

Sakana Fuguの提供開始に関する詳細調査と分析

エグゼクティブサマリー

Sakana AIは2026年6月22日に、複数の外部LLMを背後で協調させながら、利用者からは**単一モデルAPI**として使える新製品 **Sakana Fugu** の正式提供開始を公表しました。公式の製品ページ、リリース、コンソール文書、料金表、利用規約、プライバシーポリシーから確認できる範囲では、現時点の商用提供形態は**クラウド型のマネージドAPI**であり、モデルは **fugu** と **fugu-ultra** の2系統です。βテストは2026年4月24日に始まり、6月22日の発表はGA相当の一般提供開始と読むのが妥当です。 1

実務上もっとも重要なのは、「**高性能なマルチモデル協調を、OpenAI互換APIでそのまま使える**」点と、「**データ保護の初期設定が必ずしも保守的ではない**」点です。前者については、Sakana FuguはChat Completions / Responses / Models APIに対応し、OpenAI SDK互換、Codex CLI向け導線、組み込みの **web_search** ツールにも対応しています。後者については、利用規約上、入力・出力コンテンツは**自社モデルの学習・改善・分析**に使われ得て、人手レビューもあり得るうえ、外部LLMプロバイダーやクラウド事業者への開示も明記されています。したがって、企業導入では**秘密情報・個人情報・高機密研究データをそのまま投入する運用は推奨しにくく**、少なくともopt-out、入力制限、契約精査、社内利用基準が必要です。 2

性能面では、Sakana AIの公式コンソール掲載値によれば、**Fugu Ultra** はSWE Bench Pro 73.7、Terminal Bench 2.1 82.1、LiveCodeBench 93.2、LiveCodeBench Pro 90.8、Humanity's Last Exam 50.0、GPQA Diamond 95.5などを記録し、Opus 4.8、Gemini 3.1、GPT-5.5の比較対象を多くの項目で上回っています。一方で、**全項目で一律最強ではありません**。たとえばMRCrV2ではGPT-5.5が94.8で首位、CTI-REALMではOpus 4.8が69.6で僅差首位です。つまり、Sakana Fuguの価値は「万能単独モデル」ではなく、**難問に応じて複数モデルの強みを束ねることで総合性能を押し上げる**ことにあります。 3

価格は公式に**USD建て**で提示されています。月額サブスクはStandard \$20、Pro \$100、Max \$200。従量課金の**Fugu Ultra** は標準コンテキストで**入力 \$5 / 出力 \$30 / キャッシュ入力 \$0.50 (各1M tokens)**、272K超コンテキストで**入力 \$10 / 出力 \$45 / キャッシュ入力 \$1.00**です。ユーザー要請どおりJPYでも推定すると、**1 USD ≈ 161.289 JPY**の参考レートではStandard 約¥3,226、Pro 約¥16,129、Max 約¥32,258、Fugu Ultraの標準入力は約¥806/100万tokens、標準出力は約¥4,839/100万tokensです。なお、**Fugu Ultraはオーケストレーション由来のトークンも課金対象**で、OpenAI系の感覚より請求が膨らむ可能性があります。 4

結論として、Sakana Fuguは**研究プロダクトではなく、すで実運用可能な“オーケストレーション層”の商用API**です。ただし、現時点の公開資料だけでは**完全なエージェントプール一覧、各ワーカーモデルの固定構成、正式なDPA/SOC2/ISO情報、EU向け専用コンプライアンスパック**は確認できません。したがって、導入判断は「性能」だけでなく、**データ利用条件・規制適合・説明責任・監査可能性**を含めて行うべきです。 5

事実関係の確認と前提

まず、ユーザー提示文の中核事実——「Sakana AIは2026年6月22日、複数のAIモデルを組み合わせでタスクをこなすAIシステム『Sakana Fugu』の提供を始めると発表した」——は、**公式リリースと公式製品ページで裏づけ可能**です。Sakana AIの2026年6月22日付ブログ「Sakana Fugu: One Model to Command Them All」は、新製品としてSakana Fuguを紹介し、単一のモデルAPIからアクセスできる**マルチエージェント・オーケ**

ストレーションシステムだと説明しています。日本語側の公式文面でも「提供を開始します」と明記されています。日本語メディアの gihyo.jp も同日、「正式提供を開始」と報じています。 ⁶

ただし、**時系列には注意が必要です**。Sakana Fuguは突然現れた製品ではなく、2026年4月24日に**βテスト開始**が公式ブログで公表されていました。したがって、2026年6月22日は「初登場」ではなく、**βから一般提供への移行日**と整理するのが正確です。 ⁷

不明点もあります。まず、**提供地域**について、ユーザー要請どおり「国内のみかグローバルか」は発表文だけでは明示されていません。公開利用規約では、サービスは**EEA、英国、スイス以外の地域**向けとされており、これを文字どおり読むと、日本国内限定ではなく、少なくともそれ以外の相当広い地域を想定した提供です。ただし、公式リリースで公開された「ローンチ地域一覧」は確認できませんでした。したがって、“**国内のみ**”とは言えない一方、“**完全グローバル**”とも断定しないのが厳密です。 ⁸

