ChatGPT

Sakana AI「Fugu Ultra v2」「Fugu Max」徹底分析レポート

エグゼクティブサマリ

- **Sakana AIが2026年9月11日に公開したFugu MaxとFugu Ultra v2は、単体の巨大基盤モデルではなく、複数の外部・オープンウェイト・専門モデルを動的に選択・協調させる「学習済みオーケストレーター」である。** 両者は同じ中核オーケストレーション思想を共有し、Fugu Maxは「必要十分なモデルを選んでコスト効率を最大化」、Fugu Ultra v2は「複雑な多段タスクで絶対性能を最大化」という異なる最適化目標を持つ。Maxは入力$2／出力$6、Ultra v2は通常コンテキストで入力$5／出力$30（いずれも100万トークン当たり）である。SakanaはMaxについて10ベンチマーク中7つでコスト性能のPareto frontierを拡張、Ultra v2について8ベンチマーク中5つで首位または同率首位と主張している。(公式リリース) [1]

- **性能面では非常に有望だが、「公開初日のベンダー公表値」であることを強く割り引いて読む必要がある。** Ultra v2はChartographyで48.3、Claude Opus 5で27.3、Claude Fable 5で29.5とSakanaは報告し、DeepSWEでは74.3としている。一方、2026年9月11日時点の本調査では、LlamaIndex、HELM、LMPerf等におけるFugu Max／Ultra v2の独立した掲載・再現結果を確認できなかった。さらにTerminal-Bench系では、同じ基盤モデルでもエージェント・ハーネスの違いだけで大幅にスコアが動くことが独立研究で示されているため、「モデル性能」と「オーケストレーション／実行環境性能」を分離して評価すべきである。 [2]

- **技術的な最大の強みは「モデルを買う」のではなく「モデル市場全体を利用するルーターを買う」点にある。** 6月公開のFugu技術報告では、通常Fuguは軽量な選択ヘッドでワーカーモデルを高速選択し、Fugu Ultra系はConductorを基礎として、サブタスク分割、モデル割当、通信トポロジー、共有メモリを学習する設計が説明されている。ただし同技術報告はMax／Ultra v2より約3か月前のもので、**v2の正確なパラメータ数、オーケストレータのバックボーン、全ワーカープール、最大コンテキスト、実測レイテンシ、サーバーGPU要件、v2固有の学習データ量はいずれも未公開**である。(Technical Report) [3]

- **企業導入上の最大の注意点は、性能ではなくデータガバナンスと契約条件である。** 標準利用規約では、ユーザー本人を含む「personal information」の入力を禁止し、健康・金融等のセンシティブ情報を送らないよう明記している。またコンテンツをモデル学習等に利用できる条項があり、学習利用のオプトアウトは可能だが過去の学習への遡及的効果は保証されない。第三者サービス利用、日米を含む国外移転もあり、さらに可用性・稼働時間・応答時間の保証はない。したがって金融・医療・人事等で個人データを直接処理する本番導入は、**現行標準規約のままでは不適切であり、匿名化または個別の法人契約・DPA・SLAが前提**になる。(Terms、Privacy Policy) [4]

- **戦略的には「Maxをデフォルト、Ultra v2を品質エスカレーション先」にする二層構成が最も合理的である。** Maxの価格はGPT-5.6 TerraやClaude Sonnet 5と同じ入力$2だが、出力はそれぞれ$12、$10に対し$6である。一方Ultra v2の出力$30はClaude Opus 5の$25より高いため、全リクエストをUltraへ送る経済合理性は乏しい。企業は内部評価で「Maxで十分なタスク」と「Ultraで追加品質が費用を上回るタスク」を定量分類し、さらに直接OpenAI／Anthropic等へのフォールバックを維持するのが妥当である。 [5]

公式情報と技術仕様

Sakana AIの2026年9月11日公式リリースは、従来の「より大きな単一モデルを作る」というスケーリング競争ではなく、「どの問題にどの計算資源・モデルを割り当てるか」という**能力×コストの二次元最適化**をFuguの中心命題に置いている。Fugu MaxとUltra v2は別々の基盤モデルというより、同一のオーケストレーション基盤を異なる目的関数で最適化したサービスと理解するのが正確である。 6

製品の系譜もこの解釈を裏付ける。Sakanaは2026年4月にFugu beta、6月にGAとUltra v1、7月にFugu-CyberおよびClaude Codeインターフェース、8月にSakana ChatとNVIDIA連携、9月にMaxとUltra v2という短いサイクルで拡張している。8月にはNVIDIA NemotronをFuguの専門エージェントとして組み込み、両社で評価・改善する方針を正式発表しており、Maxでは実際にNemotron系を含む最大規模のオープンウェイト／専門モデル群を利用するとされる。(NVIDIA連携発表) 7

ただし、6月のTechnical Reportを**Fugu Max／Ultra v2そのものの設計書として読むのは誤り**である。同文書が説明するのはFuguおよび初代Fugu Ultraの基礎技術である。通常Fuguでは、事前学習済みLMの隠れ状態に軽量な選択ヘッドを付け、ワーカープールの各モデルに対応するlogitを直接出して選択する。テキストでルーティング判断を生成しないため、オーケストレーション自体のレイテンシ低減を狙う。一方UltraではConductorを発展させ、自然言語でサブタスク・担当モデル・通信関係を定義するエージェントワークフローを生成し、共有メモリを保持する。初代Ultraの学習設定では最大5ステップのワークフローを用い、GRPOによる強化学習が使われていた。 3

