THAKI Cloud

METIS

Contents

1.
Customer Pain Points

Customer Pain Points

AI Infrastructure Operations: Cost and Complexity Are Blocking Enterprise Growth

Single-vendor lock-in, workload optimization burden, operational difficulty, and data sovereignty regulations are major barriers to enterprise AI transformation.

GPU-Centric TCO Surge

- Single-vendor dependency
- Supply chain constraints
- Rapidly increasing infrastructure costs

Hardware Optimization Complexity

- Different optimal accelerator combinations required for each workload (training, inference, fine-tuning)

Multi-Cluster Recovery Management

- Downtime occurs due to manual workload switching during cluster failures.

AI Platform Operations Burden

- Scalable operation challenges for multi-cluster, multi-tenant, and multi-agent systems

Data Sovereignty & Compliance

- Local data residency requirements
- Industry-specific regulations
- On-premise/sovereign cloud deployment needs

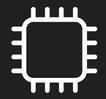
HPC-AI Workload Separation

- Slurm-based HPC and Kubernetes-based AI are separated, leading to lower GPU utilization

2. Metis

Metis

Cloud-Native AI Platform for Unified Multi-xPU, Multi-Cluster Operations



xPU Heterogeneous Orchestration

- Cloud-native AI/ML platform that efficiently orchestrate various accelerators (GPU/NPU/TPU)

Complexity Abstraction

- Improved developer experience by eliminating Kubernetes, multi-cluster scheduling, and xPU management complexity

End-to-End Workflow

- Unified platform for inference, fine-tuning, and evaluation supporting the entire AI lifecycle

Developer-Centric DX

- Self-service portal, templates, and agents
- Workflows that let you focus on AI development, instead of infrastructure

Hybrid/Sovereign Cloud

- Data sovereignty guaranteed with priority support for on-premise data centers and domestic cloud providers

Project-Based Resource Management

- Resources organized by group/project for project development efficiency

3. Competitive Differentiation

Competitive Differentiation

Metis Provides Clear Technical and Operational Differentiation

Category	Competitor Strengths	Competitor Limitations	THAKI CLOUD
Global LLM/API	• SOTA models • Global infrastructure • Managed services (RAG, Agents, Evaluation)	• Limited on-prem/sovereign options • Lack of xPU abstraction • Data sovereignty/cost for regulated workloads	• Control plane that operates regardless of xPU architecture and usage • GPU/NPU hosting with on-prem and sovereign cloud support
GPU LLM Infrastructure	• High-performance GPU infrastructure Serverless/dedicated endpoints • Strong developer experience (DX)	• Mostly GPU-centric • Own cloud-centric • Multi-cluster/BYOC and domestic NPU integration limitations	• Model platform UX + xPUaaS multi-cluster + BYOC architecture • Domestic NPU and on-prem integration
Domestic CSP/AI	• Domestic regions and compliance • Strong enterprise sales/support • Public/financial references	• Fragmentation between AI service products • Limited heterogeneous accelerator integration • Lack of LLM platform experience	• Unified AI platform layer on top of CSP • Consistent AI platform experience • Local compliance/data residency inheritance
Backend Performance	• Overcoming limitations of synchronous processing • Performance tuning and reduction of expansion costs	• Synchronous processing	• Go Event-Driven • >10,000 events/sec, <1 sec latency
Deloyment Standardization	• Reduced deployment time • increased onboarding speed for new personnel • rollback capability during failures	• Manual environment configuration	• Docker template-based one-click deployment • 100% reproducibility

4.
Platform Positioning & Target Users

Platform Positioning & Target Users

Optimal Development & Operations Environment for Engineers

Metis spans the entire IaaS-PaaS-SaaS stack, providing an optimal development and operations environment for Data Scientists, ML Engineers, and Platform/DevOps Engineers.

