Laiye AI | Red Hat

Create a secure, consistent and scalable AI environment – from edge to core

Enterprise AI can be powerful, but managing it is complex

As more organisations adopt AI, IT teams face mounting complexity: balancing resources while training models, optimising performance, monitoring systems, and managing recovery across dispersed environments.

They must debug without disruption, detect anomalies, integrate applications with data sources, and deliver a cloud-like experience – yet still struggle to maintain consistency across solutions from a single ISV.

Technology used

- Red Hat OpenShift AI
- NVIDIA AI Enterprise

Simplify your AI journey with Laiye AI Foundry

Laiye AI Foundry, co-developed with Red Hat and NVIDIA, is an integrated, out-of-the-box solution that streamlines enterprise AI adoption at scale.

It reduces operational complexity and accelerates time to value with pre-configured tools, powerful infrastructure, and MLOps capabilities.

- **AI-ready from day one –** Pre-integrated Red Hat and NVIDIA stack for fast deployment.
- **Faster model development –** Built-in MLOps and full lifecycle control.
- **Optimised performance –** Intelligent tuning for CPU, GPU and DPU resources.
- **Consistent environments –** Unified across cloud, edge and on-prem environments.
- **Seamless integration –** Certified plug-ins for broad application compatibility.
- **Trusted and proven –** Built on validated, widely adopted solutions.
- **Edge or datacentre AI –** Flexible and scalable for diverse use cases and industries.
- **Rich model ecosystem –** Plug-and-play support for commercial, semi-commercial and open-source models.

How the solution addresses common AI challenges

Category	Key Challenge	Solution
Development	OS stability under extreme loads (e.g. heat, resource pressure, node failure)	Stable performance at large scale (up to 10,000 CPU cards)
	Keeping up with frequent NIM service updates and optimising low-level configurations	Seamless optimisation and delivery of latest NIM services
Operations	Difficulties managing, recovering, and maintaining 10,000 GPU cards across hybrid environments	End-to-end GPU card support at massive scale
	Complex GPU and Infiniband tuning required for peak AI performance	Custom tuning for GPU, network, and OS performance
Management	Fragmented AI resources across on-prem, cloud, and regions make coordination difficult	Unified cross-region support and hybrid resource integration

Built by leaders in AI innovation

Laiye AI Foundry brings together the strengths of three leading technology providers: **Red Hat, NVIDIA and Laiye,** to solve the most complex challenges in enterprise AI. Each brings deep domain expertise and a commitment to delivering scalable, enterprise-ready solutions.

Red Hat

Red Hat OpenShift AI is at the heart of this solution: an enterprise-grade container platform optimised for AI workloads. It runs securely in hybrid environments, supports CI/CD pipelines for faster model development and offers tools that accelerate development and monitoring.