技術項目	Fugu Max	Fugu Ultra v2	評価・注意点
基本設計	Fugu共通の学習済みオーケストレーション。最大規模のオープンウェイト／専門モデルプールから、タスクを解ける「最も軽量なモデル」へ動的ルーティング。Nemotron系を含む。 7	同じ中核オーケストレーション思想を、複雑な多段推論・自律研究・フルスタック開発などの絶対品質向けに最適化。 8	Max/v2固有のルーター内部構造は未公開。6月技術報告のFugu／Ultra v1は設計系譜の参考資料。
オーケストレータのパラメータ数	**未公開**	**未公開**	複合システムなので「総パラメータ数」は選択されたワーカーにより変動し、単一モデルほど意味が明確でない。
ワーカーモデル数・全モデル名	**未公開**。Nemotron採用のみ明示。 7	**未公開**。ただしFable 5、Fable 5.1、GPT-6 Astraはv2のプールに入っていないと明示。 9	正確なルーティング先を利用者から隠す設計は知財保護には有利だが、監査・データ所在地確認には不利。
学習手法・データ	Max固有データは**未公開**	v2固有の詳細は**未公開**。学習カットオフは2026-08-28。Ultra v1報告では公開データと専門家設計E2E環境の混合、数学・科学・工学・ツール利用・対話・計画等を利用。 10	データセット名、総トークン数、著作権構成、言語比率等は公開されていない。

技術項目	Fugu Max	Fugu Ultra v2	評価・注意点
推論速度	tokens/s、TTFT、P50/P95レイテンシとも**未公開**	同左。品質優先で、系譜上はより深いオーケストレーション＝追加レイテンシという設計。 11	利用規約上も応答時間の保証なし。 12
コンテキスト	料金はコンテキスト長によらず固定。ただし**最大長は未公開**。 13	272K超で高額料金帯が設定されており、少なくとも料金設計上は272K超を想定。ただし**最大長は未公開**。 13	「料金が固定」＝無制限コンテキストではない。
ハードウェア／VRAM	APIサービスとして提供。サーバー側GPU・VRAM・推論構成は**未公開**。 14	同左	自社GPUへのオンプレミス展開・モデルウェイト配布は公開資料で確認できない。
API／SDK	OpenAI互換API。既存Fuguからはmodel指定の変更で移行可能。 15	同左。公開モデルIDとして `fugu-ultra-v2` を料金表に記載。 13	専用SDKへの移行を必須としないことが導入摩擦を大きく下げる。完全なOpenAI全API機能互換まで保証された記述ではない。
入力料金	**$2 / 1M tokens**	**$5 / 1M**、272K超は$10 / 1M	13
出力料金	**$6 / 1M tokens**	**$30 / 1M**、272K超は$45 / 1M	Maxのweb_search／web_fetchは$0.007/call。 13
キャッシュ入力	$0.25 / 1M	$0.50 / 1M、272K超$1.00	13
ライセンス／ウェイト	Fugu自体はオープンウェイト公開ではなく、APIサービス利用規約。リバースエンジニアリング、抽出・蒸留等は禁止。 16	同左	内部でオープンウェイトモデルを使うことと、Fuguそのものがオープンウェイトであることは別。
データ学習利用	コンテンツの学習・改善利用を規約上許容。オプトアウト可能だが遡及的削除は保証されない。 17	同左	本番導入時はアカウント作成直後に設定確認すべき。
セキュリティ／プライバシー	アクセス制御、従業員教育、監視、委託先監督等を公表。日米等への国外移転あり。 18	同左	本調査で確認した公開資料にはSOC 2、ISO 27001等の具体的認証、個別リージョン固定、具体的保持期間の明示を確認できず。

技術項目	Fugu Max	Fugu Ultra v2	評価・注意点
提供地域	利用規約ではEEA・英国・スイス以外。[14]	同左	製品ページは「EU/EEAでは未提供」と記載しており、範囲表現に差がある。契約判断ではより明確なTermsのEEA/UK/Switzerland除外を保守的に採用すべき。[19]

