集合知による次世代AIアーキテクチャの確立：Sakana AI「Fugu」システムがもたらすマルチエージェント・オーケストレーションとAI主権の再定義

Gemini 3.1 pro

序論：モノリシック基盤モデルの限界と新たなパラダイムの台頭

2026年6月22日、日本のAIスタートアップであるSakana AIは、複数の最先端基盤モデル（フロンティアモデル）を内部で動的に連携させ、単一のAPIとして機能する自律型マルチエージェント・オーケストレーションシステム「Sakana Fugu（サカナ・フグ）」の一般提供を開始したと発表した[1]。この発表は、単に新しいAIモデルが市場に投入されたという事実にとどまらず、これまでAI業界のメインストリームであった「巨大なパラメータを持つ単一（モノリシック）のモデルをスケーリングさせる」というアプローチに対する根本的なパラダイムシフトを提示している。

近年、世界的なAI開発競争は、莫大な計算資源（CAPEX）とデータを投下して基盤モデルを学習させる「スケール則（Scaling Laws）」の追求に終始してきた[2]。ハイパースケーラーと呼ばれる巨大テクノロジー企業群は、パラメータ数を数千億から数兆規模へと拡張することで、モデルの汎用的な推論能力を引き上げることに成功してきた。しかしながら、現実世界におけるソフトウェア開発、高度なサイバーセキュリティ分析、あるいは未知の科学的発見といった「多段階でノイズの多い複雑な課題解決」においては、単一のモデルに対する「1ショット（一回きり）」のプロンプト入力による出力性能が明確な限界に達しつつあるという課題が顕在化していた[2]。

Sakana Fuguは、このアーキテクチャ上の限界を「集合知（Collective Intelligence）」すなわちスウォーム（群れ）インテリジェンスの概念を用いて突破しようとする野心的な試みである[2]。エンドユーザーや開発者から見れば、FuguはOpenAI互換の単一のAPIエンドポイントを提供する一つの言語モデルに過ぎない。しかし、その背後では、入力された難解なタスクをシステムが自律的に細かいサブタスクに分解し、世界最高峰の外部LLM（大規模言語モデル）プールから最適な専門エージェントを動的に選定・編成し、その結果を検証・統合するという極めて高度なプロセスが実行されている[1]。

さらに、本システムがエンタープライズ市場や国家機関から強い関心を集めている背景には、技術的なスコアの優位性だけでなく、近年のAIを取り巻く先鋭化した地政学的な緊張が存在する。後述するように、米国政府による特定のフロンティアモデルに対する前例のない輸出規制が現実のものとなる中、特定のプロバイダーAPIに依存する「ベンダーロックイン」は、企業の事業継続計画（BCP）における致命的な脆弱性（Concentration Risk）とみなされるようになった[6]。Sakana Fuguは、特定のベンダーに依存しない「スワップ可能（入れ替え可能）」なエージェントプールを構築し、障害や規制によるアクセス遮断を自動的に迂回する動的ルーティング機構を備えることで、組織の「AI主権（AI Sovereignty）」を保護するための堅牢なインフラストラクチャとしても機能するよう設計されている[2]。

本報告書では、Sakana Fuguを構成するアーキテクチャの根幹となる要素技術（TRINITYおよびConductor）、各種業界ベンチマークにおける定量的・定性的な性能評価、地政学的リスクへの対抗策としての構造的意義、エンタープライズ市場への導入における経済的合理性、そしてAI業界全体に与える長期的な影響について、網羅的かつ多角的な分析を展開する。

「Sakana Fugu」システムの全体像と提供モデルの仕様

Sakana Fuguの設計思想は、複雑なマルチエージェントシステムの運用負荷をユーザーから完全に抽象化することにある。従来、企業が複数のAIを連携させるためには、LangGraph、CrewAI、AutoGenといったオープンソースのフレームワークを用いて、開発者自身がハードコードされた静的なワークフローやエージェントの役割定義を設計し、保守する必要があった[3]。対照的に、Sakana Fugu自体が一つの高度な言語モデルとして機能し、自身を含むエージェントプール内の様々なLLMを再帰的に呼び出す方法を強化学習を通じて体得している[1]。ユーザーは複雑なシステムアーキテクチャを自身のコードベース内に構築する必要はなく、既存のAPIキーをFuguのエンドポイントに向けるだけで、背後にあるマルチエージェントの群れにシームレスにアクセスできる[2]。

ユーザーのワークロードの特性や要求される推論の深さに応じて、Fuguシステムは「Fugu（標準モデル）」と「Fugu Ultra（最上位モデル）」の2つのバリアントで提供されている[1]。

Fugu（標準モデル）：レイテンシと性能の最適均衡

標準のFuguモデルは、日常的な業務プロセスに最適化され、強固な推論パフォーマンスと低レイテンシ（応答速度）のバランスを追求したモデルである。このバリアントは、インタラクティブなチャットボットサービス、日々のコーディング支援、コードレビュー、さらには「Codex」のような開発ツールやIDE（統合開発環境）のバックエンドエンジンとして直接組み込まれることを想定して設計されている[1]。簡単なクエリに対しては、Fuguが背後の複雑なチームを立ち上げることなく直接回答を生成し、不要なレイテンシの増加を防ぐ自律的判断を下す[1]。

