

COMPANY INTRODUCTION 2026

Company Introduction

**Superintelligence,
at your service.**

AGI at your service.

SIKUMO K.K. · Token Factory investor and operator

AI Cloud Services | Tokyo · United States | Global Token Factory Operator



SIKUMO株式会社

SIKUMO K.K. · Tokyo · United States

WHAT WE BELIEVE

Three beliefs about the AI era

Our entire strategy rests on three long-term beliefs.

01 Tokens are the new electricity

TOKENS ARE THE NEW ELECTRICITY

Every inference, every line of code and every agent action is ultimately metered in tokens. Whoever produces tokens cheaper and more reliably holds the infrastructure of the AI economy.

02 Open models have caught up

OPEN MODELS HAVE CAUGHT UP

DeepSeek, Qwen, GLM, Llama and other open models keep closing in on the closed frontier. Models are commoditising; value is shifting from "who owns the model" to "who produces tokens efficiently".

03 Compliant compute is scarce

COMPLIANT COMPUTE IS SCARCE

High-end GPUs are constrained by export controls and supply chains. Compute that can be deployed at scale and operated continuously in compliant regions is the scarcest resource in overseas AI markets.

So we do one thing

Build and operate the most cost-efficient Token Factories in compliant regions.

Two identities

To customers: AI cloud services

AI CLOUD SERVICES

Turning advanced compute in compliant regions, open models and continuous operations into intelligence companies can use directly.

As a business: Token Factory investor and operator

TOKEN FACTORY INVESTMENT & OPERATIONS

Building and operating the most cost-efficient Token Factories in regions where NVIDIA GPU servers can be procured compliantly.

WHO WE ARE

SIKUMO: operating Token Factories for customers worldwide

SIKUMO K.K. is headquartered in Tokyo and invests in and operates Token Factories, with Japan and the United States as its core locations.

SIKUMO was founded by a group of engineers devoted to distributed computing and AI. The founders and core team have spent more than a decade in distributed computing, through the full cycle from cloud to large language models. We bring compute operations, model deployment and API operations proven in production to markets where the world's most advanced GPUs can be obtained compliantly, delivering cost-efficient AI services to customers worldwide.

10+ years

of distributed computing and AI

2

core markets: Japan · US

4

product lines, all around the Token Factory

2

factory types: training + inference · inference-first

The name: SIKUMO = SI × KUMO

SI · Superintelligence

Superintelligence

The next stage of AI. What we manufacture is not compute itself, but the tokens on the road to superintelligence.

Si · Silicon

Silicon

GPUs, chips and silicon compute — our "production line", and the origin of the hexagon mark.

Shikumi · 仕組み

Mechanism

SIKUMO sounds like the Japanese "shikumi" (mechanism): making superintelligence a mechanism every company can use.

KUMO is the Japanese word for cloud: our delivery model and our identity as a company rooted in Japan. The formal name is always written SIKUMO.

Mission and values

Make AGI create the greatest value for humanity — and make work simple and effective.

Make AGI create the greatest value for humanity — and make work simple and effective.

Integrity

Honest with customers, partners and data. Transparent pricing, promises we keep, no grey business.

Persistence

Infrastructure is a long game. We are willing to spend a decade doing what is hard and right.

Creation

Create real value with technology: lower token cost, simpler ways to use AI.

THE TOKEN FACTORY

The Token Factory: turning compute into metered intelligence

The Token Factory has four stages: compliant compute → open models → inference optimisation → token delivery.

01 Compliant compute

COMPLIANT COMPUTE

B300 training + inference in Japan, Vera Rubin inference-first in the US, all in compliant regions

02 Open models

OPEN MODELS

The world's best open models at the core, closed APIs as complement

03 Inference optimisation

INFERENCE ENGINE

Scheduling, quantisation and inference optimisation to maximise output per GPU

04 Token delivery

TOKEN DELIVERY

SIKUMO Cloud · Models · Factory · Hub

Token cost is decided by three efficiencies

Cost per token = compute cost ÷ (hardware efficiency × model efficiency × operating efficiency)

Next-generation GPUs raise hardware efficiency; open models and inference optimisation raise model efficiency; a decade of distributed operations raises operating efficiency. Multiplied together, that is SIKUMO's price advantage.