次に、**価格の通貨**については公式はUSD建てです。ユーザー要請に従いJPYでも推定を併記しますが、JPYはあくまで参考換算です。ここでは、XEの2026年6月20日時点の表示値 **1 USD = 161.289 JPY** を用いた近似を示します。市場レートは日々変動するため、執行予算には余裕を見てください。 ⁹

さらに、メディア報道には**モデル名の揺れ**があります。公式ドキュメントは **Fugu / Fugu Ultra** の2モデル表記で一貫していますが、ビジネス+IT記事には「Fugu Mini」と記された箇所があります。現時点で**公式コンソールの正式モデルID**は **fugu** と **fugu-ultra** のみであり、**fugu-mini** は公開モデル一覧に出ていません。したがって、製品名としては「Mini」ではなく「Fugu」が公式と扱うべきです。 ¹⁰

製品仕様と商用条件

Sakana Fuguの公開提供形態は、現時点では**APIサービス**です。公式のGetting Startedでは、全APIエンドポイントが <https://api.sakana.ai> にあり、OpenAI SDK互換のクライアントから `base_url` を差し替えるだけで使えると説明されています。対応エンドポイントは **Chat Completions / Responses / Models API** で、実務上は **Responses API推奨** とされています。さらに、Codex CLI への導入スクリプトも用意されています。 ¹¹

商用条件の整理

以下は、公式価格表とFAQをもとに整理したものです。料金の公式提示はUSD、JPY欄は参考換算です。 ¹²

項目	公式価格	JPY参考換算	備考
Standard	\$20/月	約¥3,226/月	軽量な日常利用向け
Pro	\$100/月	約¥16,129/月	Standardの10倍利用枠
Max	\$200/月	約¥32,258/月	Standardの20倍利用枠
Fugu Ultra 入力	\$5 / 1M tokens	約¥806 / 1M tokens	標準コンテキスト
Fugu Ultra 出力	\$30 / 1M tokens	約¥4,839 / 1M tokens	標準コンテキスト
Fugu Ultra キャッシュ入力	\$0.50 / 1M tokens	約¥81 / 1M tokens	標準コンテキスト
Fugu Ultra 入力	\$10 / 1M tokens	約¥1,613 / 1M tokens	272K超コンテキスト
Fugu Ultra 出力	\$45 / 1M tokens	約¥7,258 / 1M tokens	272K超コンテキスト
Fugu Ultra キャッシュ入力	\$1.00 / 1M tokens	約¥161 / 1M tokens	272K超コンテキスト

重要な実務ポイントは3つあります。第一に、サブスクでも両モデルにアクセス可能です。第二に、Fuguの従量課金は“使った全モデルの積み上げ”ではなく、関与した最上位モデル単価の単一レートとされています。第三に、Fugu Ultraではオーケストレーション入力・出力トークンも課金対象であり、通常の単一LLMより請求予測が難しいです。 13

API・SDK・利用条件

公式ドキュメントから読み取れる技術仕様は次のとおりです。 14

項目	公開情報
ベースURL	<code>https://api.sakana.ai/v1</code>
対応API	Chat Completions / Responses / Models
SDK互換性	OpenAI SDK互換
接続先例	Python, cURL, Codex CLI
組み込みツール	<code>web_search</code>
モデルID	<code>fugu</code> , <code>fugu-ultra</code> , <code>fugu-ultra-20260615</code>
推奨API	Responses API
推定コンテキスト長	1,000,000 tokens (Codex用公式カタログ記載)
入力モダリティ	text, image
推論努力	<code>fugu</code> : <code>high</code> , <code>xhigh</code> / <code>max</code> 。 <code>fugu-ultra</code> は高負荷前提
非対応/制限	<code>temperature</code> , <code>top_p</code> , <code>stop</code> , <code>seed</code> , <code>presence_penalty</code> , <code>frequency_penalty</code> などは受理するが無視される場合あり。 <code>previous_response_id</code> は非対応

この仕様から、OpenAIアプリケーションの置換互換性は高い一方、完全互換ではないと評価できます。特に、生成パラメータの一部が“Accepted but ignored”と明記されているため、既存アプリで温度制御や `top_p` 依存の品質調整を行っている場合、同じ期待は持てません。 15

対応モデルの公開範囲

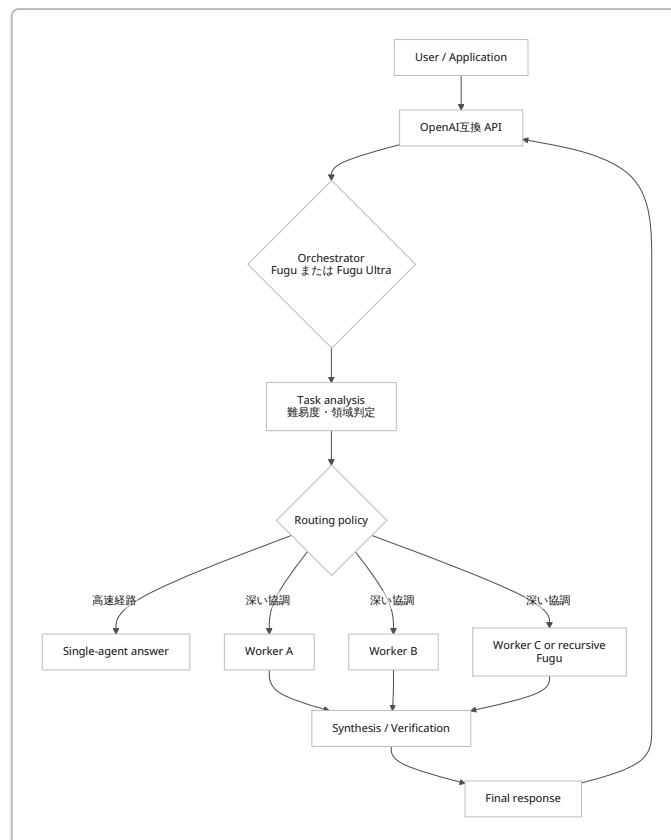
「対応モデル（各モデル名・役割・アーキテクチャ・パラメータ規模・学習データ）」については、公開されている範囲と非公開の範囲を分けて扱う必要があります。公式資料が明示している部分だけを整理すると、次のとおりです。 16

