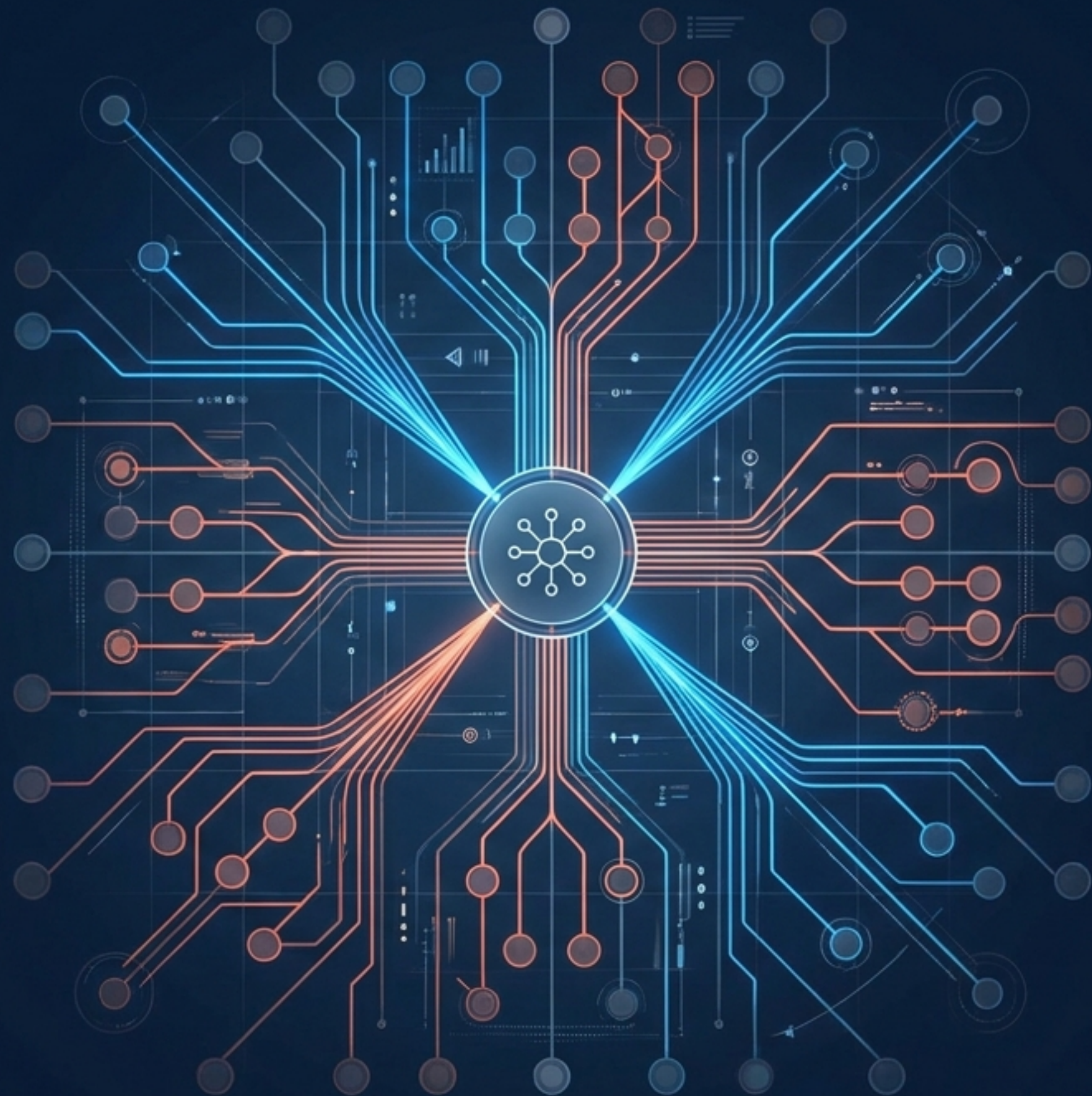


Sakana AI 「Fugu Max / Ultra v2」 徹底分析

巨大単一モデルから「オーケストレーション」へ。
エンタープライズ導入における性能、経済性、
そしてガバナンスの全貌

2026年9月 | Strategic Due Diligence Report



Executive Summary



正体 (Identity) - メタレイヤーの誕生

単体の巨大基盤モデルではなく、Nemotron等の外部・オープンウェイト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**」。



