

PRODUCT GUIDE 2026

Product Guide

One Token Factory. Four ways to use it.

One Token Factory. Four ways to use it.

01

SIKUMO Cloud

GPU Cloud

02

SIKUMO Models

Open Model API

03

SIKUMO Factory

Private Token Factory

04

SIKUMO Hub

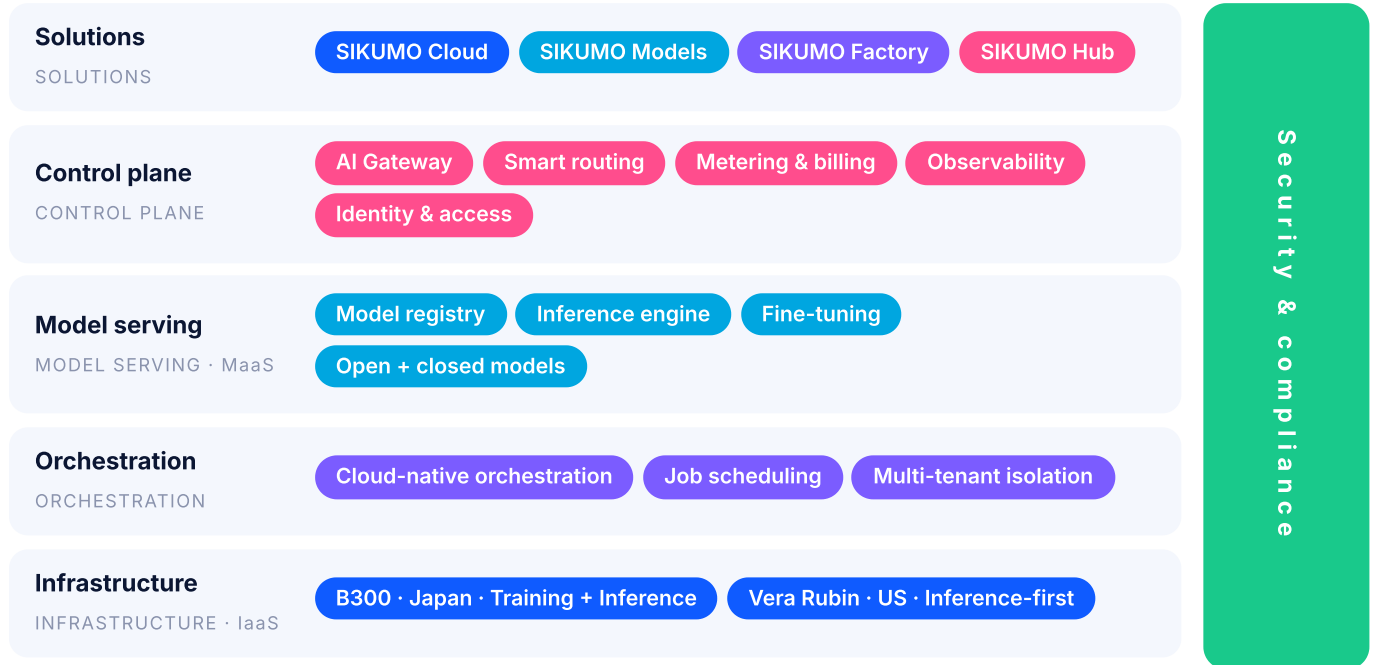
Model API Hub



OVERVIEW

Overview: One Token Factory

Every SIKUMO product runs on the same Token Factory, built as a standard layered architecture: NVIDIA next-generation GPUs (IaaS) deployed in compliant regions in Japan and the United States at the base; cloud-native orchestration, model serving (MaaS) and the AI Gateway control plane above; and four customer-facing delivery models on top.