Service Model

SaaS
Specialized AI Applications

PaaS
Metis

IaaS
K8s Clusters, GPU/NPU/xPU Infrastructure

Target Personas

Data Scientists

Requirements
Fast experimentation, reproducibility, easy dataset/model versioning

Thaki AI Platform Provides
Jupyter/Pod templates, dataset management, leaderboards, experiment templates

ML Engineers

Requirements
Model optimization, deployment pipelines, performance monitoring

Thaki AI Platform Provides
Fine-tuning studio, model registry, A/B testing, auto-scaling endpoints

Platform/DevOps Engineers

Requirements
Infrastructure automation, resource management, security

Thaki AI Platform Provides
Multi-cluster management, quota policies, observability stack, CI/CD integration

5.
Metis
Product Components

Thaki AI Platform Product Components

Serverless Interface: Fully Managed Usage-Based Inference

A fully managed inference layer that requires no infrastructure management and charges based on usage.

OpenAI-Compatible API & Model Support

- OpenAI-compatible API for easy migration from closed providers
- Support for open-source and multimodal models

Auto Scaling

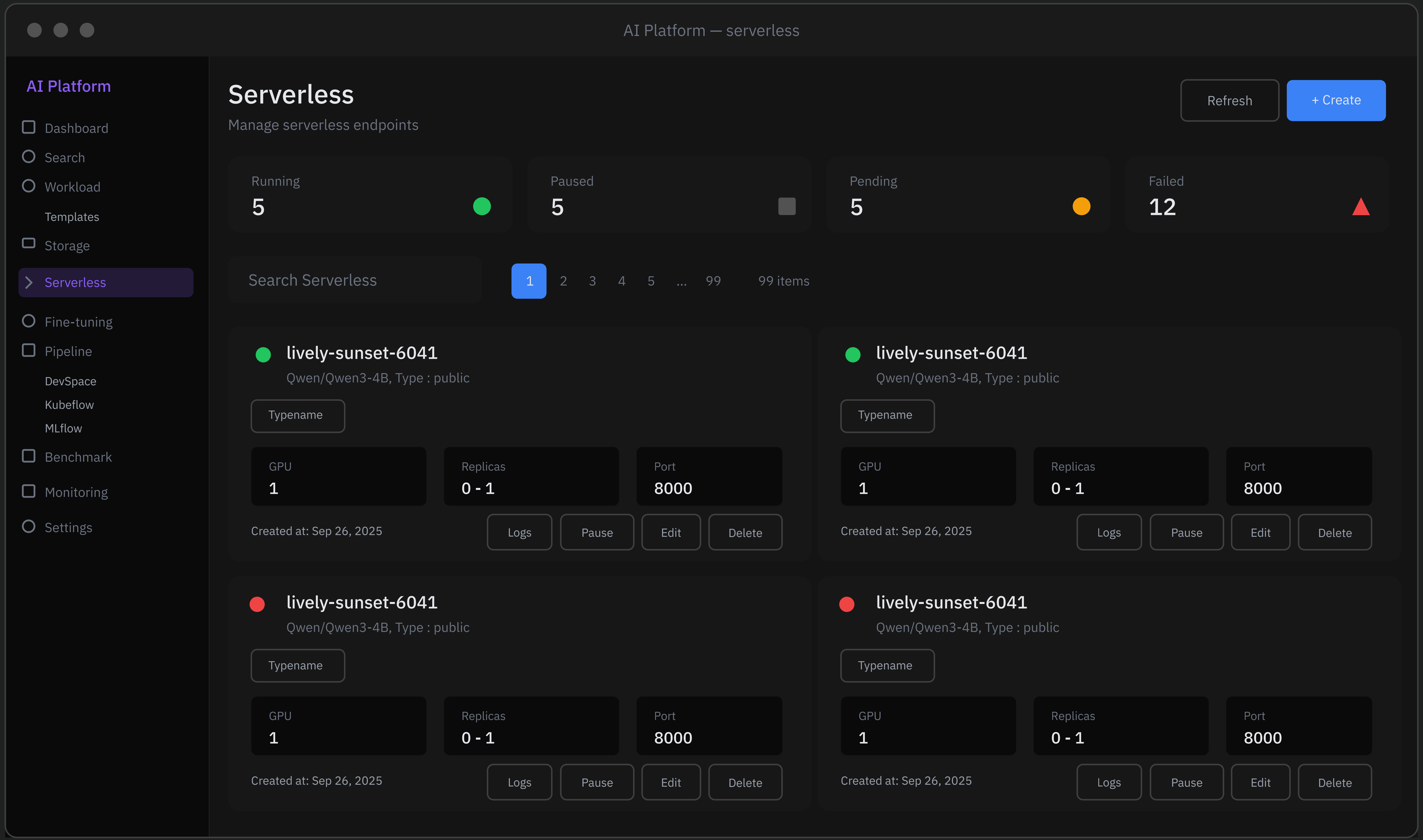
- Infrastructure optimization that automatically scales based on tokens-per-second throughput and request volume