価値 (Value) - 2つの最適化目標

- **Max**: コスト効率を極限まで高める（限界費用最小化）。
- **Ultra v2**: 複雑な多段タスクの絶対性能を追求（最終成功率最大化）。

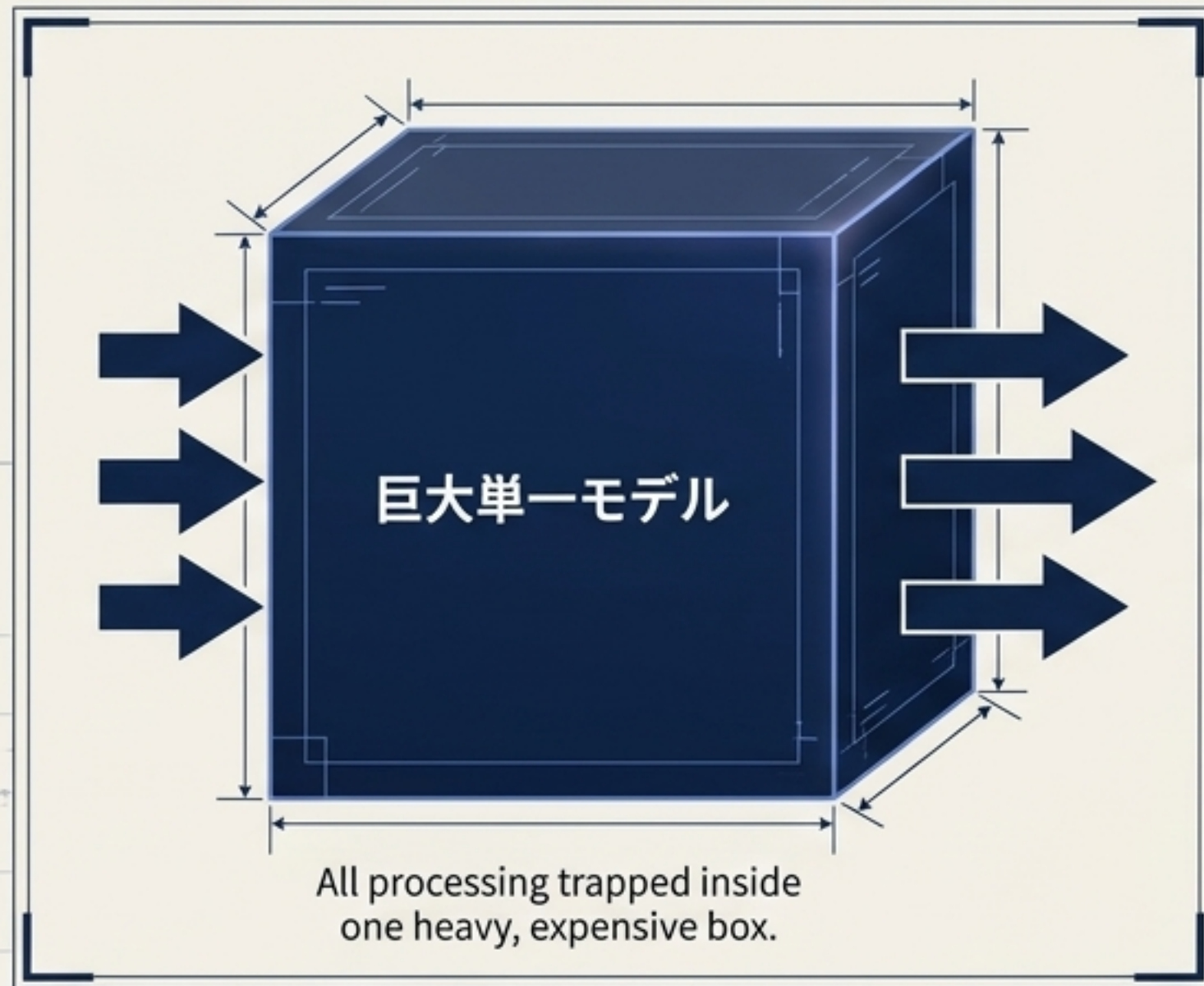


障壁 (Barrier) - ガバナンスの壁

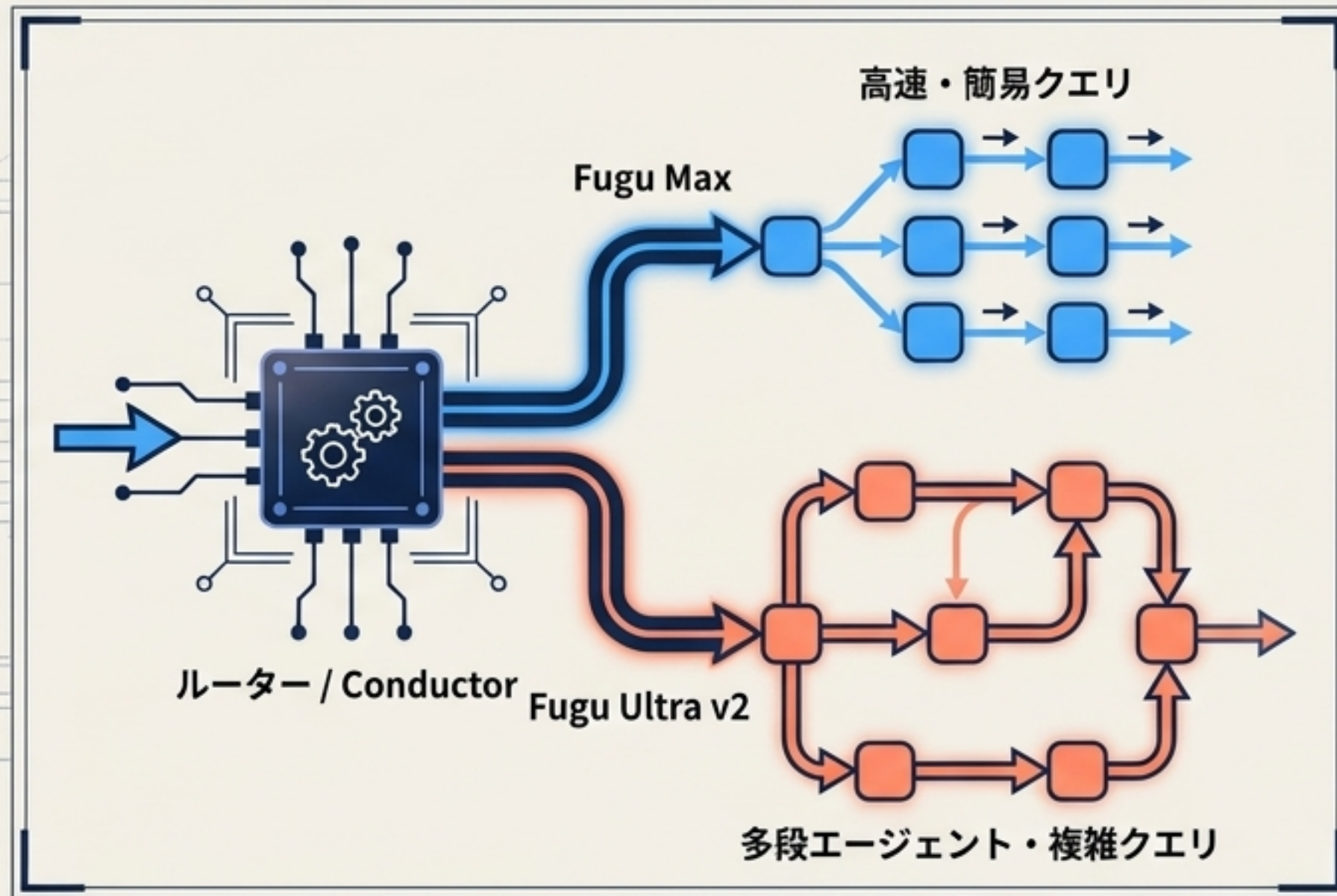
最大の導入障壁は性能ではなく「**標準規約**」。**個人情報 (PII) 入力禁止**、データ学習への利用許容、SLA（稼働時間）不在が本番展開のネックに。

Paradox Shift: モデルの「購入」から市場全体の「ルーティング」へ

従来型スケーリング (Traditional)



Fugu オーケストレーション (Orchestration)