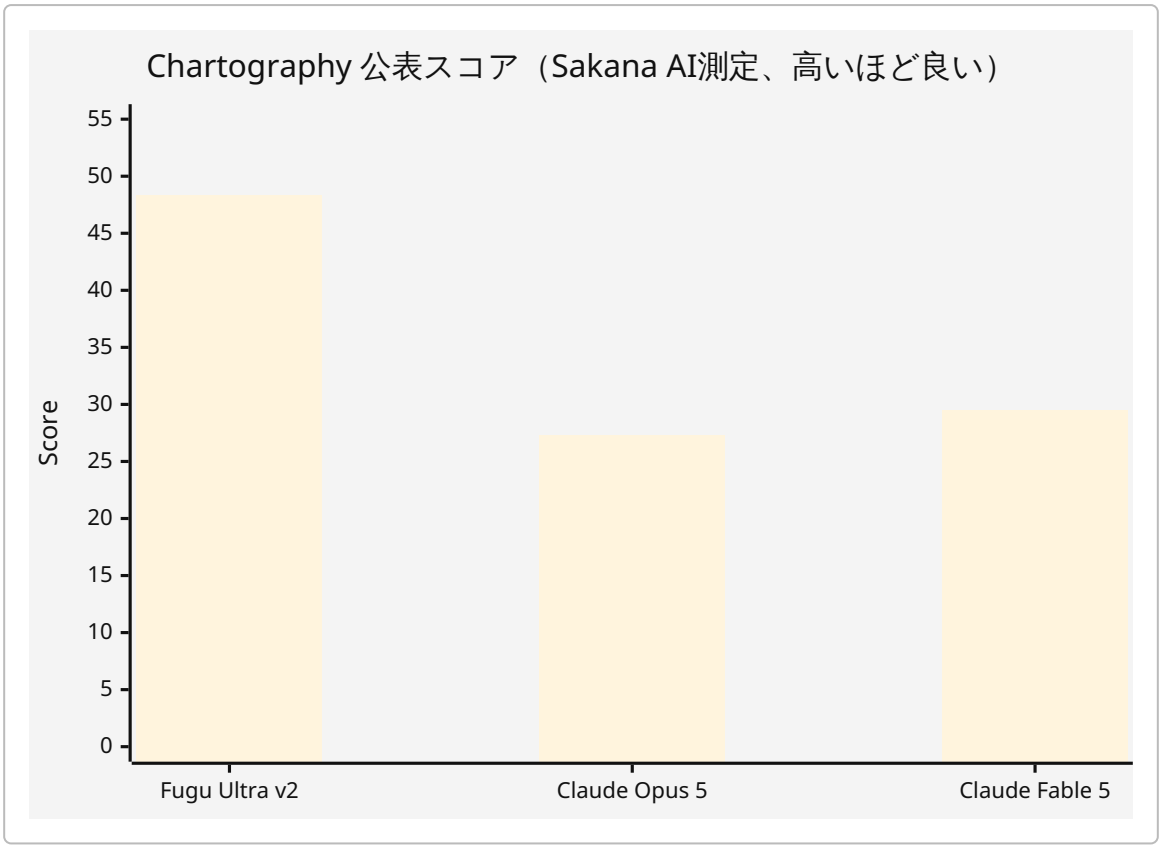
価格面で見落としやすい点は、**Ultra v2が必ずしも「安いフロンティアモデル」ではない**ことである。通常コンテキストの出力料金$30はClaude Opus 5の$25を20%上回る。Ultraの価値命題は単価の安さではなく、複数モデルを組み合わせた最終タスク成功率である。逆にMaxはSonnet 5とGPT-5.6 Terraと同じ$2の入力価格ながら、出力がそれぞれ$10、$12に対して$6であり、Sakanaのいう「出力40〜60%低価格」のうち、Sonnet比40%、Terra比50%は公表価格から確認できる。[20]

性能評価とベンチマーク

証拠の質を先に評価すると、Fugu Max／Ultra v2について現時点で最も重要な問題は「再現性」である。2026年9月11日という公開当日の調査であり、LlamaIndex、HELM、LMPerf等における両モデルの独立掲載を本調査では確認できなかった。したがって、以下のFugu v2／Max数値は原則として**Sakana AI自身が実行または集約したvendor-reported result**であり、「第三者が同じ設定で再現済みの順位」とは扱わないのが妥当である。Sakana自身の6月Technical Reportの初代モデル評価についても、比較対象の一部は各プロバイダー報告値であることが明記されている。[21]

Sakanaの公表結果では、Fugu MaxがTerminal Bench 2.1、GPQA Diamond、AA-LCR、GDP.pdf、AutomationBench、SWEFishの6ベンチマークでoverall best、10件中7件でコスト性能Pareto frontierを形成するとしている。なおSWEFishはSakana自身のコーディング課題・ユースケースを反映した**内部ベンチマーク**であり、外部公開ベンチマークと証拠価値を同一視すべきではない。[22]

Ultra v2については、GDP.pdf、Chartography、SWEFish、DeepSWE、Toolathonの5/8で首位または同率首位、7/8でトップ2というのが公式主張である。Chartographyは48.3で、Sakanaが同じ図で示すClaude Opus 5の27.3を21.0ポイント、Fable 5の29.5を18.8ポイント上回る。相対値ではそれぞれ約77%、約64%の差であり、これが第三者再現されれば特に視覚・構造化データ推論で大きな差といえる。DeepSWEは74.3とされている。[9]



ベンチマークそのものの質にも差がある。DeepSWEは113件の新規長時間ソフトウェア工学タスクを91の現役OSSリポジトリ、5言語にまたがって設計し、既存GitHub issue由来の訓練データ汚染を減らすことを狙った2026年の比較的新しいベンチマークである。手書き検証器によって任意の正しい解法を評価する設計になっており、旧来のSWE-Bench系が抱える「既知の修正を記憶しているだけではないか」という問題への対策として有力である。(DeepSWE paper) [23]

GDP.pdfも比較的意味のある実務系評価である。100件のprofessional PDF課題を10分野の実務家が作成し、単なるOCRではなく表・グラフ・脚注・クロスリファレンス・空間推論・根拠不足時の棄却などを複合的に評価する。原論文では当時の最良モデルでも100項目中15%しか完全通過しなかったと報告されており、天井効果の少ない難しい評価になっている。Fugu Ultra v2がここで首位グループに入るという主張は、金融・医療・製造等の文書処理にとって特に意味があるが、Fugu結果自体はSakana側での再評価である点は変わらない。(GDP.pdf paper) [24]