この標準モデルにおける最も重要なエンタープライズ向け機能の一つが、厳格なコンプライアンス管理のための「オプトアウト（除外）機能」である[1]。企業がAIを業務プロセスに導入する際、社内の厳格なデータガバナンスポリシーに準拠するため、あるいは特定の地域やプロバイダーへのデータ送信を制限しなければならないケースが多々ある。Fugu標準モデルでは、管理コンソール画面から特定のエージェント（特定の外部プロバイダーのLLM）をプールから明示的に除外することが可能であり、データのプライバシー要件や法的コンプライアンスを遵守しながらマルチエージェントシステムを安全に運用できるアーキテクチャを備えている[1]。

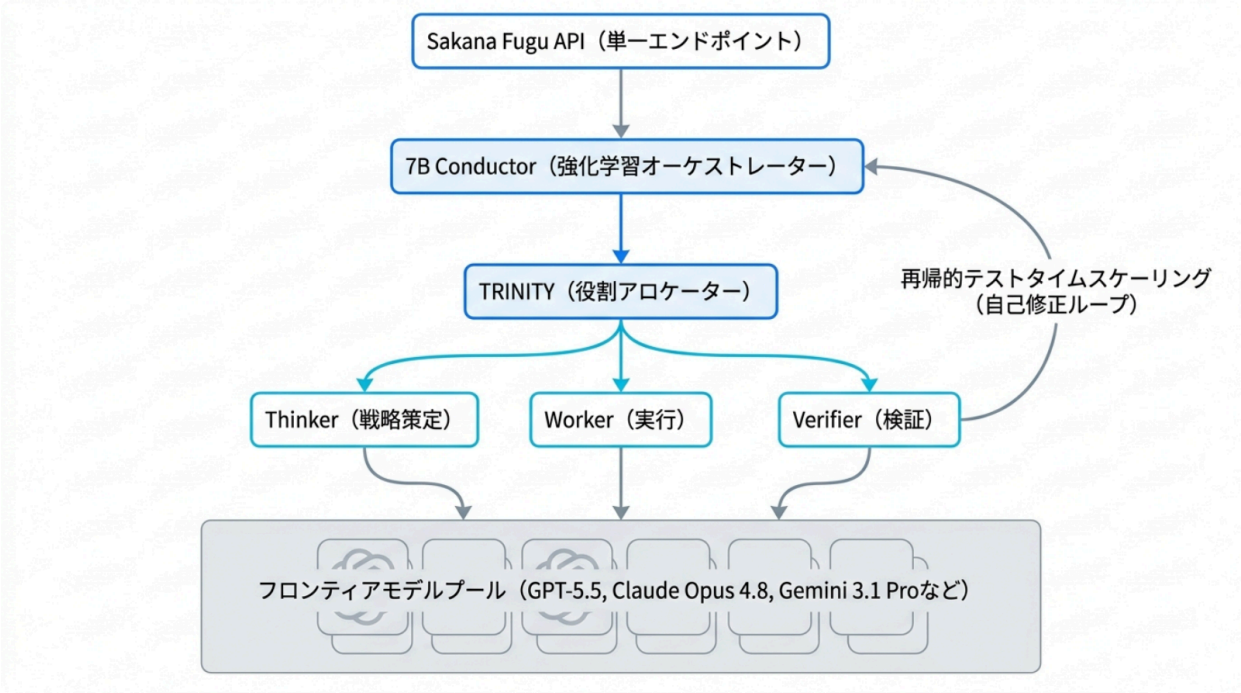
Fugu Ultra（最上位モデル）：極限の精度を追求する推論エンジン

フラッグシップモデルである「Fugu Ultra」（内部モデルID：fugu-ultra-20260615）は、ノイズが多く難解な多段階の推論タスクにおいて、回答の品質と精度を極限まで高めることに特化したシステムである[1]。AI研究の自動化、サイバーセキュリティの脆弱性分析とパッチ作成、膨大な文献・特許の調査、複雑なデータ分析など、極めて高い専門性と長期的なコンテキスト（記憶）の維持が求められる高負荷な領域で真価を発揮する[2]。

標準のFuguモデルとは異なり、Fugu Ultraは最高のパフォーマンスを担保するために、システムが利用可能なすべての中核的な専門エージェントのプールを強制的に活用するアーキテクチャとなっている。そのため、特定モデルの「オプトアウト機能」は提供されておらず、Sakana AIが厳選した固定のエージェントプールで動作する[1]。複雑な問題をブレイクダウンし、複数の専門モデル間で検証作

業を繰り返すため、標準モデルと比較して初回トークン生成までの応答時間（TTFT）や全体的なレイテンシは増加するものの、後述するベンチマークテストにおいては、現存する最強の個別基盤モデルを凌駕する目覚ましい性能を記録している[1]。

Sakana Fuguの自律的マルチエージェント協調アーキテクチャ



Fuguシステムは、単一のOpenAI互換APIの背後で、強化学習されたConductorとTRINITYコーディネーターが連携し、外部の最先端LLMに役割（思考、実行、検証）を与えて自律的に制御・修正を行う。

コアテクノロジー：TRINITYとConductorがもたらす自律的オーケストレーション

Sakana Fuguが、人間による手動のプロンプト設計や静的・ルールベースのルーティングアルゴリズムに頼ることなく、安定してフロンティアレベルの性能を発揮できる技術的根拠は、Sakana AIが2026年に開催された機械学習のトップカンファレンスであるICLR（International Conference on Learning Representations）で発表した2つの基盤的研究論文の成果に由来する。それが「TRINITY」システムと「Conductor」モデルである[1]。これらの研究は、AIモデルを単に巨大化させるのではなく、既存のAIモデル群をいかにして最適に「指揮（オーケストレート）」し、相乗効果を生み出すかというメタレベルの制御技術に焦点を当てている。