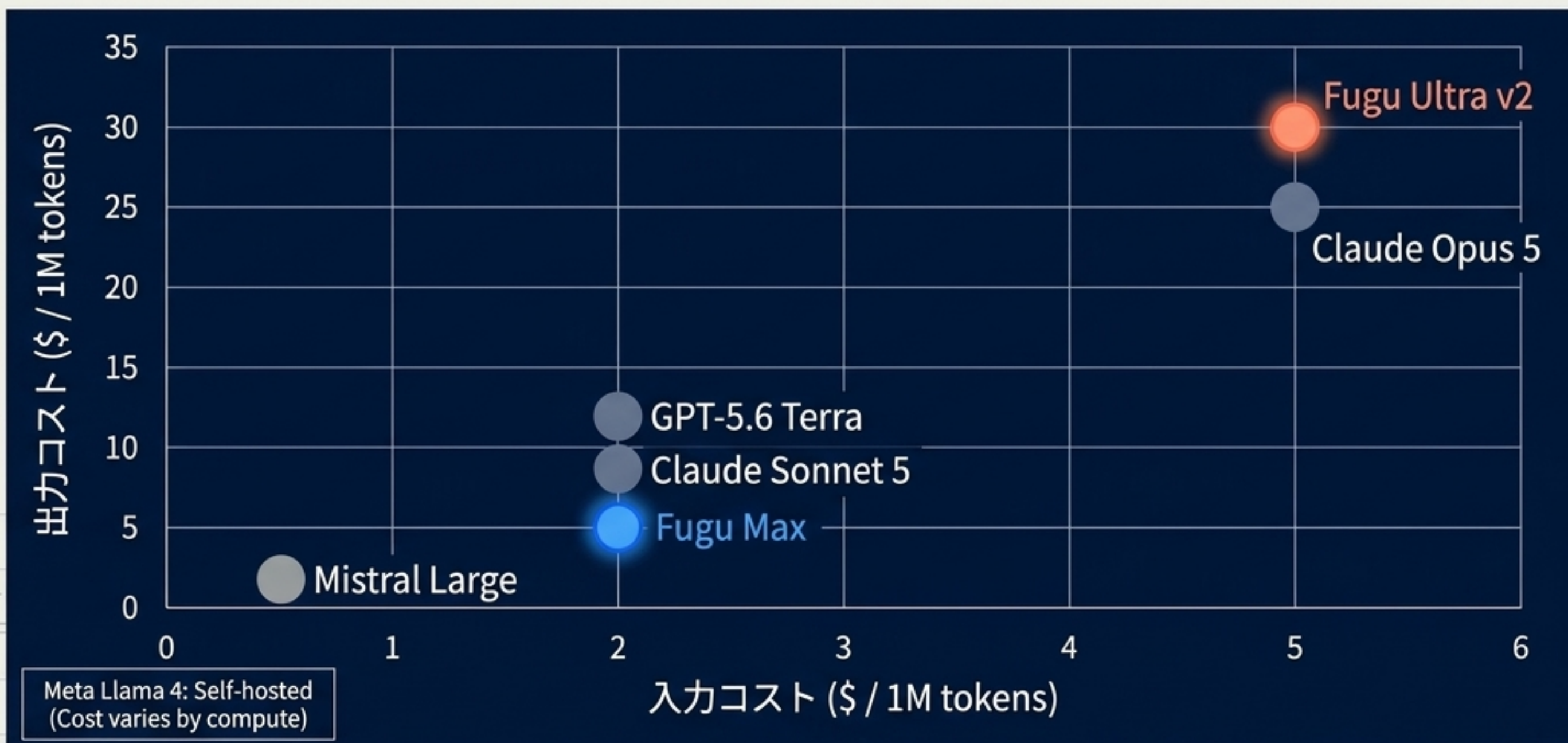


Sakana AIの命題は「モデルの巨大化」ではなく、
「どの問題にどの計算資源を割り当てるか」という能力×コストの二次元最適化である。

The Dual Engine: Max vs. Ultra v2

	Fugu Max	Fugu Ultra v2
最適化目標	タスクごとの限界費用を最小化 (Pareto frontier拡張)	追加計算を投じて 最終成功率を最大化
アーキテクチャ	軽量選択ヘッドによる 単発・高速ルーティング	Conductorによるサブタスク分割、 エージェント協調、共有メモリ管理
料金体系 (1M tokens)	入力 \$2 / 出力 \$6	入力 \$5 / 出力 \$30 (※272K超は \$10 / \$45)
想定タスク	高頻度・大量処理、 汎用的な日常業務	複雑な多段推論、自律研究、 フルスタック開発

