

次世代AIアーキテクチャ「Sakana Fugu」：集合知によるAI主権の再定義

Sakana Fugu API

7B Conductor（オーケストレーター）
強化学習により、入力された複雑なタスクを最適なサブタスクに分解し、どの専門エージェントを呼び出すべきかを自然言語で自律的に判断・指揮します。

システムの心臓部：自律的オーケストレーションの仕組み

TRINITY（役割アロケーター）
2万未満の極めて軽量のパラメータで、選定されたLLMに対して「Thinker（思考）」「Worker（実行）」「Verifier（検証）」の3つの役割を動的に割り当てます。

再帰的テストタイムスケール
推論時（Test-time）に自己修正ループを回し必要に応じて「自分自身のインスタンス」を再帰的に呼び出すことで、モデル規模に依存せず正答率を劇的に引き上げます。

集合知（Swarm Intelligence）

専門エージェントの役割分担（TRINITYによる定義）

Thinker（思考役）
複雑な問題に対する高次元の解決戦略を立案し、設計書作成に相当する深層分析を行います。



Worker（実行役）
戦略に基づき、具体的なコード法、数学的推論、データ抽出などの実務ステップを実行します。



Verifier（検証役）
生成された解決案に矛盾やバグがないかを厳格に評価。不十分な場合はThinkerやWorkerに差し戻し、修正を促します。



驚異的なパフォーマンスと導入モデル

SWE-Bench Pro (ソフトウェア開発・バグ修正)	Fugu Ultra	★ 73.7 世界最高レベル
	Opus 4.8	69.2
LiveCodeBench (最新プログラミング課題)	Fugu Ultra	93.2
	Gemini 3.1 Pro	88.5
GPQA-Diamond (大学院レベルの難問接戦)	Fugu Ultra	95.5
	GPT-5.5	93.6

Fugu（標準）vs Fugu Ultra（最上位）

標準モデルは低レイテンシと「特定プロバイダーのオプトアウト（除外）機能」を提供し、Ultraは全エージェントを強制活用して極限の精度を追求します。



フロンティアモデルプール

戦略的価値：AI主権とコスト構造



ペンダーロックインの回避とAI主権

特定のAPIが援割や曝露で停止しても、自動的に別のモデル（スワップ可能）へ経路を迂回させる動的ルーティングにより、事業継続性を確保します。

地政学的リスクへのヘッジ：2026年のAnthropic純出授勤のような事象に際しても、Fuguの抽象化レイヤーによりシステム全体の機能不全を防ぐことが可能です。

料金体系と「バックグラウンド・トークン」

月賦制(Standard 520〜)と設置費があり、内部の自己検証プロセスで消費される「目に見えないトークン」も料金対象となる点に注意が必要です。

