

COMPANY INTRODUCTION 2026

会社紹介

**Superintelligence,
at your service.**

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・トークンファクトリーの投資・運営会社

AIクラウドサービス | Tokyo・United States | Global Token Factory Operator



WHAT WE BELIEVE

AI時代の3つの判断

私たちのAI戦略は、3つの長期的な判断の上に成り立っています。

01 トークンはAI時代の新しい電力

TOKENS ARE THE NEW ELECTRICITY

すべての推論、コード、エージェントの動作は最終的にトークンで計量されます。より低コストで安定的にトークンを生産できる者が、AI経済のインフラを握ります。

02 オープンモデルは第一線に到達

OPEN MODELS HAVE CAUGHT UP

DeepSeek、Qwen、GLM、Llamaなどのオープンモデルはクローズドの最先端に迫り続けています。モデルはコモディティ化し、価値は「誰がモデルを持つか」から「誰が効率よくトークンを生産するか」へ移っています。

03 合规なコンピュータは希少な生産能力

COMPLIANT COMPUTE IS SCARCE

ハイエンドGPUは輸出規制とサプライチェーンの制約を受けます。合规リージョンで大規模に配備し継続運用できるコンピュータは、海外AI市場で最も希少な資源です。

だから、私たちはひとつのことだけをします

合规リージョンで、最もコスト効率の高いトークンファクトリーを建設・運営する。

2つの顔

お客様に対して：AIクラウドサービス

AI CLOUD SERVICES

合规リージョンの先進コンピュータ、オープンモデル、継続運用を、企業がそのまま使える知能の能力へ。

経営として：トークンファクトリーの投資・運営

TOKEN FACTORY INVESTMENT & OPERATIONS

NVIDIA GPUサーバーを合规に調達できるリージョンで、最もコスト効率の高いトークンファクトリーを建設・運営します。

WHO WE ARE

SIKUMO：世界のお客様のためにトークンファクトリーを運営

SIKUMO株式会社は東京に本社を置き、日本と米国を中核拠点として、トークンファクトリーへの投資と運営を行っています。

SIKUMOは、分散コンピューティングとAI技術の発展に情熱を注ぐエンジニアによって設立されました。創業者とコアチームは分散コンピューティング分野で10年以上の経験を持ち、クラウドから大規模言語モデルまでの技術サイクルを一貫して経験しています。本番環境で長年検証してきたコンピュータ運用、モデル配備、API運営の能力を、世界最先端のGPUを合規に入手できる市場に持ち込み、世界のお客様にコスト効率の高いAIサービスを提供します。