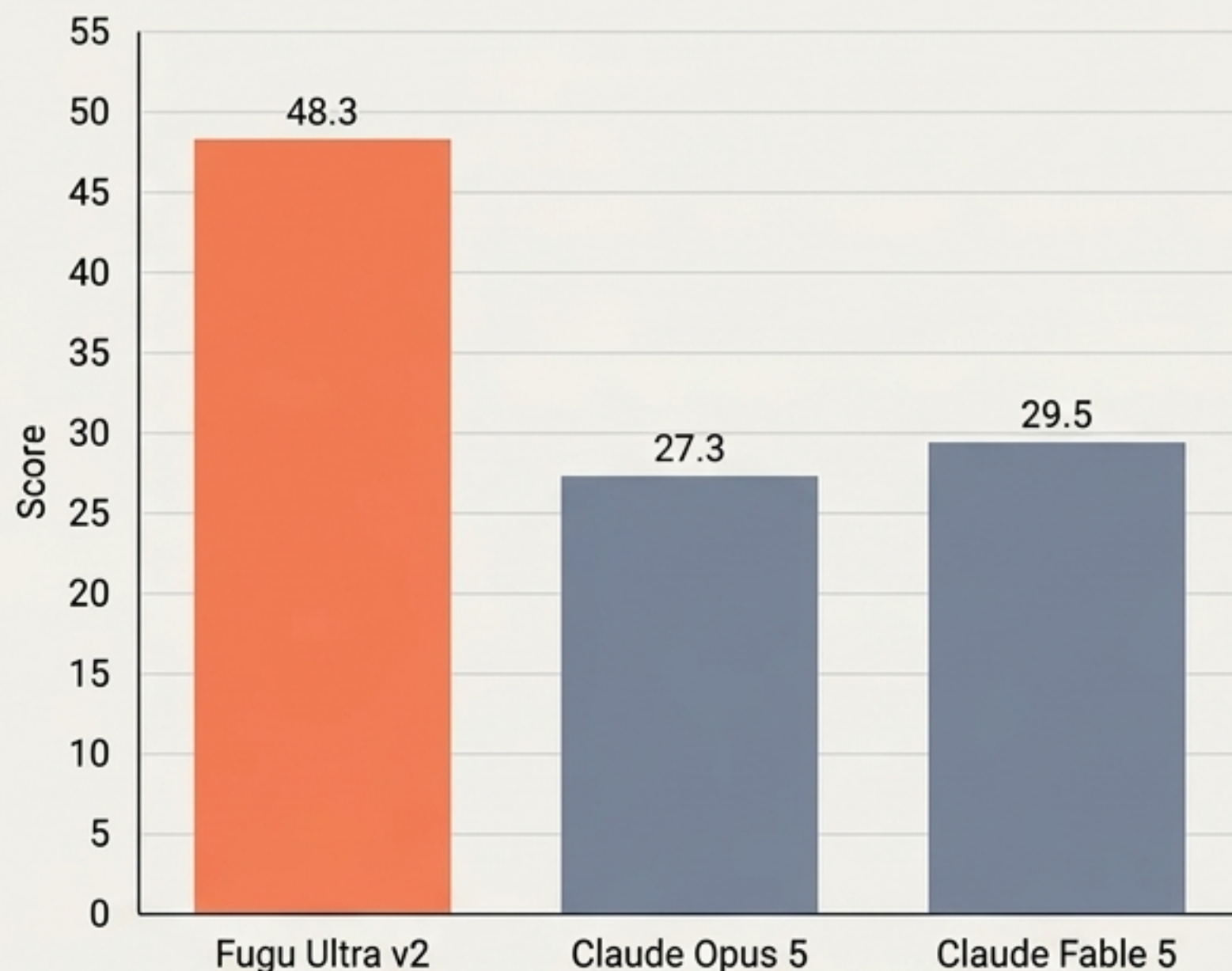
Unit Economics: トークン単価の現実とポジショニング



MaxはGPT-5.6 TerraやSonnet 5と比較して出力コストで40～50%の優位性を持つ。
しかし、Ultra v2はOpus 5の出力を20%上回る「プレミアム価格帯」であり、
全リクエストをUltraに流す経済合理性はない。

Performance Claims: 視覚・構造化データ推論におけるブレイクスルー

Chartography 公表スコア
(Sakana AI測定、高いほど良い)