TRINITY：進化型スモール言語モデル（SLM）による役割分担

TRINITYシステムは、基礎となる各AIモデルの重み（ウェイト）パラメータに一切の変更を加えることなく、テスト時（推論実行時）に複数の最先端モデルの強みを動的に構成・融合するためのアプローチ

である[9]。Fuguの内部において、TRINITYはクエリを複数ターンにわたる対話的プロセスとして処理する。
このシステムの心臓部には、20K（2万）未満の学習可能パラメータしか持たない極めて軽量なコーディネーター（SLM: Small Language Model）が配置されている[10]。このコンパクトなコーディネーターは、各ステップにおいて会話の全トランスクリプト（履歴）を読み取り、利用可能なプール内から最適なLLMを選択する。そして選択されたLLMに対して、現在のタスクの進捗状況に応じて以下の3つの明確な役割（Role）のいずれかを適応的に割り当てる[2]。

1. **Thinker**（思考役）: 複雑な問題に対して高次元の解決戦略を立案し、現在のシステムの状態や前提条件を深く分析する役割を担う。コードを書く前の設計書作成に相当する。
2. **Worker**（実行役）: Thinkerが策定した戦略や計画に基づき、具体的なコードの記述、数学的推論の展開、あるいはデータの抽出といった実務的な問題解決ステップを直接実行する。
3. **Verifier**（検証役）: Workerが生成したソリューション（解決策）が、当初の要件を完全に満たしているか、論理的な矛盾がないか、コードにバグが含まれていないかを客観的かつ厳密に評価・検証する。

この役割分担のプロセスは一過性のものではない。Verifierが現在のソリューションに欠陥を見出した場合、そのフィードバックは再びThinkerやWorkerに差し戻され、Verifierが最終的な解決策として受け入れ可能だと判断するまで反復して継続される[10]。ICLR 2026での論文発表時点において、この進化型コーディネーターによるオーケストレーションは、GPT-5、Gemini 2.5-Pro、Claude 4-Sonnetといった構成モデル単体のパフォーマンスを平均して上回る成績を残しており、異なるモデルの長所を組み合わせることで生じる「能力の底上げ効果」を学術的に実証している[9]。

Conductor：強化学習が導き出した自然言語によるエージェント統率

TRINITYと並び、Fuguアーキテクチャのもう一つの強力な技術的支柱となるのが、強化学習（Reinforcement Learning: RL）を用いてエンドツーエンドで訓練された7B（70億パラメータ）のオーケストレーション特化型言語モデル「Conductor」である[11]。
Conductor自身は、直接的に数学の問題を解いたりソフトウェアコードを実行したりする能力を持たない。その代わり、GPT、Gemini、Claude、さらには多様なオープンソースモデルからなるフロンティアモデルのプールを統制する「敏腕マネージャー」あるいは「メタ・プロンプトエンジニア」として機能する[11]。従来、AI業界に存在したNot Diamond、Martian、RouteLLMなどの一般的なルーティングプラットフォームは、入力されたプロンプトのテキスト特徴を分析し、最も適した単一のモデルに対して処理を「1ショット（一回のみ）」で割り当てる静的なアプローチを採用していた[3]。
しかし、FuguのConductorは、複数ラウンド（マルチラウンド）にわたる動的パラダイムを採用している。入力された複雑な課題に対して、Conductorはプログラムコードではなく、自然言語による協調ワークフローを自ら出力してシステムを制御する。具体的には、任意のクエリに対してConductorが自律的に以下の3点を決定する[11]。

- エージェントの選定: 自身のプールの中から、どの特性を持ったワーカーモデルを呼び出すべきか。
- タスクの細分化と指示: 選定したモデルに対して、どのような具体的なサブタスクを与えるか。この際、Conductor自身がエキスパートレベルのプロンプトエンジニアとして振る舞い、ワーカーが最大のパフォーマンスを発揮できるようなカスタム指示文（プロンプト）を動的に設計・付与する。
- コンテキストの管理: 呼び出されたエージェントのコンテキストウィンドウ（記憶領域）に、これま

でのタスク履歴や他のエージェントの出力のうち、どの部分を含めるべきかを判断する。これにより、不要な情報によるハルシネーション（AIの幻覚）を防ぎ、トークン消費を最適化する。

Conductorの強化学習プロセスは、タスクの完了という最終的な報酬の最大化のみを目的としており、その結果として、自然言語を介した極めて効果的なコミュニケーション・プロトコルやエージェント間の連携トポロジーが「創発的」に発見されている[12]。簡単な事実確認のクエリに対しては即座に単一モデルで回答し、難易度の高いコーディング課題に対しては、プランナー、コーダー、レビュアーで構成される多段階のパイプラインをオンザフライ（即座）に構築するなど、タスクの複雑さに応じて自律的に処理経路を柔軟に変化させる能力を持つ[11]。特筆すべきは、Conductorが戦略策定の権限をGemini 2.5 Proに完全に委譲し、残りのプールへのサブタスク指示を当該モデルに決定させるといった、高度な適応性を示すケースも確認されている点である[13]。

再帰的テストタイムスケーリング（**Recursive Test-Time Scaling**）による自己修正能力

Conductorのアーキテクチャ設計の中で、最も画期的かつ推論パラダイムを根本から変える概念が「再帰的テストタイムスケーリング」の実装である[11]。Conductorは、オーケストレーションを行うプロセスにおいて、外部のモデルだけでなく、ワークフローの実行役として「自分自身のインスタンス」を再帰的にワーカープールの中から選択することが可能である[1]。