A mature technology stack: four platforms, continuously iterated

Compute Platform

COMPUTE PLATFORM · GPU IaaS

Cloud-native orchestration and multi-tenancy, GPU as a service

Inference Platform

INFERENCE PLATFORM · MaaS

Inference, fine-tuning and model registry

AI Gateway

AI GATEWAY · API ROUTER

Unified API, routing, billing and observability

Private AI Stack

PRIVATE AI STACK

Dedicated and on-prem deployment, security and compliance

The founding team's decade of distributed computing has matured into a stack proven in production. We follow international architectures and open standards, stay technology-neutral, and iterate global R&D into every Token Factory.

Model strategy: open at the core, closed as complement

Core: the world's best open models

DeepSeek

Qwen

GLM

Llama

Kimi

MiniMax

We track the latest open-source releases and deploy, optimise and fine-tune them ourselves: controllable cost, customisable capability, no vendor lock-in.

Complement: leading closed-model APIs

Through SIKUMO Hub we connect authorised closed models, combine them with open models where frontier capability is needed, and let routing policy pick the best model automatically.

Result: same quality at lower token cost; more choice, less lock-in.

PRODUCTS

Four product lines: four ways to deliver the Token Factory

One Token Factory, four delivery models. Product names follow "SIKUMO + one English noun".

01 · Delivering the "production line"

SIKUMO Cloud

GPU Cloud

NVIDIA next-generation compute as a service — on-demand, reserved, annual or dedicated clusters.

Compute Japan B300 training + inference · US Vera Rubin inference-first

Forms GPU instances · bare metal · clusters · dedicated clusters

For AI companies · model teams · developers

02 · Delivering the "product"

SIKUMO Models

Open Model API

The world's best open models on our own and partner compute, as per-token APIs.

Models DeepSeek · Qwen · GLM · Llama · Kimi · MiniMax and more

API Mainstream-compatible · streaming & batch · tool calling

For AI apps · SaaS · internet companies

03 · Delivering the whole factory

SIKUMO Factory

Private Token Factory

Private deployment of compute and models for data-sensitive industries, with data boundaries defined by contract.

Forms Dedicated hosting · on-premises · hybrid

Scope Compute + models + inference platform + fine-tuning + operations

For Finance · healthcare · public sector · manufacturing

04 · One endpoint for every authorised model

SIKUMO Hub

Model API Hub

Unified access, billing, smart routing and governance for open and closed model APIs.

Coverage Our open models + authorised closed-model APIs

Govern One bill · budgets · usage monitoring · audit

For Multi-model apps · platform teams · agents

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](#)

See the SIKUMO Product Guide for detailed specifications, billing models and target customers.

GLOBAL FOOTPRINT

Global footprint: two factory types, each with its role

We build Token Factories only where NVIDIA GPU servers can be procured compliantly, choosing the underlying compute to fit each region's role.

JAPAN · TOKYO

SIKUMO Tokyo

NVIDIA B300 · Training + Inference

Built around B300 for large-model training, fine-tuning and high-throughput inference, serving Japan and Asia-Pacific. Data and models stay in Japan.

UNITED STATES

SIKUMO US

NVIDIA Vera Rubin · Inference-first

Built around Vera Rubin for large-scale, low-latency inference and API services, serving North America and the world.

MORE REGIONS

More compliant regions

NVIDIA next-generation GPUs · Expanding on demand

Continuously expanding the Token Factory network in markets where NVIDIA GPU servers can be procured compliantly.

Open standards, global R&D collaboration

Japan

Japan R&D

United States

US R&D

Open standards

Open Standards

Factory SKUs and project stages are labelled planned / building / testing / commercial, based on procurement, delivery and acceptance records. This page does not commit to GPU counts or go-live dates.

WHY SIKUMO

Why SIKUMO

1 A decade of technology

Long experience in distributed computing and AI; running compute is what we have always done.

2 Mature platforms

Compute, inference, gateway and private stack — ready to use and continuously iterated.

3 Advanced compute in compliant regions

B300 training + inference in Japan, Vera Rubin inference-first in the US.

4 Open-first cost structure

Open models at the core make token cost structurally lower.

5 From an API to a whole factory

One stack covers every delivery model, from public API to private factory.

6 Honest business values

Transparent pricing and promises we keep — a partner customers can trust for the long run.

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

Superintelligence, at your service.

AGI at your service | Integrity · Persistence · Creation | Manufacturing Intelligence, Token by Token.

SIKUMO株式会社 · SIKUMO K.K.

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

