

PRODUCT GUIDE 2026

产品手册

一座 Token 工厂，四种交付方式

One Token Factory. Four ways to use it.

01

SIKUMO Cloud

GPU算力云 · GPU Cloud

02

SIKUMO Models

开源模型API · Open Model API

03

SIKUMO Factory

行业Token工厂 · Private Token Factory

04

SIKUMO Hub

模型API聚合 · Model API Hub



OVERVIEW

产品全景：一座 Token 工厂

SIKUMO 的所有产品运行在同一座Token工厂上，并按国际通行的分层架构构建：底层是部署在日本、美国合规区域的 NVIDIA新一代GPU（IaaS），之上是云原生编排、模型服务（MaaS）与AI网关控制平面，最上层是面向客户的四种交付方式。



如何选择

我要训练或自建推理，需要GPU	→ 01 SIKUMO Cloud
我要直接调用高性价比的模型	→ 02 SIKUMO Models
我的数据不能出域，需要专属AI能力	→ 03 SIKUMO Factory
我要同时用多家模型，统一管理	→ 04 SIKUMO Hub

开源为核心，闭源为补充：同等质量，更低的 Token 成本。 单位Token成本 = 算力成本 ÷ (硬件效率 × 模型效率 × 运营效率)

01 SIKUMO Cloud

GPU算力云 · 以 GPU 即服务的方式，交付 NVIDIA 新一代算力

客户无需采购GPU、建设机房或搭建计算环境，即可获得训练、推理、微调与AI应用运行所需的算力。资源来自SIKUMO自有的Token工厂与合作算力，部署在日本、美国等可合规采购NVIDIA GPU服务器的区域。

算力规格

NVIDIA B300

日本 · 训推一体

大规模模型训练、微调与高吞吐推理

NVIDIA Vera Rubin

美国 · 推理为核心

大规模、低延迟的模型推理与API服务

产品形态

GPU实例

按需配置，适合开发、测试与短期推理

裸金属服务器

整机独享，适合生产环境与模型部署

GPU集群

多机高速互联，适合分布式训练与大规模推理

专属集群

为大型客户划分专属资源池，长期稳定保障

使用与计费模式

按需

ON-DEMAND

弹性、阶段性任务

预留

RESERVED

月度／季度承诺，成本更优

包年

ANNUAL

长期稳定的训练与生产

专属

DEDICATED

整集群专属，定制交付

适用客户

AI初创公司与模型团队

需要训练或微调的企业

推理服务与AI应用平台

科研与高校机构

了解价格与开通方式，请联系 SIKUMO 客户团队。可用地区、型号与交付周期以实际核实为准。

02 SIKUMO Models

开源模型API · 全球优秀开源模型，一个 API，按 Token 计费

客户不必自行部署与运维模型。SIKUMO 在自有及合作算力上部署全球优质开源模型，经推理优化后以标准API交付，按实际Token用量结算，持续跟进开源社区最新版本。

模型目录（持续更新）

DeepSeek 系列

推理 · 代码 · 通用对话
高性价比的前沿推理模型

Qwen · GLM · Llama · Kimi · MiniMax 等

多语言 · 长上下文 · Agent
覆盖中、英、日等多语言场景

产品形态

标准API

兼容主流API规范，现有应用可平滑迁移

流式与批量

在线实时调用与离线批处理两种模式

工具调用

支持Function Calling与结构化输出（视模型）

专属推理端点

独享吞吐与延迟保障，适合高并发业务

使用与计费模式

按量

PAY-AS-YOU-GO
按输入／输出Token计费

套餐

COMMITTED
承诺用量，单价更低

专属端点

DEDICATED
按GPU时长计费

企业

ENTERPRISE
定制价格与服务等级

适用客户

AI应用与Agent开发者

SaaS与互联网企业

多语言内容与客服业务

需要替换高价闭源API的团队

了解价格与开通方式，请联系 SIKUMO 客户团队。模型版本、上下文与能力范围以模型卡为准。

03

SIKUMO Factory

行业Token工厂 · 把整座 Token 工厂，交付给客户自己的环境

面向金融、医疗、政企、制造等数据敏感行业，SIKUMO 将算力、开源模型、推理平台、微调与运维打包，以专属集群、本地部署或混合部署的方式交付。按合同定义数据边界，模型可定制，成本可预期。