Red Hat's AI/ML platform for predictive and gen AI applications

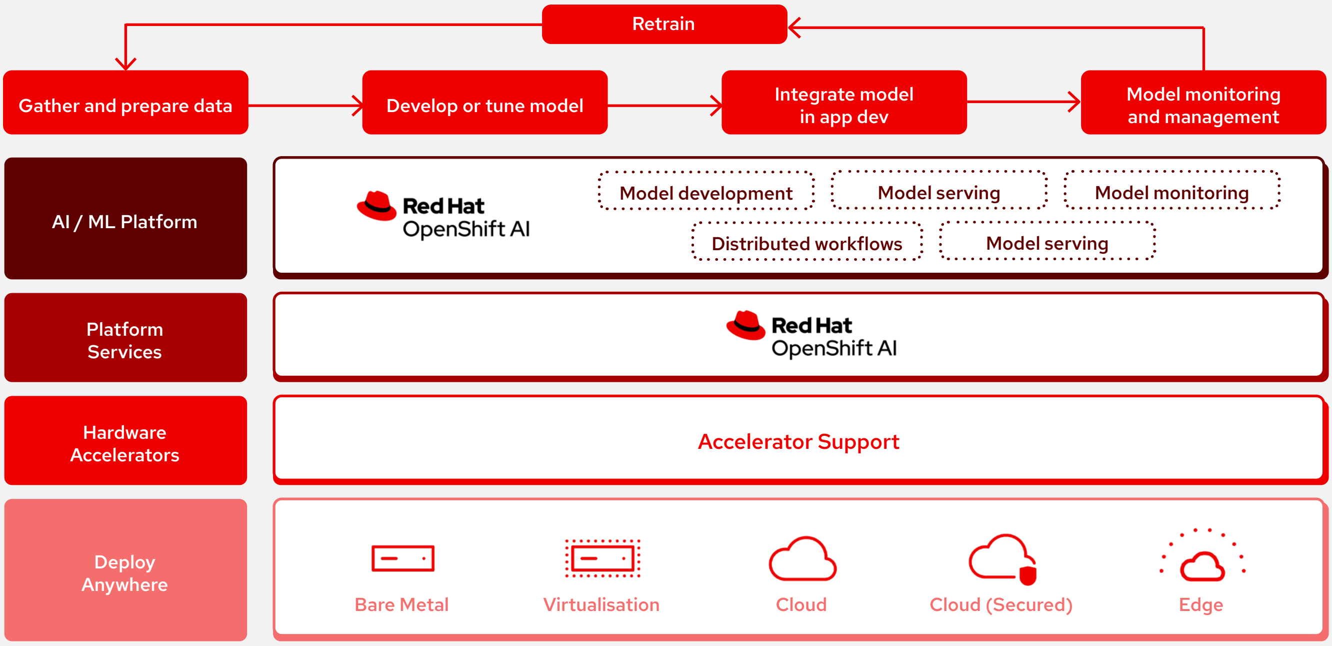


Figure 1: The Red Hat OpenShift AI platform

NVIDIA

NVIDIA is the world's leading provider of GPU technology, powering the most demanding AI workloads across industries. NVIDIA AI Enterprise provides containerised tools to deploy generative AI and ML quickly across cloud, data centre, or edge; fully integrated with open-source frameworks.

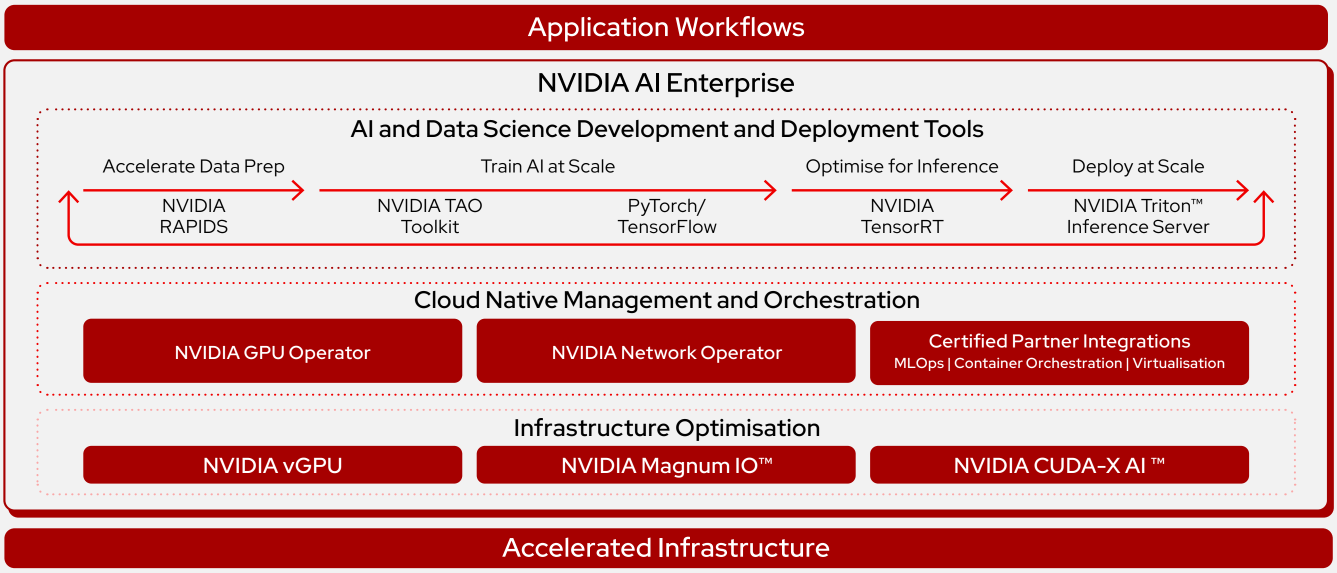


Figure 2: NVIDIA AI Enterprise

Beijing Laiye Information Technology Co., Ltd.

Laiye is a full-stack LLM technology company with expertise spanning high-performance computing, robotics, NLP and autonomous systems. Backed by talent from top-tier R&D organisations including NVIDIA, Qualcomm, Huawei and Alibaba - it delivers out-of-the-box AI solutions; reducing cost and complexity without compromising performance.