How to choose

I need GPUs to train or run my own inference	→ 01 SIKUMO Cloud
I want to call cost-efficient models directly	→ 02 SIKUMO Models
My data cannot leave my boundary; I need a private AI stack	→ 03 SIKUMO Factory
I use many models and want one place to manage them	→ 04 SIKUMO Hub

Open models at the core, closed APIs as complement: same quality, lower token cost. Cost per token = compute cost ÷ (hardware efficiency × model efficiency × operating efficiency)

01 SIKUMO Cloud

GPU Cloud · NVIDIA next-generation compute, delivered as a service

Get the compute you need for training, inference, fine-tuning and AI applications without buying GPUs, building data halls or standing up a compute environment. Capacity comes from SIKUMO's own Token Factories and partner compute, deployed in regions such as Japan and the United States where NVIDIA GPU servers can be procured compliantly.



Compute

NVIDIA B300

Japan · Training + Inference

Large-scale training, fine-tuning and high-throughput inference

NVIDIA Vera Rubin

US · Inference-first

Large-scale, low-latency inference and API services

Form factors

GPU instances

On-demand configuration for development, testing and short-term inference

Bare metal

Dedicated physical servers for production and model deployment

GPU clusters

High-speed interconnect for distributed training and large-scale inference

Dedicated clusters

Reserved resource pools for large customers with long-term stability

Usage & billing

On-demand

ON-DEMAND

Elastic, short-term workloads

Reserved

RESERVED

Monthly / quarterly commitment, better cost

Annual

ANNUAL

Long-term stable training and production

Dedicated

DEDICATED

Whole-cluster dedication, custom delivery

Who it is for

AI startups & model teams

Enterprises that train or fine-tune

Inference services & AI platforms

Research institutes & universities

For pricing and onboarding, contact the SIKUMO customer team. Available regions, SKUs and lead times are confirmed case by case.

02 SIKUMO Models

Open Model API · The world's best open models, one API, per-token billing

No need to deploy or operate models yourself. SIKUMO runs the world's leading open models on its own and partner compute, optimises them for inference and delivers them through standard APIs, billed on actual token usage and continuously tracking the latest open-source releases.

Model catalogue (continuously updated)

DeepSeek family

Reasoning · code · general chat
Frontier reasoning at an efficient price

Qwen · GLM · Llama · Kimi · MiniMax and more

Multilingual · long context · agents
Covers Japanese, English, Chinese and more

Form factors

Standard API

Compatible with mainstream API specs; migrate existing apps smoothly

Streaming & batch

Real-time online calls and offline batch processing

Tool calling

Function calling and structured output (model-dependent)

Dedicated endpoints

Guaranteed throughput and latency for high-concurrency workloads

Usage & billing

Pay-as-you-go

PAY-AS-YOU-GO
Billed per input / output token

Committed

COMMITTED
Commit usage for a lower unit price

Dedicated

DEDICATED
Billed per GPU-hour

Enterprise

ENTERPRISE
Custom pricing and service levels

Who it is for

AI app & agent developers

SaaS & internet companies

Multilingual content & customer service

Teams replacing expensive closed APIs

For pricing and onboarding, contact the SIKUMO customer team. Model versions, context and capability scope are stated on each model card.

03 SIKUMO Factory

Private Token Factory · The whole Token Factory, delivered into your own environment

For data-sensitive industries such as finance, healthcare, public sector and manufacturing, SIKUMO packages compute, open models, the inference platform, fine-tuning and operations, delivered as a dedicated cluster, on-premises or hybrid. Data boundaries defined by contract, customisable models, predictable cost.