一方でTerminal-Bench 2.1のようなエージェント評価は、**基盤モデルだけではなく実行ハーネスに非常に敏感**である。2026年8月のStateM研究では、GPT-5.5 xhighの参考値83.1%を状態管理ハーネスによって92.1%へ引き上げ、GPT-5.6 Sol xhighで95.3% raw accuracyに到達した。つまりFuguのような「オーケストレーション込みシステム」と単体LLMを比較する際には、これは欠点というよりFuguの狙いそのものではあるものの、**「Fuguのルーティング学習が優れている」のか「実行環境・ツール設計が優れている」のかを公開スコアだけから分解できない**。(StateM paper) [25]

多言語対応については、初代Technical Reportに古典日本語の仮名書状の読順復元という専門家作成の25ページ評価がある一方、この課題自体に公開標準ベンチマークは存在しないと著者ら自身が説明している。Ultra v2の発表には、MGSMやMMLU多言語版等に相当する包括的な標準多言語スコアは示されていない。したがって「日本企業発だから日本語で競合を大幅に上回る」といった推論は現時点では証拠不足である。 [21]

長文性能についても同様で、Ultra v2に「272K超」の料金体系があることから長コンテキスト入力を明確に想定していることは分かるものの、最大ウィンドウ、Needle-in-a-Haystack、長文QA、長文での精度劣化率は未公開である。Maxも「context lengthに関係なく固定料金」としているが、最大コンテキスト自体は公表していない。したがってGPT-5.6 Terraの公開1,050,000トークンやMeta Llama 4 Scoutの10Mと、仕様値として直接比較することはまだできない。 [26]

安全性についてはさらに情報が少ない。9月11日のリリースは能力・コスト評価が中心で、専用の有害性、jailbreak、prompt injection、CBRN、サイバー悪用、バイアス等の定量安全ベンチマークを提示していない。6月Technical Reportでも安全性専用評価の詳細なセクションを確認できない。この点は、Llama 4がLlama Guard、Prompt Guard、CyberSecEvalやred-team手法を具体的に公開していることと比べると、**Fuguの公開安全性エビデンスは現時点では薄い**。 [27]

以上から、現段階の性能評価は次のように整理できる。

評価軸	現時点の判断	確信度
複雑なエージェント／コーディング	Ultra v2は極めて有望。DeepSWE 74.3、複数ベンチで上位。 [28]	**中**：数値は未独立再現
コスト性能	Maxは公表価格上、Sonnet 5／Terraより出力単価が明確に安く、公式評価では10件中7件でPareto frontier。 [20]	**中〜高**：価格は検証可能、性能部分は未再現
PDF・視覚構造化推論	Ultra v2のChartography/GDP.pdfは注目に値する。 [29]	**中**
日本語・多言語	系譜上の実験はあるが、v2標準多言語評価不足	**低**
長文	272K超利用を想定するが最大長・長文精度未公開。 [13]	**低**
安全性	定量的な公開安全評価不足	**低**
レイテンシ	数値未公開、SLAなし。 [12]	**低**

設計思想・差分・市場インパクト

Fugu MaxとUltra v2の違いは、単純な「小型版／大型版」ではない。**Maxは「タスクごとの限界費用を最小化するルーター」、Ultra v2は「複数エージェントの協働に計算を追加して最終成功率を最大化するオーケストレーター」**と捉えると分かりやすい。Maxでは能力の低いモデルでも十分な課題に高価なモデルを使わないこと自体が知能の一部とされ、Ultraでは逆に複雑な問題分解・検証・ツール利用へ追加計算を使う。 [30]

したがってトレードオフは次のようになる。Maxは高頻度・大量処理で最も魅力的だが、ルーターがタスク難易度を過小評価して安価なワーカーを選んだ場合には品質損失が起こり得る。Ultraはその逆で、複数エージェントによる分解・検証によって難問の成功率を高められる一方、追加の呼び出し・共有状態管理によってレイテンシと内部計算量が大きくなりやすい。後者の性質は初代UltraのConductor設計にも明示されている。 [31]

競合比較では、Fuguは単一モデルを正面から代替するだけではなく、OpenAI、Anthropic、Meta、Mistral、NVIDIA等を「部品化」できる点が根本的に異なる。これは潜在的には非常に強いポジションだが、同時にAmazon Bedrock、Google Vertex AI、Azure、各種AI Gateway、あるいはモデルプロバイダー自身の自動ルーティング機能と競合する「メタレイヤー」でもある。

サービス／モデル	公開標準価格・1M tokens	主な構造・特徴	Fuguとの戦略的比較
Fugu Max	**入力$2 / 出力$6** [13]	複数のオープン／専門モデルを動的ルーティング。Nemotron系含む。	単一モデルより「価格×最終タスク成功率」を最適化することが主眼。
Fugu Ultra v2	**$5 / $30**。272K超は$10 / $45。 [13]	複雑な多段タスク向けオーケストレーション	出力単価そのものは安くない。難問成功率へのプレミアム。
OpenAI GPT-5.6 Terra	**$2 / $12**、キャッシュ$0.20。 [32]	1.05M context、128K max output、function calling・MCP・web/file search等を公開。	Maxと入力単価同額、出力はMaxが50%安い。OpenAIはツール・エージェント基盤の垂直統合が強い。
Anthropic Claude Sonnet 5	**$2 / $10** [33]	汎用／コーディング、Anthropic APIに加えてBedrock・Google Cloud配布。 [33]	Maxは出力40%安い。Anthropicはクラウド調達・データルーティング選択で優位。
Anthropic Claude Opus 5	**$5 / $25** [34]	高能力単体モデル。Fast modeは$10/$50。	Ultraは入力同額だが標準出力は20%高い。Fuguの価値は単体価格でなくオーケストレーション品質に依存。
Mistral Large	**$0.5 / $1.5**。batchは50%引き。 [35]	非常に低単価な単体API	純粋なtoken costではMaxより大幅に安い。Fuguは高い成功率で再試行・人手を削減できるかが勝負。
Meta Llama 4 Scout／Maverick	中央集権的API定価ではなく、自己ホスト／提供事業者コストに依存	2025年の公式open-weight参照例。Scout 17B active/109B total、10M context、Int4で単一H100、Maverick 17B active/400B total。 [36]	自社運用・完全制御を求める組織にはMeta型が優位。Fuguはインフラ管理を捨てて複数モデル活用を買う選択。
NVIDIA Nemotron	モデル／提供形態による	オープンモデル群。SakanaとFugu内専門エージェントとして共同最適化。 [37]	競合であると同時にFuguの供給部品。これがFuguの「モデル市場の上位レイヤー」という性質を象徴する。