vLLM-Based Engine

- Optimal performance guaranteed through high throughput
- Low latency, and efficient KV cache utilization

Reduced Management Burden & Fast Prototyping/Serving

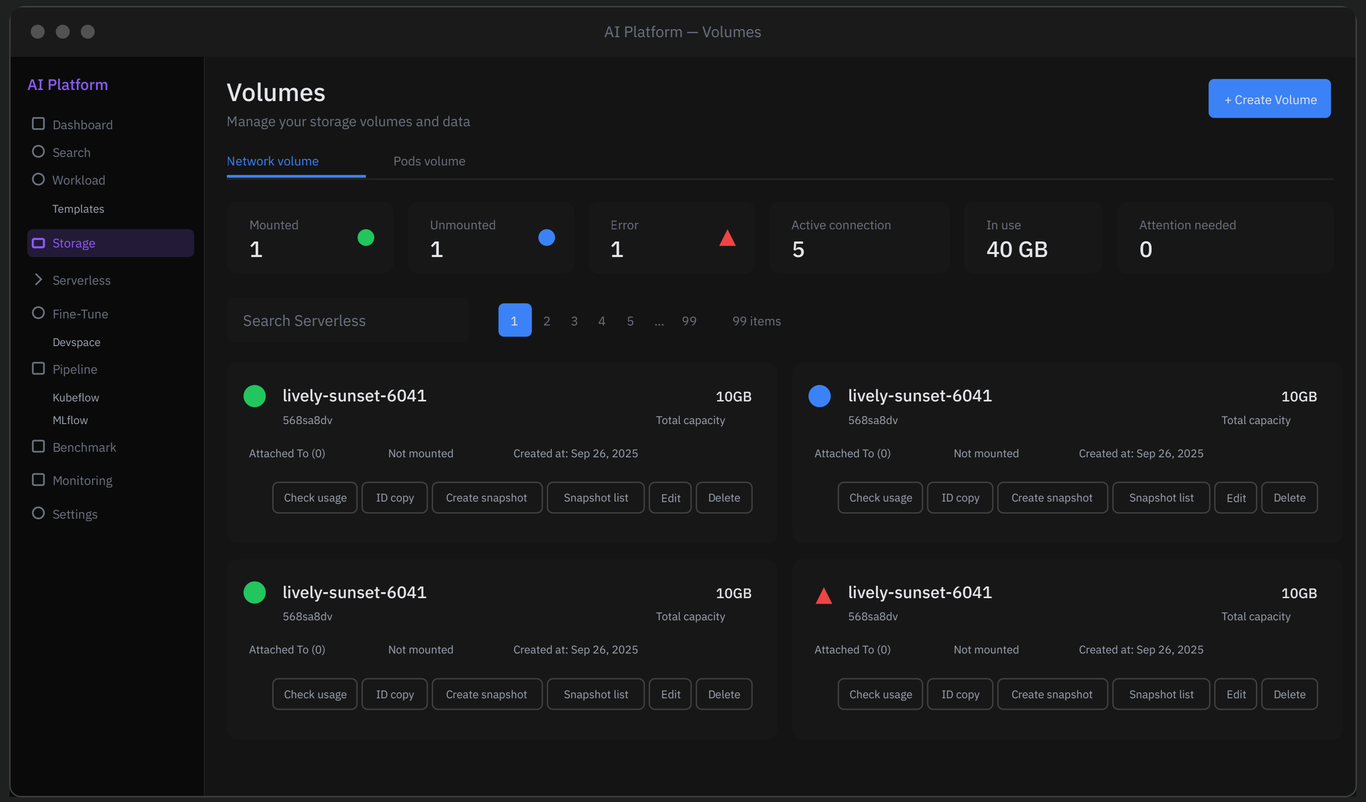
- No infrastructure management burden
- Fast prototyping and production-grade serving on the same unified stack



Metis Product Components

Dedicated Endpoints: Consistent Performance with Dedicated xPU Capacity

Dedicated GPU/xPU nodes for high-priority or latency-sensitive services.



