enforcing data segregation, reducing the risk of cross-department data mixing and leakage, and complying with financial-industry data governance and customer data protection requirements.

7. One-click Model Deployment

7.1 Through one-click import of APMIC financial models, an enterprise's own models or approved open-source models (integrated with Hugging Face, supporting direct download of the latest open-source models), with the inference engine and GPU resources configured automatically, users can bring a model online within minutes, thereby greatly shortening deployment timelines, reducing reliance on specialist MLOps staff and accelerating the adoption of AI applications in financial scenarios.

---

8. Model Lifecycle Management

8.1 Through full lifecycle model management covering model publishing, version control, updates and parallel testing of multiple versions, users can govern every model version in an orderly manner and manually roll back to a previously stable version when necessary, thereby establishing a traceable Model Governance process, reducing the risk of version changes and meeting financial regulators' requirements for model management and auditability.

---

9. Sensitive Data Stays On-prem

9.1 Through an on-premises architecture that keeps models, documents, vector databases, logs and financial data entirely within the internal network, the organisation's own data centre or an air-gapped environment, users can

run AI services without sending sensitive data to the cloud and without depending on public cloud connectivity, thereby satisfying financial-industry data residency and personal data protection regulations, reducing the risk of data leakage and ensuring that sovereignty over customer data remains entirely with the institution.

---

10. GUI-based Operation Console

10.1 Through a graphical operation interface reachable remotely from a browser, users can centrally manage models, API endpoints, API keys, apps, GPU resources, system resources, process monitoring, access logs, event logs and storage mount management, thereby removing the need for complex command-line operations, lowering the operational barrier and allowing administrators without a technical background to operate the platform safely within their authorised scope.

10.2 System configuration: a centralised management console provides category pages for system information, general information, users and groups, network, security, terminal, updates (supporting remote updates in OTA mode), system reset, notification settings and accounts, so that administrators can get started easily and locate the relevant configuration settings quickly, thereby lowering the barrier to system configuration and the chance of human misconfiguration, shortening configuration and handover time, and further reducing multi-site operating costs while accelerating the rollout of security patches and version updates.

---

11. Admin Dashboard & System Monitoring

11.1 Through a management dashboard that monitors GPU, VRAM, GPU temperature, CPU, RAM, network, storage, disk temperature, per-process resource usage, errors and alerts in real time, users can detect resource bottlenecks and abnormal signs early, thereby reallocating resources ahead of time, reducing the risk of service interruption and keeping financial business systems highly available at a stable service level.

11.2 A single page presents all system services centrally (including name, category, status and port) to help system administrators quickly look up the service port information of the relevant applications, thereby speeding up troubleshooting and fault localisation and shortening the time spent diagnosing service interruptions.

---

12. Security Control

12.1 Through allowlists and blocklists, automatic blocking of IP addresses that attempt unauthorised or abnormal access and logins, and API access control, together with security mechanisms for API documentation and system logs, users can effectively block unauthorised access while recording access behaviour completely and traceably, thereby passing financial-industry information security inspections and supply chain risk reviews smoothly, and accelerating internal audit and deployment approval processes.

12.2 Separate user accounts can be created and assigned different roles and permissions to separate the access scope of the system and of applications, so administrators can enforce Segregation of Duties and least privilege,

thereby reducing the risk of internal privilege abuse and operational error, and meeting financial-industry internal control and information management compliance requirements.

12.3 Through system logging, all system changes and user operation activity are recorded in full to form a traceable Audit Log, thereby supporting regulatory compliance and evidence for internal and external audits, and providing a basis for investigation should a security incident occur.

12.4 By identifying and masking sensitive data in real time at the LLM security gateway layer, on both inbound and outbound conversation traffic (covering personally identifiable information (PII) and sensitive strings such as API keys and credentials, with support for restoring the original values after the model responds, and with the restoration mapping held only on premises), users can achieve centralised content protection without installing endpoint protection software on every client. Sensitive data is processed in memory only, is never written to disk and is released as soon as processing completes; system logs record only masked or tokenised content, so IT administrators cannot see the original personal or sensitive data in the logs. The service is also deployed in isolation using containers (Docker / Kubernetes), thereby meeting the financial industry's control requirements for personal and sensitive data, reducing the risk of sensitive data leakage, and delivering this module as a one-click installable app.

12.5 For each system version the manufacturer provides SBOM (Software Bill of Materials) supporting documentation disclosing software composition, in order to meet the relevant information management compliance requirements, thereby accelerating financial institutions' supply chain security reviews and deployment approvals, and reducing security blind spots arising from third-party components.