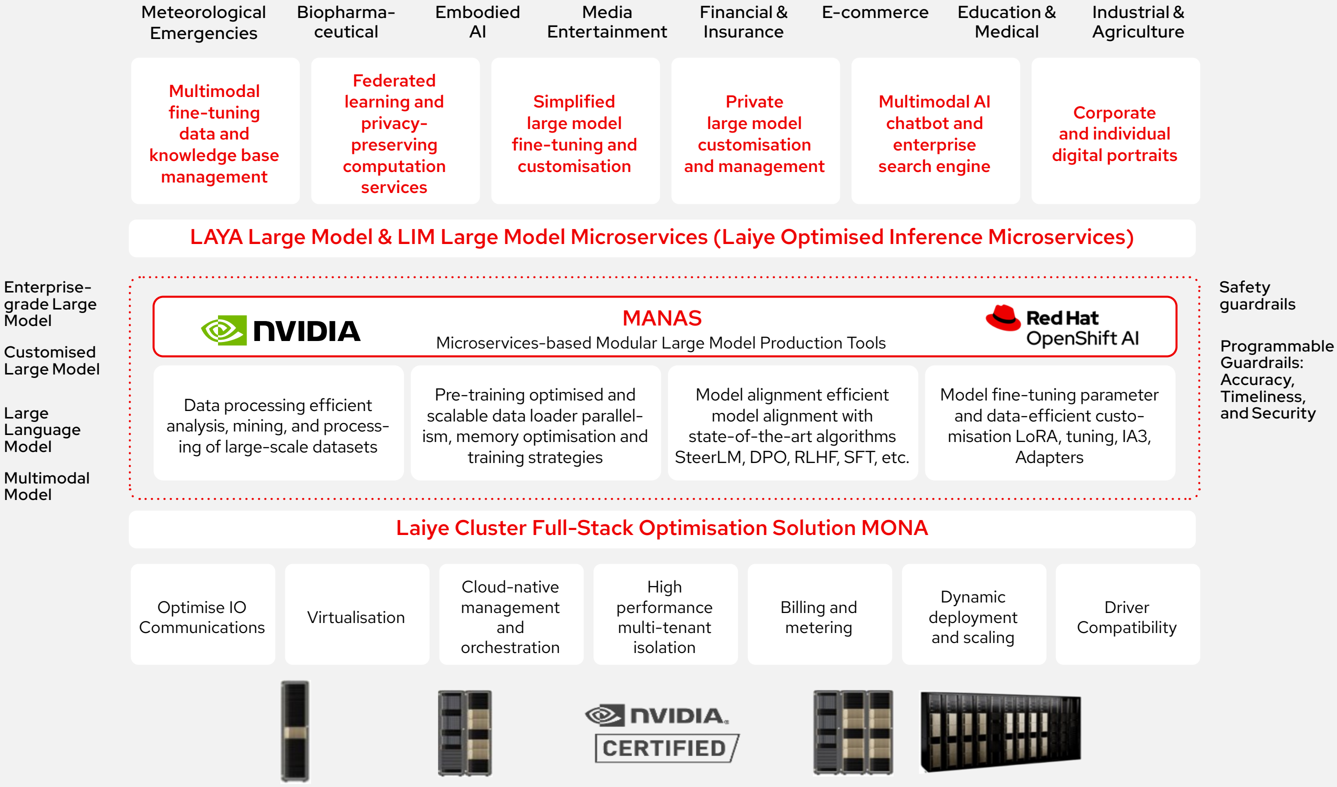

Meteorological Emergencies
Biopharma-ceutical
Embodied AI
Media Entertainment
Financial & Insurance
E-commerce
Education & Medical
Industrial & Agriculture
Multimodal fine-tuning data and knowledge base management
Federated learning and privacy-preserving computation services
Simplified large model fine-tuning and customisation
Private large model customisation and management
Multimodal AI chatbot and enterprise search engine
Corporate and individual digital portraits
LAYA Large Model & LIM Large Model Microservices (Laiye Optimised Inference Microservices)
Enterprise-grade Large Model
Customised Large Model
Large Language Model
Multimodal Model
NVIDIA
MANAS
Microservices-based Modular Large Model Production Tools
Red Hat OpenShift AI
Data processing efficient analysis, mining, and processing of large-scale datasets
Pre-training optimised and scalable data loader parallelism, memory optimisation and training strategies
Model alignment efficient model alignment with state-of-the-art algorithms SteerLM, DPO, RLHF, SFT, etc.
Model fine-tuning parameter and data-efficient customisation LoRA, tuning, IA3, Adapters
Safety guardrails
Programmable Guardrails: Accuracy, Timeliness, and Security
Laiye Cluster Full-Stack Optimisation Solution MONA
Optimise IO Communications
Virtualisation
Cloud-native management and orchestration
High performance multi-tenant isolation
Billing and metering
Dynamic deployment and scaling
Driver Compatibility
NVIDIA CERTIFIED