ブレイクスルー

視覚・構造化データ推論（Chartography）で、Opus 5に対し相対値で約77%の圧倒的リード。



実務文書解析

GDP.pdf（10分野のプロフェッショナルPDF・空間推論）で首位グループ入りを主張。



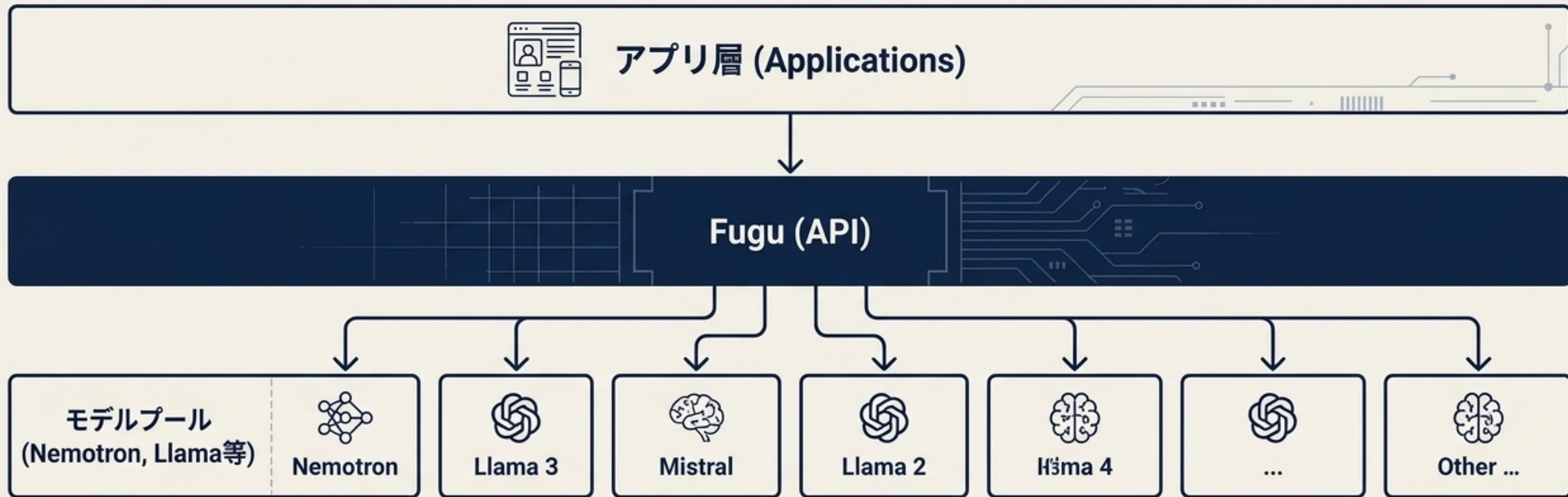
コーディング

DeepSWE（新規長時間ソフトウェア工学タスク）で74.3を記録。

Reality Check: 現段階でのエビデンス評価

Confidence Matrix		
 High/Mid	コスト性能	Maxの出力単価は公表価格で立証済み。10件中7件のPareto frontier拡張主張は有力。
 Mid	複雑なコーディング/ 視覚推論	DeepSWEやChartographyのスコアは極めて有望だが、第三者再現待ち。
 Mid	エージェント性能の 純粹評価	独立研究（StateM論文）が示す通り、スコア向上は「モデル自体の能力」か「実行ハーネスの優秀さ」かの分解が困難。
 Low	長文処理・安全性・ レイテンシ	272K超の料金設定はあるが最大コンテキスト長、長文精度劣化率、P50/P95レイテンシ、定量的な安全性ベンチマークがいずれも未公開。日本語標準多言語評価も不足。