この表から重要なのは、**Fugu Maxが市場で最安ではない**という点である。Mistral等の低価格モデルや自社ホストのオープンウェイトは、純粋なトークン価格ならさらに安い。Maxの経済価値が成立するためには、「安価なモデルを利用者自身で選び分ける運用コスト」や「弱いモデルで失敗して再試行するコスト」を、Sakanaの学習済みルーターが十分削減できなければならない。これはSakanaが今後第三者評価で証明すべき核心的なunit economicsである。 [38]

逆にFuguの強力な競争優位になり得るのは、**モデル更新をユーザーアプリケーションから切り離すこと**である。OpenAI互換APIを維持したまま内部ワーカープールを入れ替えられるなら、新しいオープンモデル、NVIDIA Nemotron、あるいは将来の専門モデルの改善をアプリ側改修なしに取り込める。これはモデル陳腐化の速い市場で大きな価値を持つ。 [39]

ただしこれは「ロックインの消滅」ではなく**ロックイン場所の移動**でもある。特定のOpenAI／Anthropicモデルへの依存は下げられる一方、どのモデルがどの入力を処理したかをユーザーが完全には把握できないオーケストレーション層自体への依存が増える。さらに規約はService Configurationのリバースエンジニアリングやモデル抽出を禁じているため、ルーティング知識を自社へ持ち出すことはできない。 16

企業市場では、この構造は**Sakanaにとって機会と弱点を同時に生む**。SaaSベンダーや一般企業にとってOpenAI互換は統合を容易にする一方、規制業種では「誰がサブプロセッサなのか」「どの国で処理されたか」「特定モデルを排除できるか」「モデルバージョンを固定できるか」が重要になる。標準Fuguでは一部モデルの除外という思想が示されているが、Max／Ultraは固定プール型であるため、透明性よりSakana側の最適化を優先した商品設計とみられる。 40

スタートアップにとってはさらに契約上の注意がある。利用規約はFuguを利用してServiceと競合する製品・サービス、特にAI service/APIを開発・提供する行為を禁止している。一般業務SaaSへのAI機能組み込みまで直ちに禁止されるとは限らないが、**AI-nativeなモデルゲートウェイ、AI API、汎用AIアシスタント事業者は契約上の競合該当性を事前確認すべき**である。 41

法規制・倫理・安全性

日本では「人工知能関連技術の研究開発及び活用の推進に関する法律」、いわゆるAI法が2025年9月1日に全面施行され、政府のAI戦略・「信頼できるAI」形成の制度枠組みが整備されている。もっとも、企業のFugu導入実務ではAI法だけを見ればよいわけではなく、個人情報、知的財産、消費者保護、金融・医療等の業法上の義務を、具体的なユースケースごとに評価する必要がある。日本政府はAI法の全面施行を2025年9月1日と公式に説明している。 42

Fuguについて最も即時性の高い規制上の問題は、**標準Termsと個人データ処理の不整合リスクを利用企業自身が管理しなければならない点**である。Privacy Policy上、Sakana自身はアカウント情報、プロンプト、アップロード内容、出力、IP、ログ等を収集し得る。一方、ユーザーがAPIのInputに本人または第三者のpersonal informationを入れることは禁止され、健康・金融等のセンシティブ情報を提出しないよう明示されている。したがって「個人情報を含む融資審査」「患者カルテをそのまま解析」「学生個人データを含む評価」のような用途は、現行標準契約では避けるべきである。 43

データ主権にも注意が必要である。Privacy Policyは、基盤インフラ等の認定サービスプロバイダーへの情報開示、および日本・米国等を含む国外へのデータ移転の可能性を明示している。保持期間は一律の日数ではなく「目的に合理的に必要な期間」とされ、具体的なスケジュールは別途文書化され得るとしている。したがって、厳格なデータ所在地・削除期限を必要とする企業は標準Privacy Policyだけで要件充足と判断すべきではない。 44

入力データの学習利用についても、Termsはコンテンツを Sakanaのモデル訓練・改善・評価等に利用できるとし、場合によって人間のレビュアーが処理できるとしている。オプトアウトは可能だが、すでに学習に使われたデータの効果を後から完全に取り消せない場合がある。企業利用では、**本番データ投入前にtraining opt-outを有効化し、さらに法人契約上で「no training」を明文化する**のが安全側の運用である。 45