Conductorが自身を再帰的に召喚した場合、自身の編成したチームが過去に出力した結果を俯瞰的に読み取り、どのアプローチが行き詰まったのか、あるいは検証プロセスで何がエラーの根本原因となったのかを自己認識（メタ認知）する[11]。そして、そのフィードバックに基づいて、問題を解決するための全く新しい修正ワークフローの立ち上げや、アプローチの異なる別のモデルへのタスクの再割り当てをその場で動的に実行する。

この再帰的な自己修正ループの存在は、AIの性能を向上させるための「新たなスケーリングの軸」を提供する。これまでの基盤モデル開発では、より賢いAIを作るために「事前学習時（Pre-training）」に投入するデータ量と計算機リソース（GPU）を指数関数的に増やす必要があった。しかし、Fuguのアーキテクチャでは、問題を解く「推論時（Test-time / Inference）」における計算量と試行時間をシステム側で動的に拡張（スケーリング）し、自己修正を繰り返すことで、基盤モデル単体のパラメーター規模に依存することなく、非常に困難な課題に対する正答率を劇的に引き上げることが可能となったのである[11]。これは、人間が難問に直面した際に、時間をかけて何度も推敲を重ねることで正答に辿り着くプロセスをアーキテクチャレベルで模倣したものと言える。

定量および定性パフォーマンス：フロンティアモデルとの徹底比較

Sakana Fuguは、基盤となるプール内にOpenAIの「GPT-5.5」、Googleの「Gemini 3.1 Pro」、Anthropicの「Claude Opus 4.8」といった各社が誇る最先端のフロンティアモデル（SOTA）を含んでおり、これらの強力な「集合知」をConductorとTRINITYによって増幅させることで、複雑で多段階のタスクにおいて圧倒的な優位性を示している[2]。

定量的な評価基準である主要な高度業界ベンチマークにおいて、Fugu Ultraおよび標準のFuguモデルは、基盤となっている各フロンティアモデル単体の性能を凌駕、あるいはそれに匹敵するトップスコアを記録している[1]。以下の表は、最新の評価データに基づくベンチマークスコアの比較である。

ベンチマーク指標	タスクの性質・領域	**Fugu Ultra**	**Fugu (標準)**	主要競合モデルのスコア（単体実行時）
SWE-Bench Pro	複雑なソフトウェアエンジニアリング、実世界のGitHubリポジトリにおけるバグ修正や自律的エージェントタスク	**73.7**	59.0	Claude Opus 4.8: 69.2 OpenAI GPT-5.5: 58.6
LiveCodeBench	定期的に更新される最新のプログラミング課題解決（訓練データへの汚染を防ぐ動的テスト）	**93.2**	92.9	Gemini 3.1 Pro: 88.5
GPQA-Diamond	大学院生やドクターレベルの生物学・物理学・化学領域における極めて難解な多肢選択式推論	**95.5**	**95.5**	Claude Mythos Preview: 94.6 OpenAI GPT-5.5: 93.6
TerminalBench 2.1	システム管理、ターミナル環境の操作、およびコマンドラインを介したエージェント操作能力	**82.1**	80.2	OpenAI GPT-5.5: 78.2
SciCode	科学計算アルゴリズムの実装、物理モデリングのコーディング能力	58.7	**60.1**	Gemini 3.1 Pro: 58.9

Humanity's Last Exam	人類が到達し得る最高難易度の学際的推論テスト	50.0	-	Claude Fable 5: 53.3 Claude Opus 4.8: 49.8

これらの客観的なベンチマークデータから読み取れる最も注目すべき成果は、ソフトウェアエンジニアリング能力の最高峰を評価する「SWE-Bench Pro」において、Fugu Ultra（73.7）がOpus 4.8やGPT-5.5に対して明確かつ圧倒的なスコア差をつけていることである[3]。このベンチマークは、実際のコーディング環境における複数のファイルにまたがるバグ修正や、大規模なリポジトリのナビゲーションを要求する。このような長期的なコンテキストの保持と、コード生成後の複数ステップにわたるコンパイル・テスト検証が不可欠な領域において、Conductorによる厳密な役割分担（Thinker/Worker/Verifier）と再帰的な自己検証ループが、単一モデルの能力の限界を突破する上で極めて有効に機能していることを強く示唆している。
一方で、留意すべき特異点として、「SWE-Bench Pro」においてAnthropicの制限付き最上位モデルである「Claude Fable 5」がスコア80.0を記録しており、Fugu Ultra（73.7）がこれに届いていない事実がある[3]。同様に、極限推論を測る「Humanity's Last Exam」においても、Fable 5（53.3）がFugu Ultra（50.0）を上回っている[3]。しかし、これはFuguのマルチエージェント・アーキテクチャ自体の欠陥や劣後を意味するものではない。後述する米国政府による強力な輸出規制措置により、Fable 5やMythos 5といったAnthropicのトップモデルは一般公開が遮断されており、Fuguのエージェントプールにこれらの SOTAモデルを組み込むことが物理的・法的に不可能であったという外部要因に起因するものである[2]。すなわち、Fuguは「利用可能な」モデルリソースの範囲内において、オーケストレーションによって理論上の最高性能を叩き出していると言える。
定量的なスコアの優位性に加え、約500名のエンタープライズベータユーザーから実務テストを通じて寄せられた定性的なフィードバックは、ベンチマークテストでは測りきれない現実の実務環境（Messyなユースケース）におけるFuguの圧倒的な有用性を裏付けている[2]。