The Meta-Layer Paradox: ロックインの移行と知財のブラックボックス化



The Advantage モデル陳腐化の回避

アプリ側のコード改修なしで、常に最新のオープンモデルの恩恵を享受。特定ベンダーへの依存を軽減。



The Paradox ブラックボックス化

どのモデルが処理したか不透明。リバースエンジニアリング禁止により、自社にルーティングの知見（知財）が蓄積されない。

Enterprise Governance: 本番導入を阻む「標準規約」の壁

避けるべきユースケース

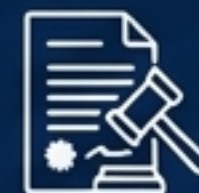


個人情報を含む融資審査

患者のカルテ・医療データの
直接解析

学生の個人データを含む
評価システム

標準規約 (Terms & Privacy Policy) の制約



PIIの入力禁止: ユーザー本人・第三者の個人情報、健康・金融等のセンシティブ情報入力が明文で禁止。

学習利用とオプトアウトの限界: コンテンツの学習利用が許容。オプトアウトは可能だが、過去の学習データへの遡及的削除は保証されない。

データ主権とSLA: 可用性 (Uptime) や応答時間の保証なし。日米等へのデータ越境移転が含まれる。

Regulatory Compliance: グローバルAI法規制への対応

1 EU (AI Act)

現在EEA・UK・スイスでは未提供。
将来参入時、複数事業者をまたぐ
GPAIにおいて、プロバイダー責任
やサブプロセッサの透明性をどの
レイヤーが担うかが焦点に。

US (NIST AI RMF)

モデル単体ではなく、「ルーター
+ ワーカー + ツール + 人間」の
システム全体としてのリスク評
価・管理が要求される。

2 Japan (AI法)

2025年9月1日全面施行。AI法
だけでなく、業法上の義務と照ら
し合わせ、「Terms上のPII入力
禁止」との不整合を企業側が管理
する責任。

Risk Mitigation Playbook: 企業が要求すべき緩和策

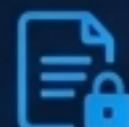
Fugu特有のリスク

推奨される緩和策（法人契約等）

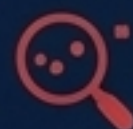


情報漏洩 / 規約の壁

第三者サービス経由でのPII処理、
国外移転。



匿名化・DLPの導入。標準規約での利用を
避け、法人契約で「No-training」「サブプ
ロセッサ明示」「データ所在地」を確約。



幻覚・誤情報

複数エージェントの合意が事実とは
限らない。



RAGによる一次資料固定、Human-in-
the-loop（人間によるレビュー）、出力の
独立確認プロセスの必須化。



挙動ドリフト

バックグラウンドでのモデル入れ替え
による予期せぬ出力変化。



自社Evals（Golden Set）による継続試
験、Sakana への「バージョン/スナップ
ショット固定機能」の要求。

Business Impact: トークン単価を超えた「真のROI」方程式



ROI算出モデル

前提: 知識労働者20人、時給6,000円、年間220日稼働。初年度導入コスト600万円。

効果: 1人1日たった30分削減できれば、年間時間価値は1,320万円。初年度ROIは **120%** に達する。

損益分岐: 1人1日13.6分の削減で投資回収。

Fuguの真の価値は数ドルの節約ではなく、「成功率向上による人間レビュー時間の削減」と「再試行率の低下」で決まる。

Industry Scenarios: 業界別導入シナリオとAPI概算

金融 (Finance)

Use Case

公開決算・市場資料
分析(Max)、複雑な
PDF横断(Ultra)。

API Estimate

月間2万件 ≒ 4.8万
円(Max) / 16.5万円
(Ultra)。

Condition

個人金融情報入力不
可。有資格者レビュー
必須。

医療 (Healthcare)

Use Case

ガイドライン比較
(Max)、精密な文献
統合(Ultra)。

Condition

カルテ等のPII投入
厳禁。専門家関与
必須。

製造 (Manufacturing)

Use Case

マニュアルQA
(Max)、複雑な故障
解析(Ultra)。

Condition

機密設計図・営業
秘密の扱いを法人
契約で明確化。

教育 (Education)

Use Case

教材生成・採点補
助(Max)、高度
STEM(Ultra)。

Condition

学生PII保護。高リ
スク評価の自動化
注意。

コンテンツ (Content)