レイヤ	名称	役割	アーキテクチャ / 方式	パラメータ規模	学習データ・学習方法の公開範囲
商用 オーケストレータ	fugu	低レイテンシ寄りの標準モデル。タスクに応じて最適モデルヘルディング	単一モデルAPIとして見せる 商用オーケストレーションLM	非公開	β説明では Trinity / Conductor を基盤に改良と説明。商用版の詳細学習データは非公開
商用 オーケストレータ	fugu-ultra / fugu-ultra-20260615	高精度重視。1〜3 エージェントを問題ごとに連携	より深いエージェントプールを使う商用オーケストレーションLM	非公開	同上。オーケストレーション専用トークン計上あり
研究 ベース	Trinity coordinator	軽量ルータ / 役割割当	SLM + light head、Thinker/ Worker/ Verifier、フル対話履歴を入力	0.6B SLM + 約10K head、学習対象合計 20K未満	研究論文で開示。Qwen3-0.6Bベース、進化戦略で最適化
研究 ベース	Conductor	自然言語でワークフローを設計するコーディネータ	最大5ステップの workflow を自然言語 + Python list で出力	7B	研究論文で開示。RLでサブタスク、モデル選択、access list を学習
外部 LLM群	OpenAI / Anthropic / Google 系	ワーカーエージェント	外部LLMを順次または並列に利用	各社とも多くが非公開	利用規約で外部LLM例として明示
参考 ベース ライン	GPT-5.5 / Opus 4.8 / Gemini 3.1 Pro	公開ベンチ比較対象	各社独自	多くが非公開	公式比較表・定性比較で開示
非エー ジェン トプー ル比較 対象	Fable 5 / Mythos Preview	“肩を並べる”比較対象	各社独自	非公開	公式にFuguの agent poolには含まれないと明示

ここで大事ななのは、商用Fuguの完全な現在ワーカープール一覧は公開されていないことです。公式は外部LLMの例として OpenAI / Anthropic / Google を挙げ、Gihyo は GPT-5.5 / Claude-Opus-4.8 / Gemini-3.1-Pro を挙げていますが、どのモデルが今この瞬間の本番プールに常時入っているかは公開資料だけでは断定できません。逆に、Fable 5 と Mythos Preview は agent pool に入っていないことは公式が明記しています。

技術アーキテクチャと推論フロー

Sakana Fuguの技術的中核は、Sakana AIがICLR 2026で公表した **TRINITY** と **Conductor** の2本柱です。公式のβ記事および本リリースは、Fuguがこれら2論文をベースに改良されたと説明しています。つまり、公開資料から断定できるのは「**Fuguは learned orchestration を製品化したもの**」という点までです。以下のアーキテクチャ図は、それを踏まえて、公式資料に整合する範囲で再構成したものです。なお、**Fugu** と **Fugu Ultra** の内部差分の完全図は非公開のため、低レイテンシの `fugu` は Trinity 寄り、深い協調を使う `fugu-ultra` は Conductor 寄り、という部分は**推定**を含みます。 18