交付内容

算力 + 平台

专属GPU集群与推理运营平台

基于算力基础设施平台与模型推理服务平台

模型 + 微调

开源模型私有部署与行业微调

结合企业知识、术语与流程的专属能力

部署形态

专属托管

部署在 SIKUMO 机房的客户专属集群

本地部署

部署在客户自有机房，完全自主可控

混合部署

敏感数据本地处理，弹性负载上云

持续运营

监控、升级、模型迭代与运维支持

交付流程

评估

ASSESS

场景、数据与合规评估

设计

DESIGN

算力与模型方案设计

部署

DEPLOY

集群、模型与平台上线

运营

OPERATE

微调迭代与持续运维

适用客户

金融机构

医疗与生命科学

政企与公共部门

大型制造企业

了解价格与开通方式，请联系 SIKUMO 客户团队。数据边界、权限与审计范围在合同中明确。

04 SIKUMO Hub

模型API聚合 · 一个接口，统一接入获准提供的开源与闭源模型

企业不必逐一对接多家模型供应商。SIKUMO Hub 统一接入自有开源模型与获授权的主流闭源模型API，提供统一网关、统一鉴权、统一账单，并按成本、质量、延迟与合规要求智能路由。

平台能力

统一接入

开源 + 闭源，一套接口

新模型上线即可调用，无需改代码

智能路由

成本 · 质量 · 延迟 · 合规

按策略自动选择最优模型，支持故障切换

治理能力

统一账单

所有模型用量合并计费，一张账单

预算控制

按团队、项目、应用设置额度与告警

用量监控

实时查看调用量、延迟、错误率与成本

审计日志

满足企业合规与内控要求

使用路径

接入

CONNECT

一个API Key接入全部模型

路由

ROUTE

配置路由与容灾策略

治理

GOVERN

预算、权限与审计

优化

OPTIMIZE

成本与质量持续优化

适用客户

多模型应用企业

企业IT与AI平台团队

AI Agent开发者

需要统一管理AI支出的组织

了解价格与开通方式，请联系 SIKUMO 客户团队。闭源模型仅在授权、地区与用途条件允许时接入。

INFRASTRUCTURE & TECHNOLOGY

算力基础设施与技术底座

SIKUMO 只在可合规采购NVIDIA GPU服务器的区域建设Token工厂，并以创始团队十余年分布式计算运营经验作为运营底座。

JAPAN · TOKYO

SIKUMO Tokyo

NVIDIA B300

训推一体 · Training + Inference
大模型训练、微调与高吞吐推理，服务日本及亚太客户

UNITED STATES

SIKUMO US

NVIDIA Vera Rubin

推理为核心 · Inference-first
大规模、低延迟推理与API服务，服务北美及全球客户

四大技术平台（国际通行架构）

算力基础设施平台

Compute Platform · GPU IaaS，云原生编排与多租户

模型推理服务平台

Inference Platform · MaaS，推理、微调与模型仓库

AI网关

AI Gateway · 统一API、路由、计费与可观测

私有化AI栈

Private AI Stack · 专属与本地部署、安全合规

开放标准，全球研发协作

日本

Japan R&D

美国

US R&D

开放标准

Open Standards

遵循国际通行架构与开放标准，保持全球技术中立；研发成果持续迭代到每一座Token工厂，服务全球客户。

工厂型号及项目阶段分别标注"规划／建设／测试／商用"，以采购、交付与验收记录为准。

SERVICE & SUPPORT

服务与支持

标准服务

STANDARD

在线文档与控制台
工单支持
用量与账单查询

企业服务

ENTERPRISE

专属客户经理
优先响应与技术支持
定制价格与合同条款

专属服务

DEDICATED

专属技术团队
部署、微调与运维全程支持
专属服务等级约定

开通流程

1 沟通需求

场景、规模与合规要求

2 方案与报价

产品组合与价格方案

3 测试验证

开通测试额度或POC

4 正式上线

签约、开通与持续服务

我们的承诺

正直

价格透明，承诺可兑现，对客户与数据诚实

坚持

长期投入基础设施，做客户可以依赖十年的伙伴

创造

持续降低Token成本，让AI的使用更简单、有效

Superintelligence, at your service.

让AGI为您服务 | スーパーインテリジェンスを、あなたのために。

SIKUMO株式会社・SIKUMO K.K.

日本东京・Web sikumo.ai・Email _____・Tel _____