Dedicated Nodes, VPC/Private Options

- Independent network environment
- Isolated infrastructure for security-critical workloads

SLA: Availability, Latency & Capacity Guarantees

- Enterprise-grade service level agreements with uptime, latency, and capacity guarantees

Fine-Grained Control of Version/Scale Range/Rollout

- Detailed configuration for model versions, scaling limits, and deployment strategies

Predictable Performance & Cost

- Consistent performance and clear cost structure in a stable production environment

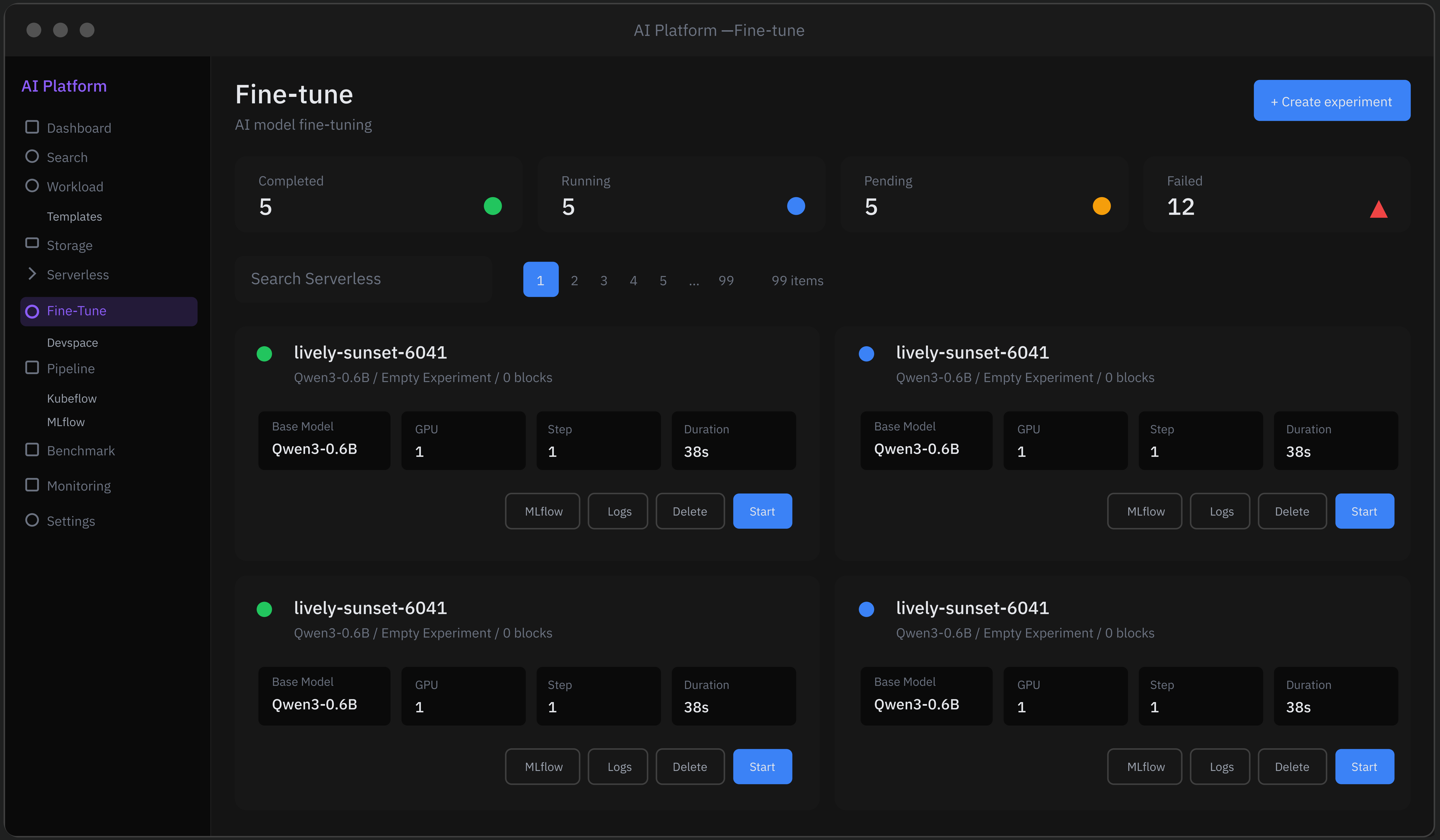
Multi-Cluster Deployment