マルチモデル連携の仕組み

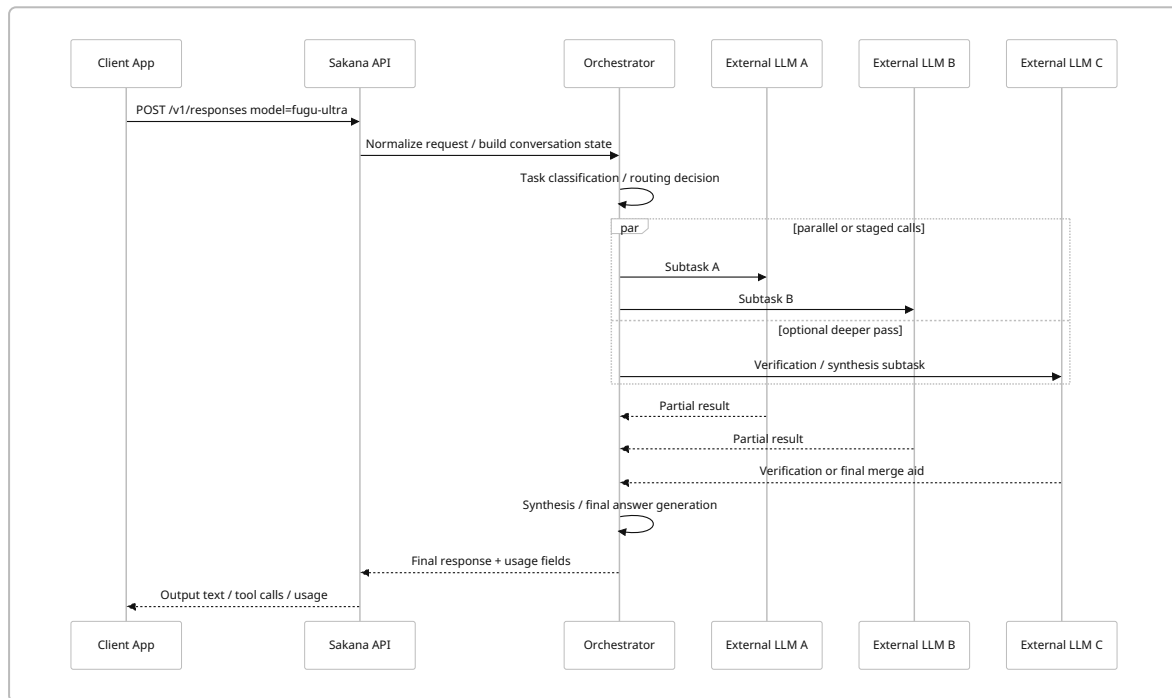
Trinity は、0.6Bの小型モデルの隠れ状態に軽量ヘッドを載せ、各ターンで「どのLLMを」「どの役割で」呼ぶかを決める方式です。役割は **Thinker / Worker / Verifier** の3つで、各ターンの入力には**元質問+過去ターンの全文脈**が入ります。学習対象は2万パラメータ未満で、軽量ながら高い文脈把握能力を持つのが特徴です。 19

Conductor は、より表現力の高い方式で、ワークフローそのものを自然言語で設計します。論文では、Conductorは**最大5ステップ**の workflow を、`model_id`、`subtasks`、`access_list` の3つのPythonリストとして出力し、どのワーカーに何をさせ、どの先行結果を見せるかを定義します。これにより、単純な sequential chain だけでなく、**並列・木構造・再帰的自己呼び出し**も可能になります。 20

Sakana Fugu商用版は、この研究系の特徴を実用面に落とし込んでいます。公式コンソールによれば、`fugu` は**低遅延と性能のバランス**を狙い、`fugu-ultra` はより深いエージェントプールを使って高精度化します。さらに、`fugu-ultra` は**1~3エージェントを問題ごとにルーティング**すると明記されており、従来の単なる1本ルータではなく、**複数呼び出し前提のオーケストレータ**です。 21

APIシーケンス

公式仕様に基づくAPIレベルの概念シーケンスは次のようになります。 22



レイテンシとスケーラビリティの推定

Sakana AIは実測レイテンシ数値を公開していません。したがって、以下は公開仕様に基づく推定です。根拠は、**fugu** が低遅延志向、**fugu-ultra** がより多くのエージェント協調を行うこと、**fugu-ultra** が1~3 エージェントを使うこと、複雑タスクではクライアントタイムアウト拡大を推奨していること、Codex向けにはストリームアイドルタイムアウトを2時間に設定していることです。 23

簡易推定モデル

- 単一外部LLM 1回呼び出しの往復を **2~6秒** と仮定
- オークストレータ判断の追加オーバーヘッドを **0.1~0.5秒** と仮定
- 2エージェント並列＋統合1回なら $\max(2\sim6, 2\sim6) + 2\sim6 + 0.1\sim0.5 \approx 4.1\sim12.5$ 秒
- 3エージェントを段階的に使う hard path なら **6~20秒台**、再帰を含むと **30秒超** も十分あり得る

このため、対話UIの既定値としては **fugu** が適し、**fugu-ultra** はバックグラウンドジョブや重分析向きという公式の位置づけは合理的です。特に **fugu-ultra** を社内分析や特許調査に使う場合、SSE/streaming と長めの client timeout を前提に設計すべきです。 24

スケーラビリティについては、Fugu自体が**stateless**でリトライが冪等だとCodex向け設定で説明されており、また **supports_parallel_tool_calls** も記述されています。ただし、真のボトルネックはSakana側のオークストレータよりも、**背後の外部LLM群のレート制限・在庫・地域制限**にあります。Sakanaが「新しいフロンティアモデル公開後、約2週間で再学習・評価してロールアウト」としている点からも、Fuguは**固定重みの単一モデル**というより“**継続再訓練されるオークストレーション製品**”です。 25