EUでは2026年9月時点でAI Actの実施フェーズが大きく進んでいる。欧州委員会によればAI Actは2026年8月2日に原則適用となり、GPAIモデル関連義務は2025年8月から既に適用、AI Officeは2026年8月以降GPAIに対する文書要求・評価・是正・制裁権限を行使できる。GPAIには透明性・著作権等の義務があり、systemic riskを伴う場合にはリスク評価・緩和が要求される。(European Commission AI Act) 46

Sakanaは現在EU/EEAでFuguを提供しておらず、TermsではさらにUKとSwitzerlandもSupported Regionsから除外している。このため直近の直接的EU市場リスクは限定されるが、これは同時にOpenAI／Anthropic等に対する市場アクセス上の弱点でもある。将来EUに参入する場合、Fuguのような複数モデル・複数事業者を

組み合わせるシステムでは、GPAIのprovider/deployer責任、サブプロセッサ、著作権、モデル文書、リスク評価をどのレイヤーが負担するかを明確化する必要がある。 47

米国ではEUのような単一AI法よりも、連邦・州・業界別ルールとリスク管理フレームワークを組み合わせる運用が重要になる。NIST AI RMFは自主的利用を前提にAIのtrustworthinessを設計・開発・利用・評価へ組み込む枠組みであり、Fuguのような動的オーケストレーションではモデル単体ではなく「router＋worker＋tools＋data＋human review」のシステム全体を評価対象にすべきである。(NIST AI RMF) 48

主要リスクと緩和策をまとめると次の通りである。

リスク	Fugu特有の論点	推奨緩和策
誤情報・幻覚	複数エージェントの合意が必ずしも真実を意味しない。規約自身も出力の正確性を保証せず、独立確認を要求。 49	RAGによる一次資料固定、引用必須化、重要判断はhuman-in-the-loop、abstention評価。
Prompt injection／tool abuse	エージェント・ツール・共有メモリが増えるほど攻撃面も増える。Ultra系は複数エージェントがfunction call可能な設計系譜。 50	ツールallowlist、最小権限、sandbox、書込操作承認、ネットワークegress制限。
個人情報・秘密情報漏洩	第三者サービス、国外移転、固定／非公開のワーカープール。 51	匿名化、DLP、no-training契約、サブプロセッサー一覧、データ所在地、保持期限を契約化。
供給網	複数モデル利用で単一ベンダー停止には強くなる一方、依存先総数が増える。 9	direct-model fallback、障害訓練、モデルプール変更通知、version pinningを要求。
モデル変更による挙動ドリフト	内部プールの入替がFuguの強みである反面、同じAPI名でも挙動が変わり得る。	自社golden setによる継続Evals、変更前後の回帰試験、固定snapshot機能をSakanaへ要求。
可用性	Terms上、uptime・response time保証なし。 12	本番では個別SLA、マルチプロバイダfallback、timeout/circuit breaker。
法的責任	専門的な法律・医療・投資等の助言としての利用を制限し、有資格者関与を要求。 52	AIを意思決定者ではなく支援者に限定し、最終承認者を明確化。

ビジネスインパクトと導入シナリオ

以下の概算では比較のため、**1リクエスト＝入力5,000 token＋出力1,000 token、Ultraは272K以下、web検索等の追加料金なし、為替は実勢値ではなく計算用に1ドル＝150円**と仮定する。この条件では1リクエスト当たりFugu Maxは `0.005×$2 + 0.001×$6 = $0.016`、Ultra v2は `0.005×$5 + 0.001×$30 = $0.055` となる。価格前提はSakana公式料金による。 13

業界	有望な導入シナリオ	推奨モデル	例示月間量とAPI費用概算	最大の導入条件
金融	公開決算・市場資料分析、規程比較、匿名化済みリスク調査、コード生成	Maxを標準、複雑な投資調査・PDF横断分析だけUltra	2万件/月：Max **$320≒4.8万円**、Ultra **$1,100≒16.5万円**	個人金融情報のInputは禁止。投資助言は有資格者／責任者レビューが必要。 [53]
医療・ライフサイエンス	公開論文レビュー、ガイドライン比較、匿名化研究データ、研究コード	精密な文献統合はUltra、日常探索はMax	5千件/月：Max **$80≒1.2万円**、Ultra **$275≒4.1万円**	患者PII/health dataを標準Termsで投入しない。診断・助言は専門家関与必須。 [54]
製造	マニュアルQA、設備文書解析、設計レビュー、ソフトウェア／PLC周辺コード支援	大量文書QAはMax、複雑な故障解析はUltra	3万件/月：Max **$480≒7.2万円**、Ultra **$1,650≒24.8万円**	機密設計情報・営業秘密の扱いを法人契約で明確化。
教育	教材生成、質問応答、教員向け授業設計、採点補助	原則Max、研究・高度STEMだけUltra	5万件/月：Max **$800≒12万円**、Ultra **$2,750≒41.3万円**	標準サービスは18歳以上。学生PIIや自動的な高リスク評価用途に注意。 [55]
コンテンツ制作	リサーチ、構成、翻訳補助、脚本・コピー、制作ツール連携	高頻度生成はMax、難しい調査・整合性検証はUltra	4万件/月：Max **$640≒9.6万円**、Ultra **$2,200≒33万円**	著作権・出典確認、AI生成であることを意図的に隠して第三者を欺く利用を避ける。 [56]

この試算から分かるのは、一般的な企業AI導入では**APIトークン代そのものがROIの主要因にならない可能性が高い**ことである。例えば、知識労働者20人、年間220営業日、完全負担人件費6,000円/時、Fuguによって1人1日30分削減できると仮定すると、年間の理論的時間価値は