- Enables simultaneous deployment to multiple clusters through Virtual Cloud.

Metis Product Components

Fine-tuning Studio: Enterprise-Grade Fine-tuning

An advanced platform for enterprise-specific AI model fine-tuning.



SFT/DPO/GRPO, LoRA/QLoRA, Distributed Training

- Support for various latest fine-tuning techniques
- Efficient distribution training across multiple GPUs

PyTorch+HF, Job Templates Provided

- Validated job templates for chat, instruction-following, RAG, and domain-specific models

Kueue/Kai Scheduling: Fair & Efficient Allocation

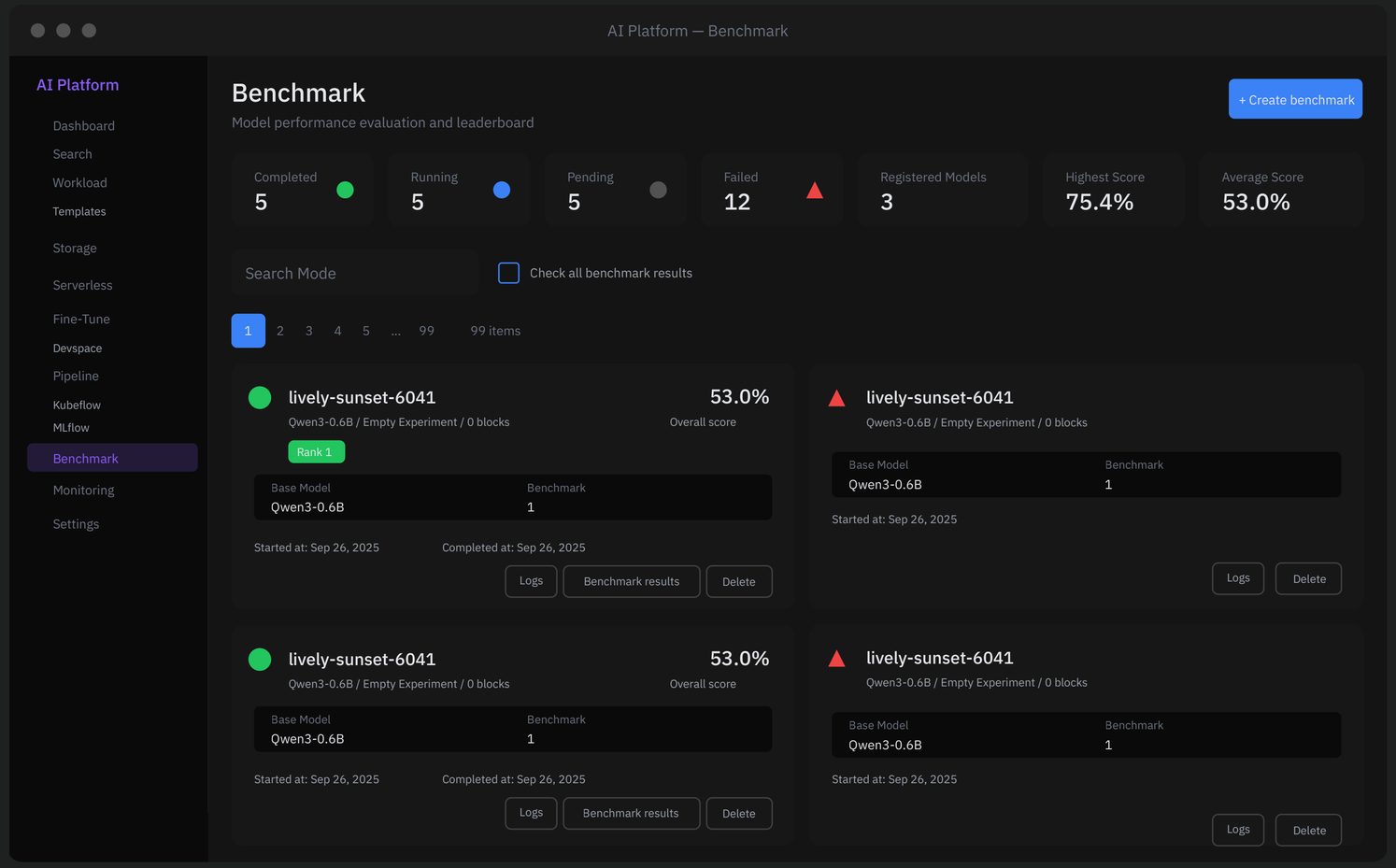
- Unified resource scheduling ensuring fair and efficient GPU allocation