- 長期的なペルソナの一貫性とエージェントの安定性（**Agent Stability**）: 一般的なLLMは、会話のコンテキストが長くなり、タスクが数時間に及ぶにつれて、AIが設定されたキャラクターや役割を忘れ、出力内容がブレる現象（ペルソナ・ドリフト）を起こす。あるプラットフォーム企業の役員による評価では、Fugu Ultraの生の出力品質がトップモデルと同等であるだけでなく、長時間のセッションにわたってもペルソナのブレが極めて少なく、一貫したアイデンティティと目的意識を保ち続けたことが高く評価された[2]。
- 徹底したコードレビューと欠陥検出の網羅性: プロのソフトウェアエンジニアからの評価では、Fugu UltraはGPT-5.5と比較して、他が見落とすバグを包括的かつ深く検出できると報告された。実際の業務コードにおいて、他の単一AIツールが約3件の問題しか指摘できなかったのに対し、Fugu Ultraは多角的な視点から20件以上の問題を検出し、潜在的な脆弱性の表面化と品質向上に大きく貢献した[2]。
- 自律的なサイバーセキュリティ診断能力: 熟練したセキュリティエンジニアの監督下で行われたテストにおいて、Fugu Ultraは破壊的なアクションを安全に回避しながら、システムへの偵察活動、クロスサイトスクリプティング（XSS）やSQLインジェクションのチェック、複雑な認証メカニズムのレビューを自律的に実行した。さらに、単に脆弱性を指摘するだけでなく、再現のため

のエビデンスと再テスト手順を含んだ完全でクリーンな報告書を生成することに成功している[2]。

- 自律型データサイエンス（**AutoResearch**）におけるブレイクスルー: 小規模なGPTモデルの学習レシピをAIエージェントに自律的に改善させる実験において、人間の介入がほとんど、あるいは全くない状態で、Fugu Ultraはアイデアの探索、実験の実行、失敗の解釈、アプローチの修正、および進捗の更新を自ら行った。その結果、検証用のBPB（bits-per-byte）スコアにおいて、平均0.9774（± 0.0019）、単一ランでの最高値0.9748を記録し、競合する他のフロンティアモデルを上回る顕著な進展をもたらした[2]。

地政学リスクの顕在化とAI主権：Anthropic輸出規制問題への最適解

Sakana Fuguのリリースが AI業界全体に与える真のインパクトは、単なるアーキテクチャの技術的な革新性のみならず、現代のAIビジネスが直面する「地政学的な脆弱性」に対する直接的かつ実用的なソリューション（ヘッジ手段）を提示したことにある。Fuguが開発され、その必要性がエンタープライズ市場で急激に高まった背景には、本システムのリリース直前である2026年6月中旬に発生した、前代未聞のAI輸出規制事件が深く関係している[3]。

2026年6月12日午後5時21分（米国東部時間）、米国政府の商務省（ハワード・ラトニック長官主導）は、国家安全保障上の重大な懸念を理由として、米国を拠点とする主要AI企業Anthropicに対して厳格な輸出管理指令（Export Control Directive）を突如として発動した。この指令により、Anthropicは同社の最高峰モデルである「Claude Fable 5」および「Claude Mythos 5」への非米国籍者（世界中の海外ユーザーおよび企業）からのアクセスを即座に遮断することを余儀なくされた[15]。Anthropicにはこの規制を実施するための猶予期間がわずか90分しか与えられなかったと報じられている[18]。

この劇的な措置の発端については、複数の要因が絡み合っている。一部の報道によれば、中国政府に関連する組織が同モデルのセーフガードをバイパス（ジェイルブレイク）し、モデルをリバースエンジニアリング（蒸留）しようとした疑いが持たれている[18]。また同時に、Amazonの研究者が「Fable 5」のセーフガードの迂回手法を発見し、それを米国当局に警告したことが決定打となったともされている[15]。トランプ政権下のホワイトハウスは、当初はAIに対する軽度な規制アプローチを推進していたものの、この事態を重く受け止め、AIモデルを「軍事や諜報活動に転用可能な兵器・軍需品」と同等の「二重用途技術（Dual-use technology）」として分類する強硬姿勢へと転じた。法的な根拠としては、大量破壊兵器（WMD）の開発に関与する可能性のある活動を制限するための包括的な権限を持つ「ECRA（2018年輸出管理改革法）」に基づく「is-informed（通知済）」レターが使用されたとみられており、これは通常、半導体製造装置の対中輸出規制などに用いられる強力な法的ツールである[16]。

しかし、サイバーセキュリティの専門家からは、この輸出規制のトリガーとなった「ガードレールのバイパス手法」に対する見解の相違も指摘されている。Luta Securityの分析によれば、指摘されたバイパス手法とは、既知の脆弱性（CVE）を含むオープンソースコードや意図的にバグを埋め込んだコードをモデルに提示し、マニュアルの多段階プロセスを経て「コードを修正し、テストスクリプトを生成させる」という、サイバー防御担当者が日常的に行っている基本的な「発見・修正・テスト（find, fix, and test）」のループに過ぎなかった。専門家は、AIモデルにバグ修正を要求する行為自体が「軍需品」として輸出規制の対象となるのは過剰反応であり、米国のサイバー防衛能力そのものを弱体化させる「見当違いの措置」であると批判している[14]。

この論争の的となった事件は、特定のAIプロバイダーが提供するAPIに深く依存して自社の基幹シス

テムやサービスを構築する「ベンダーロックイン（Concentration Risk）」が、単なるビジネス上の制約ではなく、一国における規制境界の変更、外交政策の変動、あるいは政府の突発的な安全保障上の懸念によって、「一夜にしてシステム全体が機能不全に陥る致命的な運用リスク」であることを、世界中の企業や国家機関に痛感させる結果となった[2]。リモートの基盤モデルへのアクセスを国家権力によって制限したこの初の大規模事例は、AI業界における新たな地政学的な先例を確立したのである。