10年以上

分散コンピューティングとAI技術の蓄積

2

中核市場：日本・米国

4

つの製品ライン、すべてトークンファクトリー中心

2

種類のファクトリー：学習＋推論・推論特化

社名：SIKUMO = SI × KUMO

SI・Superintelligence

スーパーインテリジェンス

AIの次の段階。私たちが製造するのはコンピュータそのものではなく、スーパーインテリジェンスへ至るトークンです。

Si・Silicon

シリコン

GPU、チップ、シリコンベースのコンピュータ——私たちの「製造ライン」であり、六角形マークの由来です。

仕組み・しくみ

メカニズム

SIKUMOの響きは日本語の「仕組み」に近い：スーパーインテリジェンスを、すべての企業が使える仕組みにする。

KUMOは日本語の「雲」：クラウドという提供形態であり、日本に根ざす会社のアイデンティティです。正式名称は常にSIKUMOと表記します。

ミッションと価値観

AGIが人類社会に最大の価値を生み出し、仕事をシンプルで効果的にする。

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

誠実・Integrity

お客様、パートナー、データに誠実に。透明な価格、守れる約束、グレーなビジネスはしない。

継続・Persistence

インフラは長期の事業。難しくても正しいことに10年をかけます。

創造・Creation

技術で本当の価値を創る：より低いトークンコスト、よりシンプルなAIの使い方。

THE TOKEN FACTORY

トークンファクトリー：コンピュータを計量可能な智能へ

トークンファクトリーは4つの工程で構成されます：合规コンピュータ → オープンモデル → 推論最適化 → トークン提供。

01 合规コンピュータ

COMPLIANT COMPUTE

日本はB300で学習+推論、米国はVera Rubinで推論特化。いずれも合规リージョンに配備

02 オープンモデル

OPEN MODELS

世界の優れたオープンモデルを中核に、クローズドAPIを補完に

03 推論最適化

INFERENCE ENGINE

スケジューリング、量子化、推論最適化でGPU 1枚あたりの産出を最大化

04 トークン提供

TOKEN DELIVERY

SIKUMO Cloud・Models・Factory・Hub

トークンコストは3つの効率で決まる

トークン単価 = コンピュータコスト ÷ (ハードウェア効率 × モデル効率 × 運用効率)

最新GPUがハードウェア効率を、オープンモデルと推論最適化がモデル効率を、10年以上の分散運用体制が運用効率を高めます。この3つの積がSIKUMOの価格優位です。

成熟した技術体系：4つのプラットフォーム、継続的な進化**コンピュータ基盤**

COMPUTE PLATFORM・GPU IaaS

クラウドネイティブ編成とマルチテナント、GPUaaS

推論基盤

INFERENCE PLATFORM・MaaS

推論、ファインチューニング、モデルレジストリ

AIゲートウェイ

AI GATEWAY・API ROUTER

統一API、ルーティング、課金、可観測性

プライベートAIスタック

PRIVATE AI STACK

専用/オンプレ配備、セキュリティとコンプライアンス

創業チームの10年以上の分散コンピューティング経験は、本番環境で長年検証された技術体系として結実しています。国際標準のアーキテクチャとオープン標準に従い、技術的中立性を保ち、グローバルな研究開発の成果をすべてのトークンファクトリーに継続的に反映します。

モデル戦略：オープンの中核に、クローズドを補完に**中核：世界の優れたオープンモデル**

DeepSeek

Qwen

GLM

Llama

Kimi

MiniMax

オープンソースコミュニティの最新版に追従し、自社配備・推論最適化・業界特化調整を実施。コストは管理可能、能力はカスタマイズ可能、ベンダーロックインなし。

補完：優れたクローズドモデルAPI

SIKUMO Hubを通じて認可済みの主要クローズドモデルに接続し、最先端能力が必要な場面でオープンモデルと組み合わせ、ルーティングポリシーが最適モデルを自動選択します。