One-Click Deployment: Serverless/Dedicated Endpoints

- Instant deployment of fine-tuned models to serverless inference or dedicated endpoints

Evaluations & Guardrails: Quality Measurement & Safety/Compliance Enforcement

A comprehensive toolset for measuring and ensuring model quality and regulatory compliance.



Model/Prompt A/B Testing

- Automated scoring based on latency, cost, quality metrics, and task-specific KPIs

HITL Evaluation Workflow

- Human expert-based evaluation system for subjective tasks

Content Filters & Guardrails

- Automated safeguards for safety checks, policy-based restrictions, and regulatory compliance

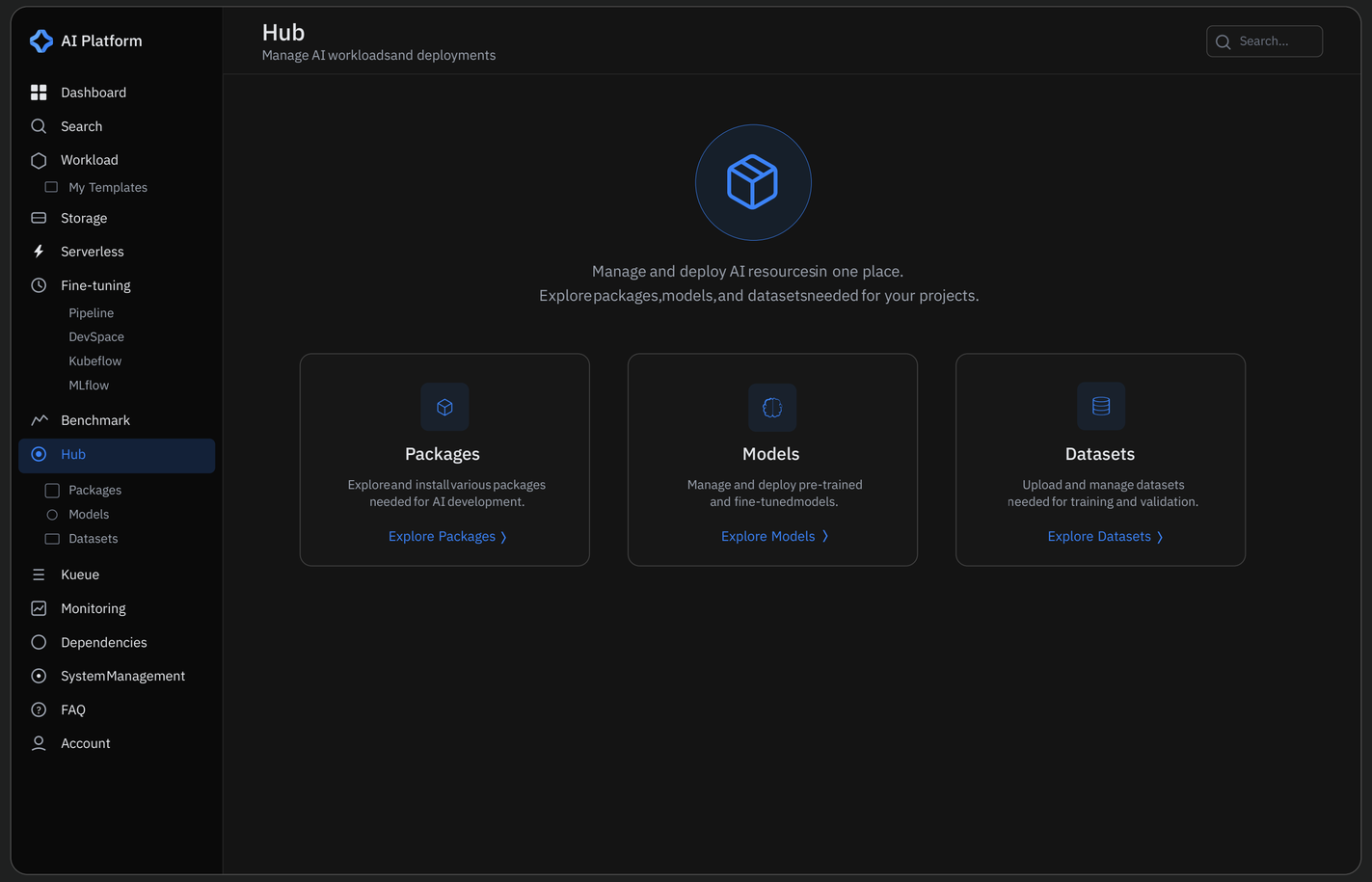
Data-Driven Decision Making