Sakana AIは、Fuguのマルチエージェント・アーキテクチャを、このような予測不可能なシングルベンダー依存を排除し、地政学的リスクを軽減するための明確な「ヘッジ手段」として位置付けている[1]。Fuguは、基盤となるエージェントプールを完全に「スワップ可能（入れ替え可能）」に設計している。もしAnthropicの事件のように、特定のプロバイダーがダウンタイムを起こしたり、政府の政策変更や予期せぬ輸出規制によって一部の高性能モデルへのアクセスが突然遮断されたりした場合でも、Fuguを採用している企業のシステム全体が停止を招くことはない。Fuguのオーケストレーターが自動的にAPIの障害やアクセス拒否を検知し、瞬時に別のプロバイダーのモデル（例えばGPT-5.5やGemini 3.1 Pro、あるいは強力なオープンソースモデル）を利用して処理経路を迂回（動的ルーティング）させる。これにより、システムは一切の運用上の混乱やコードの書き換えを伴うことなく、途切れることなく最高峰の集合知を提供し続けることができるのである[1]。

Sakana AIはこの自己修復的なルーティング能力を、企業および主権国家が自身のデータとITインフラに対するコントロールを恒久的に維持する「AI主権（AI Sovereignty）」を担保するための、最も現実的で強靭な選択肢であると説明している[2]。通常、複雑なマルチベンダー環境を自社で構築し、各社のAPI仕様の変更に追従し、障害復旧策（フェイルオーバー）を策定し、複数のプロバイダーとの契約を管理することは、企業にとって膨大な運用上のオーバーヘッド（手間とコスト）となる。Fuguは、これらの複雑なエコシステム全体の管理を「単一のOpenAI互換APIエンドポイント」という形で完全に抽象化し、顧客を地政学的な変動リスクの最前線から保護する強力な防波堤として機能しているのである[3]。

ただし、このブラックボックス化された動的データルーティング構造には、地域的な法規制との摩擦という運用上の課題も残されている。Fuguがユーザーのクエリをプール内のどの外部モデルに送信し、どのように処理を分散させるかという具体的なルーティングプロセスは、プロプライエタリ（独自の非公開技術）として設計されており、ユーザーからは隠蔽されている[3]。この構造上、データの処理拠点や移転先を厳密に特定・追跡することが困難となるため、リリース時点（2026年6月）において、厳格なデータ保護規則であるGDPR（一般データ保護規則）を適用するEU（欧州連合）およびEEA（欧州経済領域）からは、Fuguシステムへのアクセスが制限されている。Sakana AIは現在、このGDPRコンプライアンスの課題解決に向けた調整を進めている段階である[2]。

エンタープライズ導入における経済学：料金体系と「オーケストレーション・コスト」の理解

最先端のマルチエージェント・オーケストレーションを自社の業務プロセスに組み込むにあたり、企業のIT部門やCIOが最も注視するのが、コストパフォーマンスと詳細な料金体系である。Sakana Fuguは、広く普及しているOpenAI互換の単一APIフォーマットを採用しているため、企業は既存のクライアントアプリケーションやコーディング環境（SDK）からの煩雑な移行作業（マイグレーション）を行う必要がなく、接続先のエンドポイントとAPIキーを変更するだけで即座にフロンティアレベルのマルチエージェント群を統合できるという、極めて低い導入障壁を実現している[1]。

Fuguの料金プランは、個人のインディー開発者から、大規模な計算リソースを継続的に消費するグ

ローバルエンタープライズまで、あらゆる層のニーズをカバーするため、主に「月額サブスクリプション」と「従量課金制（Pay-as-you-go）」の2軸で展開されている[2]。

月額サブスクリプションプラン（定額制）

日常的および集中的なハンズオン利用を想定したサブスクリプションプランであり、いずれの階層を契約しても、標準の「Fugu」と最上位の「Fugu Ultra」の両モデルにアクセスが可能である。予測可能なランニングコストを重視するユーザーに向けた設計となっている[2]。

- **Standard**（月額**20**ドル）：日常的な軽量タスク、個人的なコーディング支援、学習用途に最適化されたエントリープラン。
- **Pro**（月額**100**ドル）：Standardプランの10倍の利用枠（トークンキャパシティ）が提供される。週に数回、集中的に高負荷な推論作業やデータ分析を行う開発者や小規模なリサーチチームに適している。
- **Max**（月額**200**ドル）：Standardプランの20倍という膨大な利用枠を誇り、長時間にわたって自律的なエージェントタスクやスクレイピング、継続的なコード監査などの高負荷なワークフローを稼働させ続けるヘビーユーザー向けである。
- ※ローンチプロモーションとして、2026年7月31日までにいずれかのサブスクリプションプランに契約した初期ユーザーには、2ヶ月目の利用料が無料となるインセンティブが提供されている[3]。

従量課金プラン（エンタープライズ向け・コンサンプションベース）

より柔軟なキャパシティの拡張性（エラスティシティ）と、サブスクリプションユーザーよりも高い処理優先度を求める企業向けに、使用したトークン量に応じて厳密に課金される従量課金制が用意されている。このプランの最大の特徴は、提供される2つのバリアント（FuguとFugu Ultra）による課金ルールの根本的な違いである[2]。