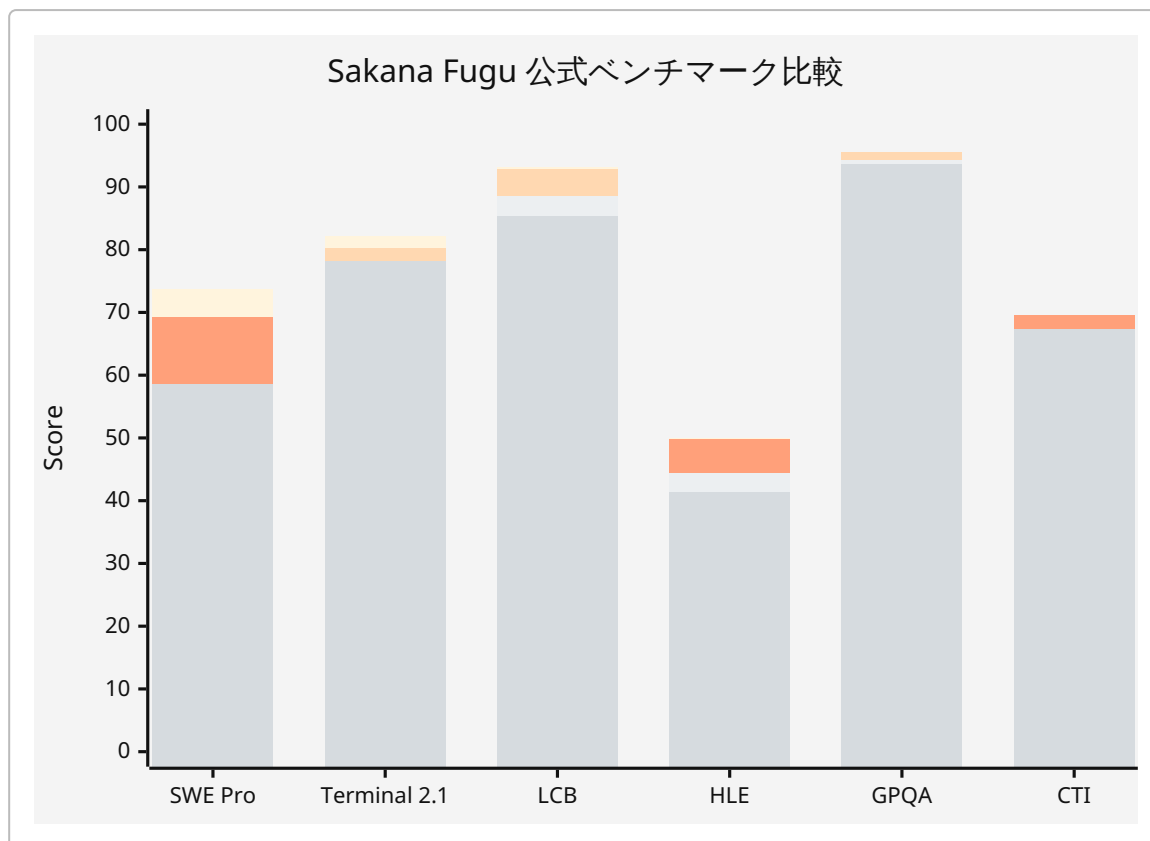
性能評価と競争ポジション

公開ベンチマークの読み方

Sakana AIの公式Modelsページは、Fugu Ultra / Fugu / Opus 4.8 / Gemini 3.1 / GPT-5.5 の比較表を掲載しています。重要な注意書きとして、製品ページでは一部ベースラインは**提供元公表値**を使用していること、また Fable 5 / Mythos Preview は**agent poolには入っていない**ことが明示されています。したがって、これらは**参考比較としては有用だが、すべて同一条件の独立再現試験とは限らない**という前提で読むべきです。

以下は公式表の主要数値です。

Benchmark	Fugu Ultra	Fugu	Opus 4.8	Gemini 3.1	GPT-5.5
SWE Bench Pro	73.7	59.0	69.2	54.2	58.6
Terminal Bench 2.1	82.1	80.2	74.6	70.3	78.2
LiveCodeBench	93.2	92.9	87.8	88.5	85.3
LiveCodeBench Pro	90.8	87.8	84.8	82.9	88.4
Humanity’s Last Exam	50.0	47.2	49.8	44.4	41.4
CharXiv Reasoning	86.6	85.1	84.2	83.3	84.1
GPQA Diamond	95.5	95.5	92.0	94.3	93.6
SciCode	58.7	60.1	53.5	58.9	56.1
τ ³ Banking	20.6	21.7	20.6	8.4	20.6
Long Context Reasoning	73.3	74.7	67.7	72.7	74.3
MRCRv2	93.6	86.6	87.9	84.9	94.8
CTI-REALM	69.4	67.5	69.6	56.0	67.3



この表から読み取れる実務的な結論は明確です。**Fugu Ultra**は、コーディング・科学推論・総合難問で非常に強い。一方で、長文脈や一部のメモリ系・サイバー系では既存単体モデルと拮抗または僅差で負ける項目もある。つまり、Fugu Ultraの価値は「必ずどのタスクでも勝つ」ことより、広いタスク分布で“外しにくくする”ことにあります。これは、研究論文が繰り返し強調する「各モデルの補完関係を学習して使い分ける」という思想と整合的です。 ²⁷

競合比較

競合は大きく3類型に分かれます。単一フロンティアAPI、ルータ/ゲートウェイ型、学習済みマルチエージェント・オーケストレータです。Sakana Fuguは3番目に属し、OpenRouterやNot Diamondとは似て非なる位置づけです。OpenRouterは主にモデル選択とプロバイダ選択の2層ルーティングを提供し、Not Diamondはモデルルーティングとプロンプト最適化に強い。一方Fuguは、公式説明上、複数モデルを逐次または並列に動かし、検証・合成まで内部で実施する点が差別化です。 ²⁸

製品	基本思想	API互換	協調の深さ	モデル/プロバイダ制御	価格モデル	向く用途
Sakana Fugu	学習済みオーケストレータが複数LLMを協調させる	OpenAI互換	高い。 1〜3エージェント、検証・合成あり	fugu は一部opt-out可、 fugu-ultra は固定プール	サブスク+従量	難問推論、研究、特許・論文調査、重いコーディング