`20人 × 0.5時間 × 220日 × 6,000円 = 1,320万円`

となる。これに対して初年度のシステム統合、セキュリティ、評価、教育、運用を合計600万円と仮定すれば、初年度ROIは

`(1,320万円 − 600万円) / 600万円 = 120%`

となる。この数値は**Fuguの実績値ではなく導入判断用の感度分析**である。30分ではなく15分/日しか削減できなければ便益は660万円、ROIは10%まで下がる。600万円の初年度投資を回収する損益分岐点は約13.6分/人/日の時間削減である。

したがってROIを左右するのは「100万トークンが数ドル安いか」より、**回答成功率、再試行率、人間レビュー時間、待ち時間、障害率、誤回答による損失**である。MaxがSonnet 5やTerraよりトークン単価を下げても、レビュー時間が増えれば総コストは上がる。逆にUltraが高価でも、1回で難しいソフトウェア修正や専門調査を完了できれば総所有コストは下がり得る。これこそSakanaのPareto-frontier主張を企業内部データで再検証すべき理由である。 [57]

同一の5K入力／1K出力を10万回/月実行した単純なtoken-only比較では、Fugu Maxが約**$1,600/月**、Fugu Ultra v2が**$5,500/月**、GPT-5.6 Terraが**$2,200/月**、Claude Sonnet 5が**$2,000/月**、Claude Opus 5が**$5,000/月**、Mistral Largeが**$400/月**になる。したがってMaxは「高性能帯の費用圧縮」には強いが絶対最安ではなく、UltraはOpusよりむしろ高価になる入力・出力比もある。 58

企業の実装パターンとしては、すべてをUltraへ送るのではなく、**①ルールベース／小型モデルで明らかな簡易課題を処理、②Fugu Maxを通常経路、③品質判定やタスク分類で難問のみUltraへ昇格、④障害時にOpenAI/Anthropic等へdirect fallback**という多段構成が、コスト・可用性・ベンダーロックインを最も合理的に管理できる。この設計はFugu自身の「適切な計算資源を適切な課題に割り当てる」という思想を、企業側でも一段上に適用するものである。 59

今後の展望・結論・推奨

短期・今後6〜12か月。 最も重要なのは、Fugu Max／Ultra v2の独立再現性が確立するかである。DeepSWE、Terminal-Bench、GDP.pdf等の外部leaderboardや第三者エージェント評価で同程度の優位を維持できれば、「巨大単一モデルではなく学習済みオーケストレーションでfrontier performanceを得る」というSakanaの主張はかなり強くなる。逆に順位がハーネスやSakana固有設定に大きく依存するなら、市場評価は「優秀なモデルゲートウェイ」に収斂する可能性がある。Terminal-Benchでハーネス変更が10ポイント規模の差を生み得ることは、この検証の重要性を示している。 60

短期ではNVIDIAとの連携も重要である。NemotronのようなオープンモデルをFugu内で共同最適化し、Sakanaが新しいモデルを継続的にプールへ取り込めるなら、単独モデル会社より速く「市場全体の技術進歩」を製品品質へ反映できる。これは日本発AI企業として、数兆パラメータ級の基盤モデル訓練費用で米国大手と正面衝突せずに付加価値を取る戦略として合理的である。 61

中期・1〜3年。 オーケストレーションそのものは急速にコモディティ化すると予想する。OpenAI自身もmulti-agent、tools、MCP、computer use等を垂直統合しており、AnthropicはBedrock／Google Cloudまで流通網を広げ、Meta・Mistral等はopen-weight／低価格モデルによって部品供給側を強くする。したがってSakanaの持続的な堀は単純な「複数モデルを呼ぶ機能」ではなく、**どのタスクでどのモデル・トポロジーが成功するかという独自の学習データ、評価インフラ、ドメイン特化オーケストレーター、企業ガバナンス機能**になると考えられる。 62

この期間にSakanaが企業市場で伸びる条件は、性能以上に「調達可能なサービス」へ成熟できるかである。特にSOC 2／ISO系認証、サブプロセッサー一覧、具体的retention、リージョン固定、no-training標準法人プラン、監査ログ、SLA、バージョン固定、ルーティング先の監査可能なメタデータが重要になる。現行公開Privacy Policyは一般的なアクセス制御等を示しているものの、規制産業が大規模本番導入するために通常要求するこれらの詳細情報を十分には提供していない。 63

長期・3〜5年。 Sakanaの仮説が正しければ、AI市場の価値の一部は「最高の一つのモデルを所有する企業」から、「多数のモデル、専門エージェント、ツール、計算資源を最適に編成する企業」へ移る。この世界ではOpenAI、Anthropic、Meta、Mistral、NVIDIA等のモデルが相互排他的な最終製品ではなく、時に競合し時に同一ワークフローの部品になる。Fuguはその“AI operating layer”を狙っていると解釈できる。Sakana自身が単一モデル依存を避け、Nemotronを含むオープンモデルを組み込む方針はこのシナリオと整合する。 64

ただし長期的な逆シナリオもある。基盤モデル自体が十分安価・高速になり、一つのモデルが内部でMixture-of-Experts、subagents、tools、self-verificationを高度に行えるようになれば、外部オーケストレーターが生む追加価値は縮小する可能性がある。またモデルプロバイダー自身が最適ルーティングをAPI内部に実装すれば、独立オーケストレーターには価格交渉力やデータ透明性で厳しい競争が生じる。Fuguの長期優位は「複数モデルを使えること」ではなく、**単一プロバイダーでは再現しにくい異質なモデル間の協調から、継**