結果：同等品質でより低いトークンコスト、より多い選択肢とより少ないロックイン。

PRODUCTS

4つの製品ライン：トークンファクトリーの4つの提供形態

ひとつのトークンファクトリー、4つの提供形態。製品名は「SIKUMO + 英単語ひとつ」。

01・「製造ライン」を提供

SIKUMO Cloud

GPUクラウド・GPU Cloud

NVIDIA最新コンピュートをGPUaaSとして提供。オンデマンド、リザーブド、年間契約、専用クラスター。

コンピュータ 日本 B300 学習＋推論・米国 Vera Rubin 推論特化

形態 GPUインスタンス・ベアメタル・クラスター・専用クラスター

対象 AI企業・モデルチーム・開発者

02・「製品」を提供

SIKUMO Models

オープンモデルAPI・Open Model API

自社・パートナーの算力上に世界の優れたオープンモデルを配備し、トークン課金のAPIとして提供。

モデル DeepSeek・Qwen・GLM・Llama・Kimi・MiniMax など

API 主要API仕様互換・ストリーミングとバッチ・ツール呼び出し

対象 AIアプリ・SaaS・インターネット企業

03・ファクトリーをまるごとお客様へ

SIKUMO Factory

プライベート・トークンファクトリー

データ機微性の高い業界向けに、コンピュータとモデルをプライベート配備。データ境界は契約で定義。

形態 専用ホスティング・オンプレミス・ハイブリッド

内容 コンピュータ＋モデル＋推論基盤＋ファインチューニング＋運用

対象 金融・医療・公共・製造

04・ひとつの接続で認可済みモデルを統合

SIKUMO Hub

モデルAPIハブ・Model API Hub

オープン／クローズドモデルAPIの統一接続、統一課金、スマートルーティングとガバナンス。

対応 自社オープンモデル＋認可済み主要クローズドモデルAPI

ガバナンス 統一請求・予算・使用量モニタリング・監査

対象 マルチモデルアプリ・プラットフォームチーム・Agent

製品の選び方

学習や自社推論のためにGPUが必要

→ [01 SIKUMO Cloud](#)

コスト効率の高いモデルを直接呼び出したい

→ [02 SIKUMO Models](#)

データを外に出せない、専用のAI基盤が必要

→ [03 SIKUMO Factory](#)

複数ベンダーのモデルを一元管理したい

→ [04 SIKUMO Hub](#)

詳細な仕様、課金モデル、対象顧客は『SIKUMO 製品ガイド』をご参照ください。

GLOBAL FOOTPRINT

グローバル展開：2種類のファクトリー、それぞれの役割

NVIDIA GPUサーバーを合规に調達できるリージョンにのみトークンファクトリーを建設し、リージョンの役割に応じて最適な基盤コンピュートを選びます。

JAPAN・TOKYO

SIKUMO Tokyo

NVIDIA B300・学習＋推論・Training + Inference

B300を中核に、大規模モデルの学習、ファインチューニング、高スループット推論を担い、日本とアジア太平洋のお客様に提供。データとモデルは日本国内に保持。

UNITED STATES

SIKUMO US

NVIDIA Vera Rubin・推論特化・Inference-first

Vera Rubinを中核に、大規模・低遅延のモデル推論とAPIサービスを北米とグローバルのお客様に提供。

MORE REGIONS

その他の合规リージョン

NVIDIA 最新GPU・需要に応じて展開・Expanding

NVIDIA GPUサーバーを合规に調達できる市場で、トークンファクトリー網を拡大し続けます。

オープン標準、グローバルR&D協働

日本

Japan R&D

米国

US R&D

オープン標準

Open Standards

各ファクトリーの型番と段階は「計画／建設／検証／商用」で表記し、調達・納品・検収記録に基づきます。本ページはGPU数量や稼働時期を約束するものではありません。

WHY SIKUMO

なぜ SIKUMO なのか

1 10年以上の技術蓄積

分散コンピューティングとAI技術への長年の取り組み。
コンピュータ運用は私たちの本業です。

2 成熟した技術プラットフォーム

コンピュータ、推論、ゲートウェイ、プライベートの4プラットフォーム。すぐに使え、進化し続けます。

3 合規リージョンの先進コンピュート

日本はB300で学習+推論、米国はVera Rubinで推論特化。

4 オープン優先のコスト構造

オープンモデルを中核に、トークンコストは構造的に低い。

5 APIからファクトリー丸ごとまで

同じ技術で、パブリックAPIからプライベートファクトリーまですべての提供形態をカバー。

6 誠実なビジネス価値観

透明な価格と守れる約束。長く信頼されるパートナーに。

ご利用までの流れ

1 要件ヒアリング

ユースケース・規模・コンプライアンス要件

2 提案とお見積

製品構成と価格案

3 検証

テスト枠またはPoCの開通

4 本番稼働

契約・開通・継続サービス

Superintelligence, at your service.

スーパーインテリジェンスを、あなたのために。 | 誠実・継続・創造 | Manufacturing Intelligence, Token by Token.

SIKUMO株式会社・SIKUMO K.K.

東京・Web sikumo.ai・Email _____・Tel _____