製品	基本思想	API互換	協調の深さ	モデル/プロバイダ制御	価格モデル	向く用途
OpenRouter Auto Router	単一APIでモデル/プロバイダを自動選択	OpenAI互換	中程度。 主にルーティング/フェイルオーバー	非常に高い。 70+ providers, 400+ models	routed model rate	可用性重視、コスト最適化、ベンダー抽象化
Not Diamond	学習済み/カスタムルーター+プロンプト最適化	REST/ Python/ TS	中程度。 原則はベストモデル選択	カスタム学習可	個別設計寄り	既存アプリの精度改善、モデル選択最適化
単一モデルAPI	単体モデルの能力最大化	各社独自 または OpenAI互換	低い	ベンダー依存	トークン従量	低遅延 チャット、 統制容易、 挙動予測しやすい

実務上の差は、「ルーティング」だけでよいか、それとも“サブタスク分解→複数モデル協調→合成”までサービス側に持たせるかです。前者なら OpenRouter / Not Diamond 的発想でも十分ですが、後者——たとえば論文再現、Kaggle、特許調査、サイバー分析——ではFugu Ultraの設計思想が刺さりやすいです。実際に公式FAQも、そのような高難度用途を Fugu Ultra の想定例として挙げています。 ²⁹

セキュリティ・プライバシー・コンプライアンス

セキュリティとプライバシーの実像

Sakana Fuguの公開資料を読む限り、セキュリティ面で最重要なのは“外部LLM協調サービスである”こと自体です。利用規約は、このサービスを入力を1つ以上の外部機械学習モデルに動的ルーティングし、順次または並列に制御・選択・連携させるAPIベースのAIオーケストレーションサービスと定義しています。外部LLMの例として OpenAI / Anthropic / Google を挙げています。つまり、データはSakana AIの単独環境で完結する前提ではありません。 ³⁰

加えて、規約は入力・出力コンテンツを自社AIモデルの訓練、品質改善、マーケティング、セキュリティ確保、法令遵守等に使うと明記し、必要に応じて人手レビューが内容を注釈・処理する可能性も示しています。opt-outは可能ですが、既に学習に使われたデータの効果は遡及的に取り消せないとも書かれています。したがって、Fuguは少なくとも公開規約上、「デフォルトでゼロ学習・ゼロ保持・ゼロ人手レビュー」のエントリープライズ専用サービスではないと理解すべきです。 ³¹

プライバシーポリシーは、収集対象として氏名・メール・課金情報・プロンプト・アップロードコンテンツ・出力・ログ・IPなどを挙げ、受領者としてLLM、クラウドストレージ/インフラ、分析、決済、サポート事業者等のベンダーを明記しています。個人データは日本や米国を含む国外第三者へ移転され得るともしています。セキュリティ措置としては、アクセス制御、従業員教育、監視、サービスプロバイダ監督などが挙げられていますが、SOC 2 / ISO 27001 / 専用DPA の公開ページは、今回確認した公式資料群では確認できませんでした。 ³²

日本法の観点

日本企業がFuguを使う際の主要論点は、**APPI（個人情報保護法）**と**AI事業者ガイドライン**です。PPCは、クラウド利用が第三者提供に当たるかどうかは、外部事業者が**実際に個人データを取り扱うかどうか**で判断すると説明しています。Fuguは規約上、Sakana自身も外部LLMもコンテンツを処理し得るため、単なる“保管のみクラウド”に比べて**法務評価は重く見るべき**です。加えて、METI/総務省のAI事業者ガイドラインは、事業者AIガバナンス構築とリスク低減を求めています。³³

このため、日本企業の実務対応としては、少なくとも次が必要です。**個人情報・営業秘密・未公開発明情報の投入禁止または厳格限定、opt-out設定、利用部門別ルール、ベンダー審査、契約/約款精査、出力の人間確認**です。特に特許・論文調査用途は公式が想定例として挙げていますが、そのまま**未公開の発明要旨**を投げてよいことを意味しません。むしろ、その用途こそ情報分類ルールが必要です。³⁴

EUの観点

EUについては、まずSakanaの利用規約自体が**EEA・英国・スイス向け提供を想定していない**と述べています。この時点で、少なくとも2026年6月23日時点の公開ローンは**EU正面展開の製品ではない**とみられます。⁸

他方で、EU AI Actは一般目的AIモデルに対し、**技術文書、下流開発者への情報提供、著作権ポリシー、学習内容の要約公開**などの義務を設けています。欧州委員会のガイドラインでは、これらの義務は**2025年8月2日から適用**されています。今回確認したSakana Fugu公開資料では、TRINITY / Conductor / 技術レポートへのリンクはあるものの、**EU GPAI向けのモデル文書パックや学習データ要約の専用公開**までは確認できませんでした。したがって、**EU上陸時に追加のドキュメンテーション整備が必要になる可能性が高い**と評価できます。これは断定ではなく、公開資料ベースの保守的判断です。³⁵