- Optimize model/prompt selection and reduce production deployment risks

Thaki AI Platform Product Components

AI Platform HUB: Centralized AI Resource Management

An integrated AI resource hub for exploring, managing, and deploying packages, models, and datasets needed for projects in one place.



Hub Features

- Centrally explore, manage, and version packages, models, and datasets by project
- Deploy required resources in a consistent manner

Resource Discovery

- Search and access the latest AI packages, internal/open-source models, and datasets in one place

Easy Utilization

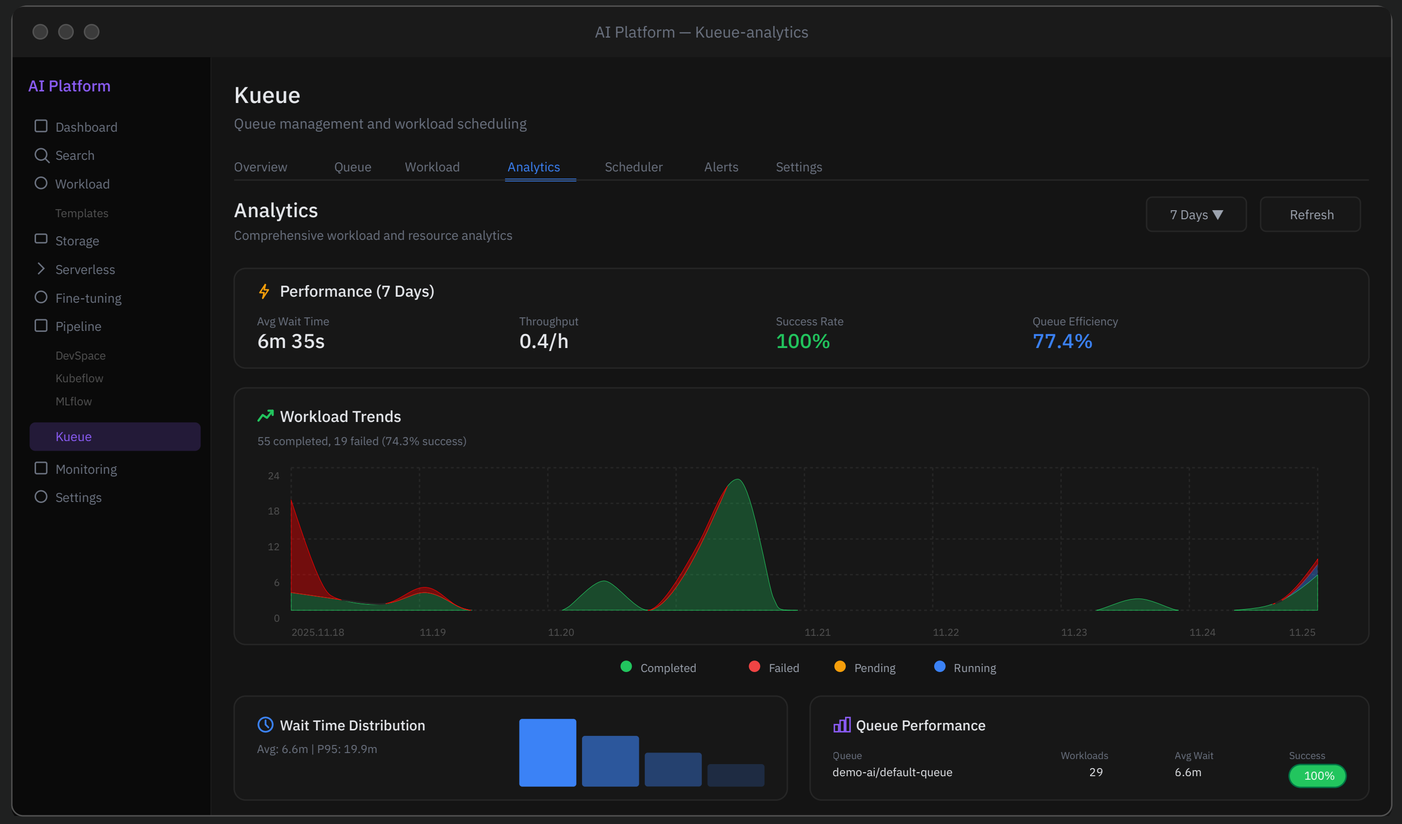
- Immediately apply packages, models, and datasets to training/inference workflows without environment setup