What is delivered

Compute + platform

Dedicated GPU cluster and inference operations platform
Built on the Compute Platform and Inference Platform

Models + fine-tuning

Private deployment of open models with industry fine-tuning
Capabilities built on your knowledge, terminology and processes

Deployment options

Dedicated hosting

Customer-dedicated cluster in a SIKUMO facility

On-premises

Deployed in your own data centre, fully under your control

Hybrid

Sensitive data processed locally, elastic load in the cloud

Continuous operations

Monitoring, upgrades, model iteration and O&M support

Delivery process

Assess

ASSESS

Use case, data and compliance assessment

Design

DESIGN

Compute and model solution design

Deploy

DEPLOY

Cluster, models and platform go live

Operate

OPERATE

Fine-tuning iterations and ongoing operations

Who it is for

Financial institutions

Healthcare & life sciences

Government & public sector

Large manufacturers

For pricing and onboarding, contact the SIKUMO customer team. Data boundaries, access rights and audit scope are defined in the contract.

04

SIKUMO Hub

Model API Hub · One endpoint for every open and closed model we are authorised to serve

No need to integrate each model vendor separately. SIKUMO Hub connects our open models and authorised closed-model APIs behind one gateway, one authentication and one bill, routing intelligently by cost, quality, latency and compliance.

Platform capabilities

Unified access

Open + closed, one interface

New models available instantly, no code changes

Smart routing

Cost · quality · latency · compliance

Policy-based model selection with automatic failover

Governance

One bill

All model usage consolidated into a single invoice

Budget control

Quotas and alerts per team, project and application

Usage monitoring

Real-time calls, latency, error rate and cost

Audit logs

Meets enterprise compliance and internal-control needs

How it works

Connect

CONNECT

One API key for every model

Route

ROUTE

Configure routing and failover policies

Govern

GOVERN

Budgets, permissions and audit

Optimise

OPTIMIZE

Continuously optimise cost and quality

Who it is for

Multi-model application companies

Enterprise IT & AI platform teams

AI agent developers

Organisations managing AI spend centrally

For pricing and onboarding, contact the SIKUMO customer team. Closed models are connected only where licence, region and use conditions allow.

INFRASTRUCTURE & TECHNOLOGY

Infrastructure & technology foundation

SIKUMO builds Token Factories only in regions where NVIDIA GPU servers can be procured compliantly, operated on the founding team's decade-plus of distributed-computing experience.

JAPAN · TOKYO

SIKUMO Tokyo

NVIDIA B300

Training + Inference

Large-model training, fine-tuning and high-throughput inference for Japan and Asia-Pacific customers

UNITED STATES

SIKUMO US

NVIDIA Vera Rubin

Inference-first

Large-scale, low-latency inference and API services for North American and global customers

Four technology platforms (standard architecture)

Compute Platform

GPU IaaS, cloud-native orchestration and multi-tenancy

Inference Platform

MaaS: inference, fine-tuning and model registry

AI Gateway

Unified API, routing, billing and observability

Private AI Stack

Dedicated and on-prem deployment, security and compliance

Open standards, global R&D collaboration

Japan

Japan R&D

United States

US R&D

Open standards

Open Standards

We follow international architectures and open standards and stay technology-neutral; R&D results flow continuously into every Token Factory for customers worldwide.

Factory SKUs and project stages are labelled planned / building / testing / commercial, based on procurement, delivery and acceptance records.

SERVICE & SUPPORT

Service & support

Standard

STANDARD

Online docs and console
Ticket support
Usage and billing lookup

Enterprise

ENTERPRISE

Dedicated account manager
Priority response and technical support
Custom pricing and contract terms

Dedicated

DEDICATED

Dedicated engineering team
End-to-end support for deployment, fine-tuning and operations
Dedicated service-level agreement

Onboarding

1 Discuss

Use case, scale and compliance needs

2 Propose

Product mix and pricing

3 Validate

Trial quota or PoC

4 Go live

Contract, onboarding and ongoing service

Our commitments

Integrity

Transparent pricing, promises we keep, honesty with customers and data

Persistence

Long-term investment in infrastructure — a partner you can rely on for a decade

Creation

Keep lowering token cost; make AI simpler and more effective

Superintelligence, at your service.

AGI at your service.

SIKUMO株式会社・SIKUMO K.K.

Tokyo, Japan · Web sikumo.ai · Email _____ · Tel _____