米国の観点

米国では、Sakanaのプライバシーポリシーが**California addendum**を含み、CCPA上のアクセス、削除、訂正等の権利を認め、**個人情報を sell/share しない**と述べています。したがって、少なくともカリフォルニア消費者向けの最低限の整合は意識されています。もっとも、FTCは近年、AI企業の**プライバシー・セキュリティ表明と実運用の不一致**に強い関心を示しており、generative AI サービスでも過度な表明は規制リスクになります。Fuguは逆に、規約で広めのデータ利用を明言しているため、“誤解を招きにくい”反面、**企業顧客には厳しめの条件**です。³⁶

リスク・制約・業界インパクト

既知の制約と導入リスク

公開資料から見える制約は、主に以下のとおりです。各項目は「欠点」というより、**採用前に必ず理解すべき条件**です。³⁷

論点	公開資料からの評価	実務インパクト
データ利用	学習利用・分析利用・人手レビューの可能性あり	機密データ投入には慎重が必要
モデルプール透明性	完全一覧は非公開	調達・監査・再現性評価が難しい
EU提供	EEA/UK/CHは対象外	欧州展開には追加確認が必要

論点	公開資料からの評価	実務インパクト
Fugu Ultra制御	プール固定	ベンダー制限やデータ主権要件に弱い場合あり
API互換性	高いが完全ではない	一部SDK移植で挙動差が出る
料金予測	オーケストレーション分も課金対象	原価予測が難しい
レイテンシ	高難度で遅くなる前提	同期UXより非同期・stream設計向き
出力責任	規約で精度・適法性は非保証	人間レビュー必須

ここで特に重いのは、“**コンプライアンスのためにFuguでは特定プロバイダを除外できるが、Fugu Ultraではできない**”点です。Fugu Ultraは性能の源泉として full agent pool を固定としているため、**高性能と統制性のトレードオフ**が存在します。金融・製薬・防衛・知財など、情報流通先を厳密に制限したい企業では、この点が採否を左右します。 ³⁸

業界インパクト

Sakana Fuguの本質的な市場インパクトは、「**モデルそのもの**」より「**オーケストレーション層**」が製品カテゴリとして独立してきたことです。Sakana AIは、巨大単一モデルのスケール競争から、**交換可能な複数モデルを束ねることでフロンティア性能を出す方向**を明確に打ち出しています。しかも同社はこれを、単なる技術論ではなく、**輸出規制・単一ベンダー依存・AI主権の問題**と結びつけています。実際、AnthropicのFable 5 / Mythos 5へのアクセス停止・制限は、Sakanaの主張に現実味を与えています。 ³⁹

この意味で、Sakana Fuguは単に「強い日本発AI」ではなく、“**ベンダー間の性能差を吸収し、規制・供給変動にもある程度耐える推論ミドルウェア**”として見たほうが実態に近いです。今後、OpenRouter のようなルータ型、Not Diamond のような学習型ルータ、そしてFuguのような協調型オーケストレータが競争することになり、差別化軸は**品質・遅延・説明性・データ統制・監査容易性**に移ると考えられます。これは研究面でもRouteLLM、MasRouter、Smoothie、RouteMoA など「ルーティングそのもの」を研究対象とする流れと一致しています。 ⁴⁰

参考論文・特許・出典

関連論文の要約と重要引用

TRINITY: An Evolved LLM Coordinator は、軽量コーディネータが多様なLLMを複数ターンにわたって制御する研究です。Qwen3-0.6BベースのSLMに軽量ヘッドを付け、役割割当（Thinker / Worker / Verifier）とLLM選択を行います。Fuguの「低レイテンシ寄りの learned router」を理解するうえで最重要です。 ⁴¹

“under 20K” の learnable parameters。 ⁴²

Learning to Orchestrate Agents in Natural Language with the Conductor は、より表現力の高いオーケストレーション研究です。7BのConductorが、最大5ステップの workflow を自然言語で設計し、どのモデルにどのサブタスクを振り、どの過去結果を参照させるかまで学習します。Fugu Ultra の「深い協調」理解に直結します。 ⁴³

“three simple Python lists” で workflow を表現。 ⁴⁴

RouteLLM は、強いモデルと弱いモデルの間でクエリごとに動的選択するルーティング研究で、**コストと品質のトレードオフ最適化**が中心です。Sakana Fuguが「単なるルータ以上」であることを比較的是っきり示す対照軸になります。 ⁴⁵

“over 2 times” cost reduction without compromising quality. ⁴⁶

MasRouter は、単一モデルのルーティングではなく、**Multi-Agent System Routing** を collaboration mode、role allocation、LLM routing を一体で扱う点で、Fuguにかなり近い研究文脈です。Fuguが研究潮流の延長上にあることを示します。 ⁴⁷

“integrates all components of MAS into a unified routing framework.” ⁴⁸

関連特許の要約

今回確認した**Sakana AI名義のFugu関連公開特許番号**は、公式リリースや製品ページでは示されていません。その代わり、近縁技術として重要な公開特許を挙げます。 ⁴⁹

US20240311405A1 — Dynamic selection from among multiple candidate generative models with differing computational efficiencies は、計算効率の異なる生成モデル群から**動的選択**する発想を扱う公開特許で、Fuguの「問題ごとに最適モデルを選ぶ」という思想に近い先行技術群です。ただし、こちらは主として**選択側**であり、Fuguのような**多段協調・検証・合成**までを必ずしも前提にしていません。 ⁵⁰