Fugu（標準モデル）の動的料金構造：標準モデルでは、基盤となる様々な外部APIを利用して処理が行われる。システムが背後で複数のエージェントを同時に稼働させ、検証ループを回したとしても、それぞれのプロバイダーモデルの料金が単純に合算（スタック）され続けるわけではない。当該タスクの解決に関与した基盤モデル群のうち、「最も料金レートの高い（最上位階層の）単一モデルのレート」のみが適用される仕様となっている。これにより、ユーザーはマルチエージェント特有の意図せぬコストの爆発的増加（ビル・ショック）を未然に防ぐことができる[2]。

Fugu Ultraの固定トークン料金構造：最高の推論性能を提供するFugu Ultra（モデルID：fugu-ultra-20260615）は、利用する入力・出力トークン数に基づく固定の独自料金体系を採用している。コンテキストウィンドウの消費規模（272Kトークンを境界とする）によって、適用されるレートが変動する階層型構造となっている[2]。

ワークロードの規模	コンテキストの要件	入力トークン（**100**万あたり）	キャッシュ入力（**100**万あたり）	出力トークン（**100**万あたり）
標準ワークロード	272Kトークン以下	$5.00	$0.50	$30.00

極限ワークロード	272Kトークン超過	$10.00	$1.00	$45.00

オーケストレーションにおける「目に見えないバックグラウンドコスト」の理解

エンタープライズ統合において、企業のIT運用担当者および財務部門が明確に理解しておくべき、極めて重要な警告事項（Caveat）が存在する。それは、マルチエージェントシステム特有の「バックグラウンド・トークン」の扱いである。
Fuguが複雑なタスクを処理する際、背後ではConductorによるサブタスクの割り当て指示、エージェント間での通信のやり取り、そしてTRINITYのVerifierによるコードの自己検証プロセスが何往復にもわたって繰り返されている。これらのシステム内部での推論と調整のプロセスで消費および生成されるトークンは、Sakana AI側で吸収・免除されるわけではなく、標準の課金対象としてユーザーの最終的なリクエスト費用に厳密に加算されるという事実である[3]。
つまり、単一の基盤モデルにプロンプトを投げて直接的な回答を一度だけ得る従来の手法と比較して、Fugu Ultraが複雑な問題を解決する過程では、ユーザーの画面に表示される最終的な出力テキストの何倍、場合によっては何十倍ものトークンが「内部でのオーケストレーション」のために消費される可能性がある。APIのレスポンスメタデータには、ユーザーに提示された「可視の生成トークン」と、内部の調整作業に費やされた「バックグラウンドトークン」を明確に分離した詳細な利用状況フィールドが含まれており、監査性を担保している[3]。
企業は、Fugu Ultraがもたらす「圧倒的な出力品質による人間の手戻り作業やバグ修正コストの削減効果」と、マルチラウンドの自律推論に伴って発生する「内部トークン消費の運用コスト（OPEX）」のトレードオフを、自社の具体的なユースケースに照らし合わせて慎重にROI（投資対効果）評価を行う必要がある。

結論：AI業界における「スケーリング則」の再定義と覇権構造の転換

Sakana AIが提示した「Fugu」のリリースとそれに伴うベンチマークの証明は、AIの知能向上における新たなアプローチの有効性を市場に突きつけた、極めて重要な歴史的転換点である。一部の開発者コミュニティやSNSプラットフォームであるRedditなどの議論においては、Fuguが独自に数兆パラメータの巨大モデルをスクラッチで学習させたわけではなく、外部の既存モデル（GPTやClaudeなど）をプールとして呼び出している構造から、「これは結局のところ、高度にラッピングされた単なるルーター（Wrapper）に過ぎず、モデルの基礎的な知能レベルが飛躍したわけではない」という懐疑的な見方や批判も一部存在した[4]。
しかし、本報告書の詳細なアーキテクチャ分析が示す通り、このシステムの本質的な価値は「ルーター」という単純な単語で片付けられるものではない。Fuguの真の革新性は、TRINITYの軽量な進化型コーディネーターやConductorの強化学習に基づく自然言語プロトコルという「メタ・インテリジェンス（調整レイヤー）」を通じて、既存モデルの能力を個別の限界以上に引き出す「推論時における計算力のスケーリング（Test-time Scaling）」を確立した点にある[11]。
これまで、OpenAI、Anthropic、Googleといったハイパースケーラーたちは、数千億円から数兆円規模の莫大なデータセンター投資（CAPEX）を行い、より多くのデータでモデルのパラメータを巨大化させる「腕力による競争」を繰り広げてきた。しかし、Fuguは他社が莫大な資本を投下して構築したこ

れらの巨大な計算資源群の上に、「オーケストレーション」というインテリジェントで軽量なソフトウェアレイヤーを被せ、適切に役割分担させることによって、最新鋭のモノリシックモデルと同等、あるいはそれ以上のパフォーマンスを、極めて資本効率の高い形で実現してしまったのである[4]。Redditの高度な議論において指摘された通り、この結果は「巨大で高価なクローズド基盤モデルの性能向上曲線がいずれピークを迎え、適切にオーケストレーションされた中・小規模モデルの動的な集合体がコストと性能の両面で市場を制覇する」という、AIエコシステムにおける新たな覇権構造の始まりを強く示唆している[4]。AI関連企業のIPOが急増している背景には、企業価値の源泉が「巨大モデルの所有」から「連携技術の所有」へと移行しつつあることへの市場の敏感な反応が透けて見える。
同時に、FuguはAIが高度な地政学的武器や二重用途技術として扱われる現代において、グローバル企業が生き残るための「レジリエンス（回復力）」の模範解答を提示した。輸出規制や突発的な政策変更によって、最も依存していたAIモデルが前触れなく使用不能になるリスクがAnthropicの事例によって現実化した今、ベンダーロックインを構造的に回避し、システムの冗長性とデータ処理の自律性を確保するアプローチは、すべての国家機関と企業にとっての必須要件となる。
Sakana Fuguは、完成され固定化されたソフトウェアパッケージではなく、強化学習によって継続的に自律進化するオーケストレーション基盤である。今後数か月のうちに、同社独自の学習モデルや日々進化する新たなオープンソースモデルが順次プールに統合され、エージェント間の協調トポロジーはさらに複雑かつ強力に進化していくことが予想される[2]。Sakana AIが掲げた「One Model to Command Them All（すべてを統べる一つのモデル）」というコンセプトは、AIの発展の歴史が「単体の規模の拡大」から「多様な専門家の集合知によるダイナミックなエコシステムの構築」へと決定的なシフトを遂げたことを告げる、最も鮮明で象徴的なマイルストーンとして評価されるべきである。