6. Observability, Security & Operations

Observability, Security & Operations

Observability, Security, Compliance & Deployment Automation for Enterprise AI Operations

Platform-embedded capabilities covering all areas for stable AI operations:
xPU resource metrics monitoring, authentication/authorization, network isolation, and CI/CD automation.



Observability

- Metrics, logs, and trace collection/monitoring
- Prometheus, Grafana, Loki, Jaeger/OpenTelemetry

Key Metrics

- xPU utilization, token throughput, latency (P50/P95/P99), error rate monitoring

Security & Authentication

- JWT-based authentication
- Roadmap: RBAC, SSO (SAML/OIDC), HashiCorp Vault integration

Isolation & Compliance

- Namespace isolation, Kubernetes network policies, encrypted secrets management

Deployment & CI/CD

- Helm chart-based component management, GitHub Actions + ArgoCD GitOps automation pipeline

7.
Business Impact

Business Impact

AI Products

Metis delivers reduced time-to-market, infrastructure cost savings, and 99.9% stability through model platform, xPU infrastructure, and multi-cluster operations optimization.

Time to Market (TTM)	Infrastructure Cost	Reliability
↓ 30-50%	↓ 20-40%	↑ 99.9%

Accelerated Time to Market

- Rapid AI product launch and adoption through turnkey model platform and GPU cloud

Unit Cost & Cost Efficiency

- TCO reduction through xPUaaS, combinatorial scheduling, MIG, and spot instance strategies

Operational Stability & Excellence

- Kubernetes-native, multi-cluster support design with robust observability and automation

Appendix #1
High-Level Architecture

Appendix #1

High-Level Architecture

Frontend

React · TypeScript · Vite · TailwindCSS-based self-service portal

MCP Control Plane

MCP Proxy · Cloud Suite MCP · Serverless MCP · Workloads MCP

Service Layer

Workloads · Finetune · Serverless · Datasets · Resources · Templates

Workload Management

Serverless workloads · Pods · Fine-tuning jobs· Dataset pipelines

Orchestration

Kubernetes · Kueue/Kai/Slurm · Tensorizer · SLO based auto-scaling

Virtualization

Virtual xPU clusters · Resource partitioning and allocation

Physical Hardware

GPU clusters(A100/H100) · NPU · InfiniBand/RoCE · NFS/PostgreSQL