“multiple candidate generative models with differing computational efficiencies.” ⁵¹

US12524210B2 — Hybrid inference system for COGS reduction は、ローカルモデルと外部LLMの間を**ルーティングモデルで振り分ける**特許で、コスト・遅延・受容率を同時に最適化する構想です。Fuguとは用途が異なりますが、「オーケストレーション層が経済性を左右する」という点で示唆が大きいです。特に、ユーザに受け入れられるかどうかをルータが予測し、ローカル／リモートを振り分ける発想は、将来の**オンプレ混在型Fugu**を考えるうえで興味深い比較対象です。 ⁵²

“predict whether output ... is likely to be accepted.” ⁵²

主要出典URL

以下は、今回の分析で特に重要だった一次・準一次ソースです。

```
https://sakana.ai/fugu-release/  
https://sakana.ai/fugu/  
https://sakana.ai/fugu-beta/  
https://console.sakana.ai/get-started  
https://console.sakana.ai/models  
https://console.sakana.ai/pricing  
https://console.sakana.ai/terms-of-service  
https://console.sakana.ai/privacy-policy  
https://console.sakana.ai/usage-policy  
https://github.com/SakanaAI/fugu  
https://arxiv.org/abs/2512.04695  
https://arxiv.org/abs/2512.04388  
https://arxiv.org/abs/2406.18665
```

<https://arxiv.org/abs/2502.11133>
<https://openrouter.ai/openrouter>
<https://openrouter.ai/blog/insights/model-routing/>
<https://docs.notdiamond.ai/docs/what-is-not-diamond>
<https://www.ppc.go.jp/en/legal/>
https://www.ppc.go.jp/all_faq_index/faq1-q7-53/
<https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers>
<https://digital-strategy.ec.europa.eu/en/factpages/general-purpose-ai-obligations-under-ai-act>
<https://oag.ca.gov/privacy/ccpa>
<https://www.ftc.gov/news-events/topics/protecting-consumer-privacy-security/privacy-security-enforcement>
<https://patents.google.com/patent/US20240311405A1/en>
<https://patents.google.com/patent/US12524210B2/>

1 5 6 16 39 49 **Sakana Fugu: One Model to Command Them All**

<https://sakana.ai/fugu-release/>

2 3 10 14 15 21 22 23 26 27 37 **Models**

<https://console.sakana.ai/models>

4 12 **Pricing**

<https://console.sakana.ai/pricing>

7 18 **Sakana Fugu: A Multi-Agent Orchestration System as a ...**

https://sakana.ai/fugu-beta/?utm_source=chatgpt.com

8 17 30 **Terms of Service**

<https://console.sakana.ai/terms-of-service>

9 **US Dollar to Japanese Yen Exchange Rate Chart**

https://www.xe.com/currencycharts/?from=USD&to=JPY&utm_source=chatgpt.com

11 25 **Get started**

<https://console.sakana.ai/get-started>

13 24 29 34 38 **Sakana Fugu — Multi-Agent System as a Model**

<https://sakana.ai/fugu/>

19 41 42 **Trinity: An Evolved LLM Coordinator**

<https://arxiv.org/html/2512.04695v3>

20 43 44 **Learning to Orchestrate Agents in Natural Language with the Conductor**

<https://arxiv.org/html/2512.04388v1>

28 **OpenRouter API and Models | OpenRouter**

<https://openrouter.ai/openrouter>

31 **Sakana Fugu Terms of Service**

https://console.sakana.ai/terms-of-service?utm_source=chatgpt.com

32 36 **Privacy Policy**

<https://console.sakana.ai/privacy-policy>

33 個人情報取扱事業者が、個人データを含む電子データを取り扱う情報システムに関して、クラウドサービス契約のように外部の事業者を活用している場合、個人データを第三者に提供したものとして、「本人の同意」（法第27条第1項柱書）を得る必要がありますか。または、「個人データの取扱いの全部又は一部を委託」（法第27条第5項第1号）しているものとして、法第25条に基づきクラウドサービス事業者を監督する必要がありますか。 | 個人情報保護委員会

https://www.ppc.go.jp/all_faq_index/faq1-q7-53

35 Rules for trustworthy artificial intelligence in the EU - EUR-Lex

https://eur-lex.europa.eu/EN/legal-content/summary/rules-for-trustworthy-artificial-intelligence-in-the-eu.html?utm_source=chatgpt.com

40 How OpenRouter Model Routing Works: Providers, Fallbacks & Auto Router — OpenRouter Blog

<https://openrouter.ai/blog/insights/model-routing/>

45 46 RouteLLM: Learning to Route LLMs with Preference Data

https://arxiv.org/abs/2406.18665?utm_source=chatgpt.com

47 48 MasRouter: Learning to Route LLMs for Multi-Agent Systems

https://arxiv.org/abs/2502.11133?utm_source=chatgpt.com

50 US20240311405A1 - Dynamic selection from among multiple candidate generative models with differing computational efficiencies - Google Patents

<https://patents.google.com/patent/US20240311405A1/en>

51 US20240311405A1 - Dynamic selection from among ...

https://patents.google.com/patent/US20240311405A1/en?utm_source=chatgpt.com

52 US12524210B2 - Hybrid inference system for COGS reduction - Google Patents

<https://patents.google.com/patent/US12524210B2/>