続的に測定可能な超過価値を生めるかにかかっている。これは現時点では合理的な仮説だが、まだ実証済みの結論ではない。 65

結論。 Fugu MaxとFugu Ultra v2は、日本発の生成AI製品として単なる「国産LLM」の延長線上にはない。Sakana AIは巨大モデル訓練の規模競争から意図的に離れ、世界中のモデルを動的に組み合わせる**collective intelligence / model orchestration layer**に競争軸を移している。Fugu Maxはこの構想の経済性を、Ultra v2は能力上限を示す製品であり、二つを同日に投入したこと自体が「速度か品質かを一つのモデルサイズで選ばせる」のではなく、「ルーティング戦略を切り替える」という製品思想を表している。 59

同時に、2026年9月11日時点では**技術透明性とエンタープライズ成熟度が性能主張に追いついていない**。パラメータ数・全ワーカープール・最大context・レイテンシ分布・v2固有学習データ・安全性評価が未公開であり、Fugu Max／Ultra v2の第三者ベンチマークも公開初日ゆえ不足している。さらに標準TermsにおけるPII入力禁止、学習利用、国外移転、SLA不在は、金融・医療・大企業導入で看過できない。 66

導入企業への推奨は明確である。 まずFugu Maxを一般業務の標準候補として、自社の実データからPII・機密情報を除いた評価セットで、GPT-5.6 Terra、Claude Sonnet 5、Mistral等との品質・レイテンシ・人間レビュー時間を比較する。その上で、Maxが失敗する高価値タスクのみUltra v2へエスカレーションする。規制業種は標準Termsのまま本番個人データを投入せず、no-training、DPA、subprocessor、data residency、retention、SLA、security certification、incident notificationを個別契約で確定してから進むべきである。 67

Sakana AIへの戦略的推奨は、第三者leaderboardへの早期参加に加えて、Fugu Max／Ultra v2専用model/system cardを公開し、P50/P95レイテンシ、最大context、標準多言語評価、日本語評価、安全・jailbreak・prompt-injection評価、サブプロセッサとデータ所在地、モデル更新ポリシーを明示することである。特にオーケストレーションの知財を開示せずとも、「どの事業者へデータが渡り得るか」「どの安全クラスのモデルが選ばれたか」を監査可能な形で返す仕組みは、規制産業で大きな差別化になり得る。現状のFuguは**技術的には非常に興味深く、経済的にもMaxは有力だが、企業基幹用途については“性能を買う前にガバナンスを確認する”べき製品**というのが本調査の最終評価である。 68

主要一次資料: Sakana AI 公式リリース 6 ／ Sakana Fugu製品ページ・料金 69 ／ Sakana Fugu Technical Report 21 ／ Sakana AI Privacy Policy 55 ／ Sakana AI Terms of Service 14 ／ Sakana–NVIDIA連携 37 ／ DeepSWE 23 ／ GDP.pdf 70 ／ EU AI Act公式解説 46 ／ NIST AI RMF 48 ／ OpenAI GPT-5.6 Terra公式仕様 32 ／ Anthropic API Pricing 71 ／ Mistral Pricing 35 ／ Meta Llama 4公式技術発表 36

---

1 2 6 7 8 9 10 15 20 22 27 28 29 30 38 39 57 59 64 68 Introducing Fugu Max and Fugu Ultra v2: Orchestrating the Pareto Frontier
https://sakana.ai/fugu-max-release/

3 11 21 31 50 65 Sakana Fugu Technical Report
https://arxiv.org/pdf/2606.21228

4 12 14 16 17 41 45 49 52 53 54 56 66 Terms of Service
https://console.sakana.ai/terms-of-service

5 13 19 26 40 47 58 67 69 Sakana Fugu — Multi-Agent System as a Model
https://sakana.ai/fugu/

18 43 44 51 55 63 Privacy Policy
https://console.sakana.ai/privacy-policy

23 DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks
https://arxiv.org/abs/2607.07946?utm_source=chatgpt.com

24 70 GDP.pdf: Benchmarking Grounded Multimodal Reasoning over Professional PDF Documents
https://arxiv.org/abs/2607.11192?utm_source=chatgpt.com

25 60 StateM: Reaching 95.3% Raw Accuracy, or a \$15 Frontier Run, on Terminal-Bench 2.1 via Harness Scaling
https://arxiv.org/abs/2608.15089?utm_source=chatgpt.com

32 62 GPT-5.6 Terra Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-5.6-terra

33 34 71 Pricing - Claude Platform Docs
https://docs.anthropic.com/en/docs/about-claude/pricing

35 Pricing | Mistral
https://mistral.ai/pricing

36 The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation
https://ai.meta.com/blog/llama-4-multimodal-intelligence/

37 61 Sakana AI Teams With NVIDIA to Advance Open Model Innovation from Japan
https://sakana.ai/nvidia-open-model-innovation/

42 城内大臣記者会見（令和7年9月2日）
https://www.gov-online.go.jp/press_conferences/minister_of_state/202509/video-301519.html?utm_source=chatgpt.com

46 AI Act | Shaping Europe’s digital future
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

48 AI Risk Management Framework | NIST
https://www.nist.gov/itl/ai-risk-management-framework