Figure 3: A high-level reference architecture view of Laiye AI Foundry

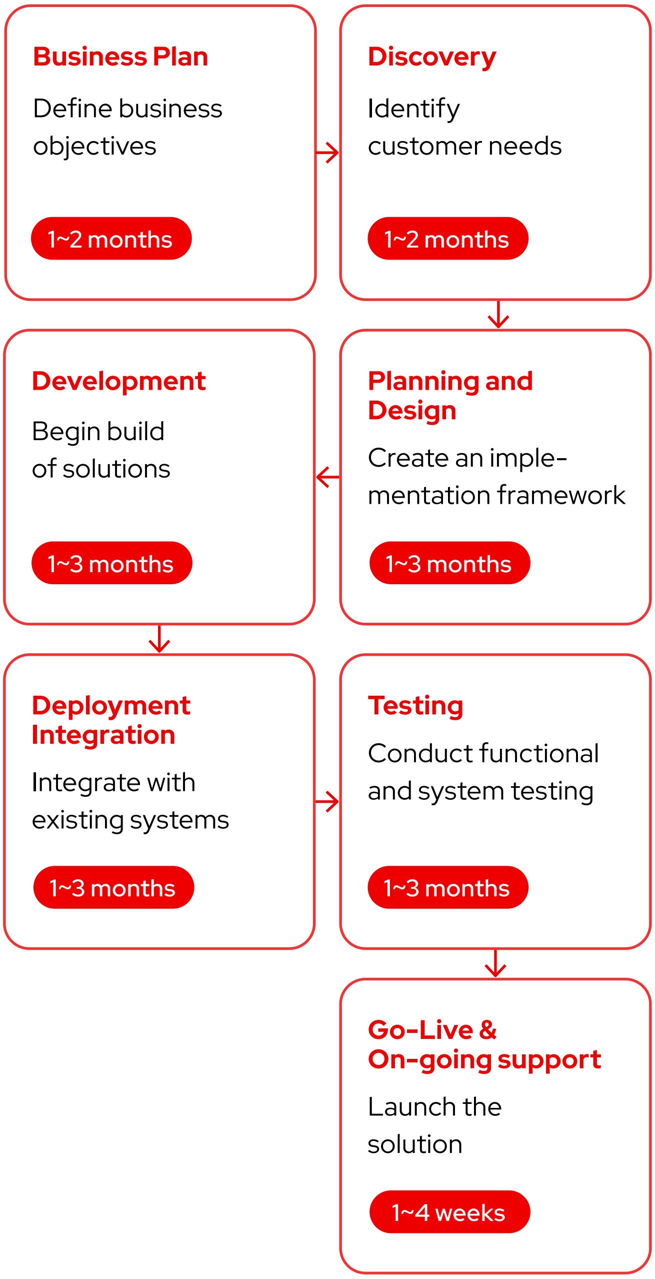
Success story: Fast, efficient AI services

A group customer set out to deploy a private, sovereign AI platform to enable internal teams and clients to train and run customised generative AI models.

During deployment, they faced challenges like the inefficient usage of RDMA networks for distributed training and underperforming model inference services.

By adopting Red Hat OpenShift AI, incorporating acceleration components from NVIDIA's NVAIE, and applying model optimisation solutions from Laiye, they built a high-performance environment for distributed training and inference; delivering fast, efficient AI services tailored to their needs.

Fast, seamless implementation



Business Plan
Define business objectives
1~2 months

Discovery
Identify customer needs
1~2 months

Planning and Design
Create an implementation framework
1~3 months

Development
Begin build of solutions
1~3 months

Deployment Integration
Integrate with existing systems
1~3 months

Testing
Conduct functional and system testing
1~3 months

Go-Live & On-going support
Launch the solution
1~4 weeks

To learn more on how this solution can accelerate your AI journey, reach out to us:

Red Hat
apj-cdsm-practice@redhat.com

Laiye
sales@laiye.ai

About Red Hat

Red Hat helps customers standardise across environments, develop cloud-native applications, and integrate, automate, secure, and manage complex environments with award-winning support, training, and consulting services.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

North America
1-888-REDHAT1
www.redhat.com

Europe, Middle East, and Africa
00800 7334 2835
europe@redhat.com

Asia Pacific
+65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com

Copyright © 2025 Red Hat, Inc. Red Hat and the Red Hat logo are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.