引用文献

1. Sakana AI Launches Sakana Fugu: An Orchestration Model That Routes Tasks Across a Swappable Pool of Frontier LLMs, 6月 23, 2026にアクセス、https://www.marktechpost.com/2026/06/22/sakana-ai-launches-sakana-fugu-an-orchestration-model-that-routes-tasks-across-a-swappable-pool-of-frontier-llms/
2. Sakana AI、ミュトス級謳う“集合知”AIモデル「フグ」 - Impress Watch, 6月 23, 2026にアクセス、https://www.watch.impress.co.jp/docs/news/2119048.html
3. No Claude Fable 5? No problem: Sakana achieves frontier performance with new Fugu multi-model, auto synthesis system, 6月 23, 2026にアクセス、https://venturebeat.com/orchestration/no-claude-fable-5-no-problem-sakana-achieves-frontier-performance-with-new-fugu-multi-model-auto-synthesis-system
4. Fable 5 has been beaten by Sakana.ai's Fugu in some tasks, claims Japanese startup in the new release - Reddit, 6月 23, 2026にアクセス、https://www.reddit.com/r/LLMDevs/comments/1uca8e3/fable_5_has_been_beaten_by_sakanaais_fugu_in_some/
5. Sakana Fugu: One Model to Command Them All, 6月 23, 2026にアクセス、https://sakana.ai/fugu-release/
6. Mitigating vendor lock-in with Sakana AI Fugu multi-agent models, 6月 23, 2026にアクセス、https://www.artificialintelligence-news.com/news/mitigating-vendor-lock-in-saka

na-ai-fugu-multi-agent-models/
7. Sakana Fugu: A Multi-Agent Orchestration System as a Foundation Model, 6月 23, 2026にアクセス、https://sakana.ai/fugu-beta/
8. Sakana AI releases Fugu Ultra system to rival top AI labs, 6月 23, 2026にアクセス、https://www.testingcatalog.com/sakana-ai-releases-fugu-ultra-system-to-rival-top-ai-labs/
9. Trinity: An Evolved LLM Coordinator - Sakana AI, 6月 23, 2026にアクセス、https://sakana.ai/trinity/
10. TRINITY: An Evolved LLM Coordinator - Sakana AI, 6月 23, 2026にアクセス、https://sakana.ai/assets/fugu-beta/TRINITY_poster.pdf
11. Learning to Orchestrate Agents in Natural Language with the ..., 6月 23, 2026にアクセス、https://sakana.ai/learning-to-orchestrate/
12. Learning to Orchestrate Agents in Natural Language with the Conductor - arXiv, 6月 23, 2026にアクセス、https://arxiv.org/html/2512.04388v1
13. How Sakana trained a 7B model to orchestrate GPT, Claude and Gemini LLMs, 6月 23, 2026にアクセス、https://venturebeat.com/orchestration/how-sakana-trained-a-7b-model-to-orchestrate-gpt-5-claude-sonnet-4-and-gemini-2-5-pro
14. The Fable 5 Export Controls Harm US Cyber Defense - Luta Security, 6月 23, 2026にアクセス、https://www.lutasecurity.com/post/the-fable-5-export-controls-harm-us-cyber-defense
15. Commerce Restricts Anthropic's Fable and Mythos Access, 6月 23, 2026にアクセス、https://letsdatascience.com/news/commerce-restricts-anthropics-fable-and-mythos-access-078b1e71
16. Legal Considerations Related to the Anthropic “Export Controls Directive” - Just Security, 6月 23, 2026にアクセス、https://www.justsecurity.org/142745/law-anthropic-export-controls/
17. Anthropic Fable 5 and Mythos 5 Pulled Offline: Inside the Export-Control Ban, 6月 23, 2026にアクセス、https://letsdatascience.com/blog/anthropic-fable-mythos-export-control-ban
18. Anthropic had 90 minutes to restrict Claude Fable 5 as White House feared Chinese access, 6月 23, 2026にアクセス、https://www.businesstoday.in/technology/artificial-intelligence/story/anthropic-had-90-minutes-to-restrict-claude-fable-5-as-white-house-feared-chinese-access-537095-2026-06-16
19. How Anthropic lost the White House's trust — and then its flagship product, 6月 23, 2026にアクセス、https://www.washingtonpost.com/technology/2026/06/15/how-anthropic-lost-white-houses-trust-then-its-flagship-product/
20. Sakana AI Labs Launches Sakana Fugu, Multi-Agent System Competes with Top Models, 6月 23, 2026にアクセス、https://www.kucoin.com/news/flash/sakana-ai-labs-launches-sakana-fugu-multi-agent-system-competes-with-top-models