13. Backup & Restore

13.1 Through a backup mechanism that covers apps, models and vector databases and supports restoring onto a different machine, users can rebuild services quickly in hardware failure or disaster scenarios, thereby shortening recovery time (RTO), ensuring business continuity and meeting financial-industry business continuity and disaster recovery (BCP / DR) requirements.

---

14. Agent / MCP / CLI Integration

14.1 By supporting diverse integration interfaces including Agent, MCP, CLI and API, users can connect the platform to existing internal systems, knowledge bases, databases and financial workflows, thereby letting the institution itself or third-party vendors build automated applications rapidly, extending the value of the platform and preserving flexibility for future extension and ecosystem integration.

14.2 AI Skills are provided for specific features and built-in applications, and can be integrated with the mainstream agentic coding tools currently on the market for developers to connect to, with no additional subscription fee, thereby lowering the barrier to development and integration as well as tool licensing costs, accelerating automation and AI application development, and optimising total cost of ownership (TCO).

---

15. Data Integration & App Market

15.1 By mounting existing remote storage such as NAS onto the platform (Remote Storage Mount), users can turn static files into conversational knowledge assets, thereby making existing data immediately usable for GenAI applications and removing repeated data migration and format conversion.

15.2 Through one-click installation of generative AI applications from the App Market, users can deploy commonly used GenAI applications quickly, thereby lowering the barrier to adoption and development and shortening the time needed to bring applications online.

15.3 Through built-in data preprocessing tools such as RAG, OCR and audio/video transcription, users can turn documents, images and audio/video content into searchable, question-answerable knowledge, thereby improving the accuracy and coverage of knowledge Q&A and accelerating the activation of unstructured data.

ARDGE - Datacenter AI Appliance

- Category: Appliance bundled with ARX OS and GPU compute
- Product Name: ARDGE ARX-300
- Target customers: private on-premises AI requirements, large enterprises, AI data centres, production environments

Reference recommended hardware specifications:

Item	Specification
AI Server	NVIDIA Blackwell (GB300) or above
Operating System	ARX OS
CPU	NVIDIA Grace 72-Core (ARM Neoverse V2) CPU or above
GPU	NVIDIA Blackwell Ultra (B300) GPU or above
Memory	Total Coherent Memory: 748GB or above - System Memory: 496GB LPDDR5X - GPU Memory: 252GB HBM3e
Storage	2 x 1TB M.2 NVMe PCIe 5.0 SSD or above 2 x 2TB M.2 NVMe PCIe 5.0 SSD or above

Item	Specification
LAN	1 x 10 Gigabit Ethernet 1 x NVIDIA ConnectX-8 SmartNIC

- Note: for project requirement planning, please contact us first.

ARDGE - Enterprise AI Appliance

- Category: Appliance bundled with ARX OS and GPU compute
- Product Name: ARDGE ARX-200
- Target customers: private on-premises AI requirements, medium and large enterprises, production environments

Reference recommended hardware specifications:

Item	Specification
AI Server	Rack Server or above
Operating System	ARX OS

Item	Specification
CPU	Intel Xeon CPU or above
GPU	2 x NVIDIA RTX PRO 6000 Blackwell or above
Memory	256 GB DDR5 or above
Storage	2 x 1TB M.2 NVMe PCIe 5.0 SSD or above 2 x 2TB M.2 NVMe PCIe 5.0 SSD or above
LAN	2 x 10 Gigabit Ethernet

- Note: for project requirement planning, please contact us first.

ARDGE - Edge AI Appliance

- Category: Appliance bundled with ARX OS and GPU compute
- Product Name: ARDGE ARX-100

- Target customers: private on-premises AI requirements, small and medium enterprises, development environments

Reference recommended hardware specifications:

Item	Specification
AI Server	NVIDIA Blackwell (GB10) or above
Operating System	ARX OS
CPU	ARM v9.2-A CPU or above
GPU	NVIDIA Blackwell GPU or above
Memory	128 GB LPDDR5x or above
Storage	1TB M.2 NVMe PCIe 4.0 SSD or above

Item	Specification
LAN	1 x 10 Gigabit Ethernet 1 x NVIDIA ConnectX-7 SmartNIC

- Note: stock is limited and prices fluctuate; please contact us first.

- Latest product news:
  https://ardge.com/#products
- Purchase orders, partnership or business enquiries:
  usasales@ardge.com

- Document release version: 2026.07.15-1