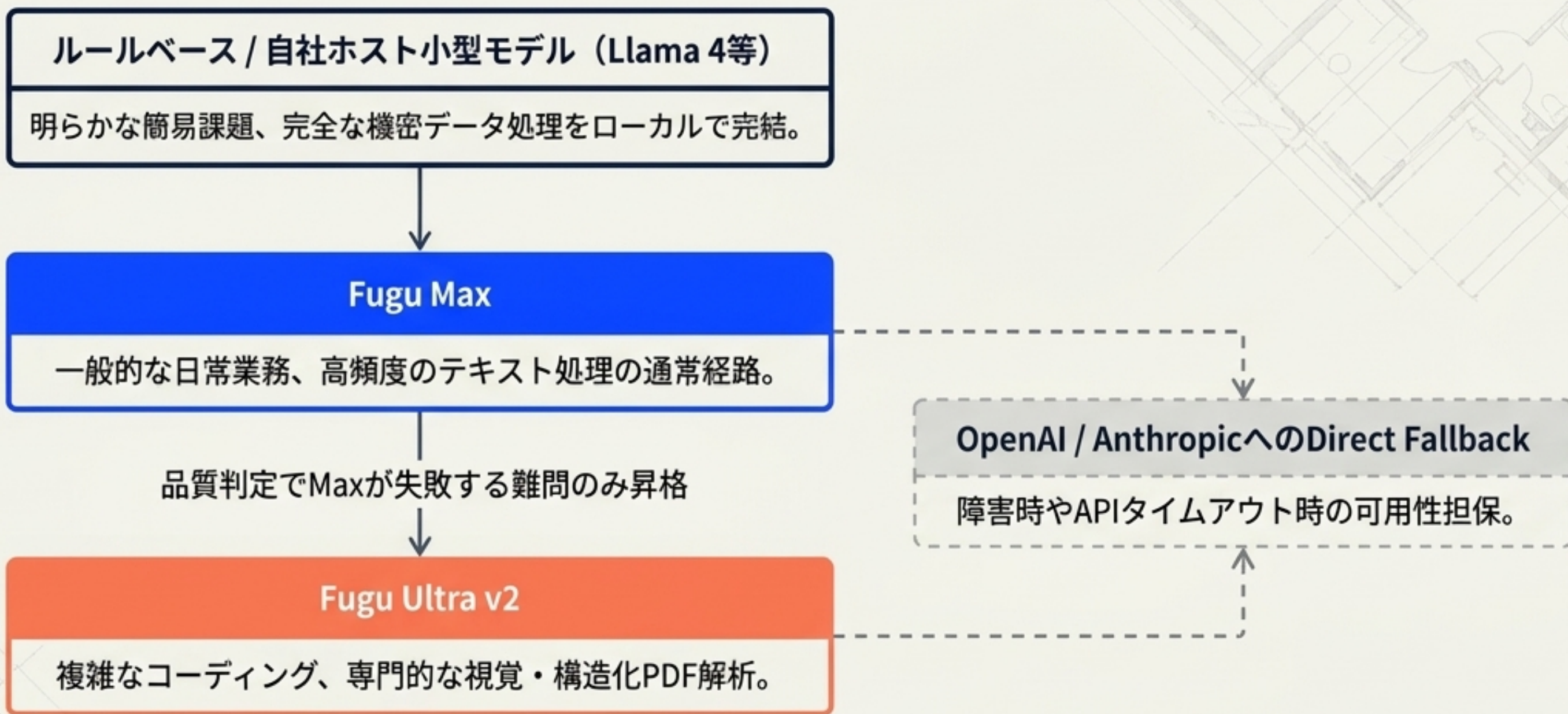
Use Case

構成・翻訳補助
(Max)、整合性検証
(Ultra)。

Condition

著作権確認。AI生
成の透明性担保。

The Enterprise Blueprint: 最適な多段フォールバック・アーキテクチャ



インサイト: 適切な計算資源を適切な課題に割り当てるFuguの思想を、企業インフラ全体へと拡張する。

Future Outlook & Final Verdict: 今後の展望と最終結論

短期
(6-12ヶ月)

第三者再現性の確立とNVIDIA
(Nemotron等) オープンモデル
エコシステムとの連携強化。

中期
(1-3年)

オーケストレーション機能のコモ
ディティ化。SOC2、サブプロセ
ッサ明示など「エンタープライズ
成熟度」が勝負の分水嶺に。

長期
(3-5年)

「単一モデルの所有」から「AI
Operating Layer」の掌握へ。

結論とアクション

「性能を買う前に、ガバナンスを確認せよ」

Fugu Max / Ultra v2は強力なアーキテクチャ
だが、標準規約での本番利用（個人データ投
入等）はリスクが高い。

まずは実データ（PII除外）でMaxをテストし、
DPA、SLA、No-trainingを盛り込んだ個別法人
契約を締結してから本格展開へ進むこと。