



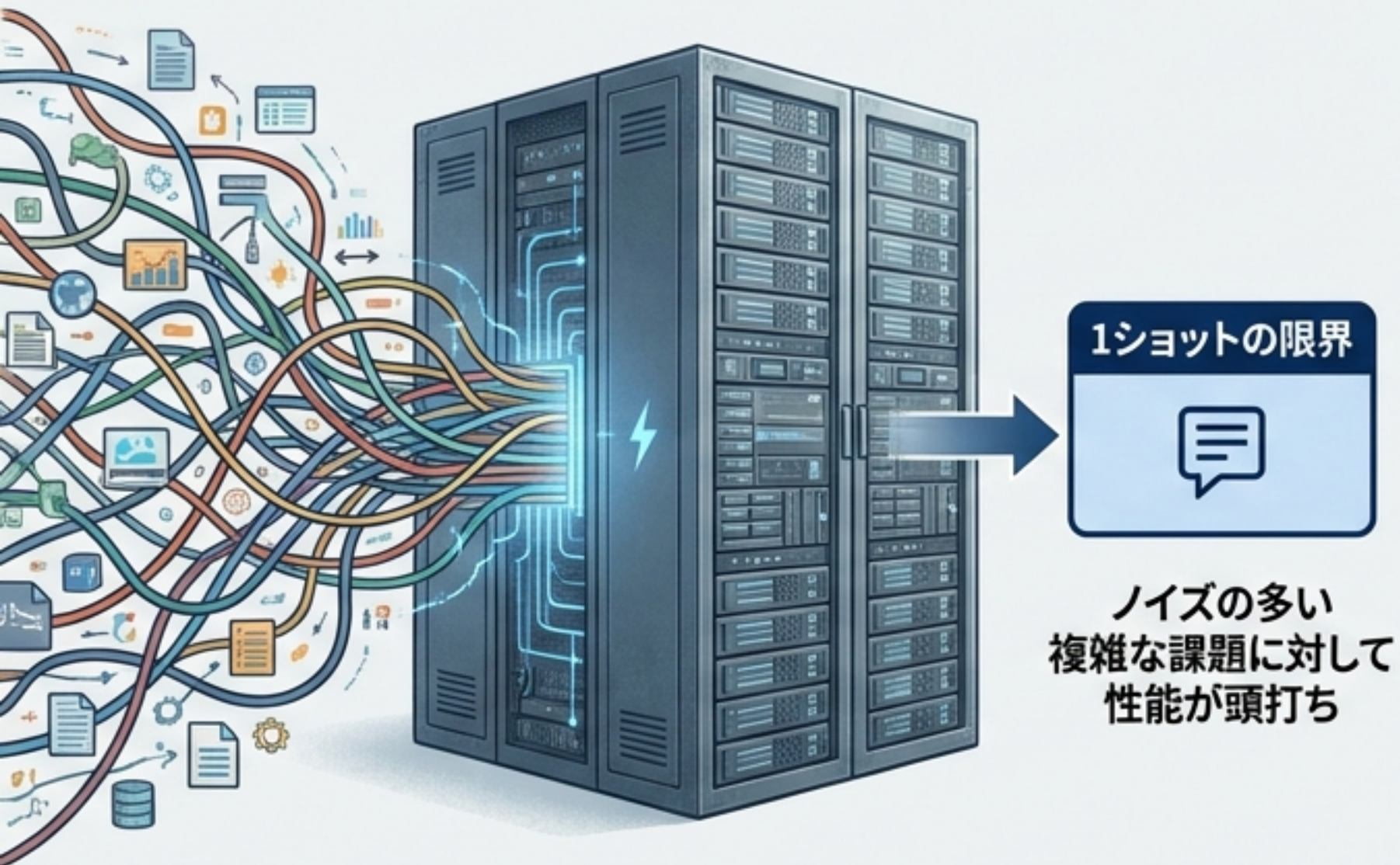
集合知による次世代AIアーキテクチャの確立

Sakana AI「Fugu」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

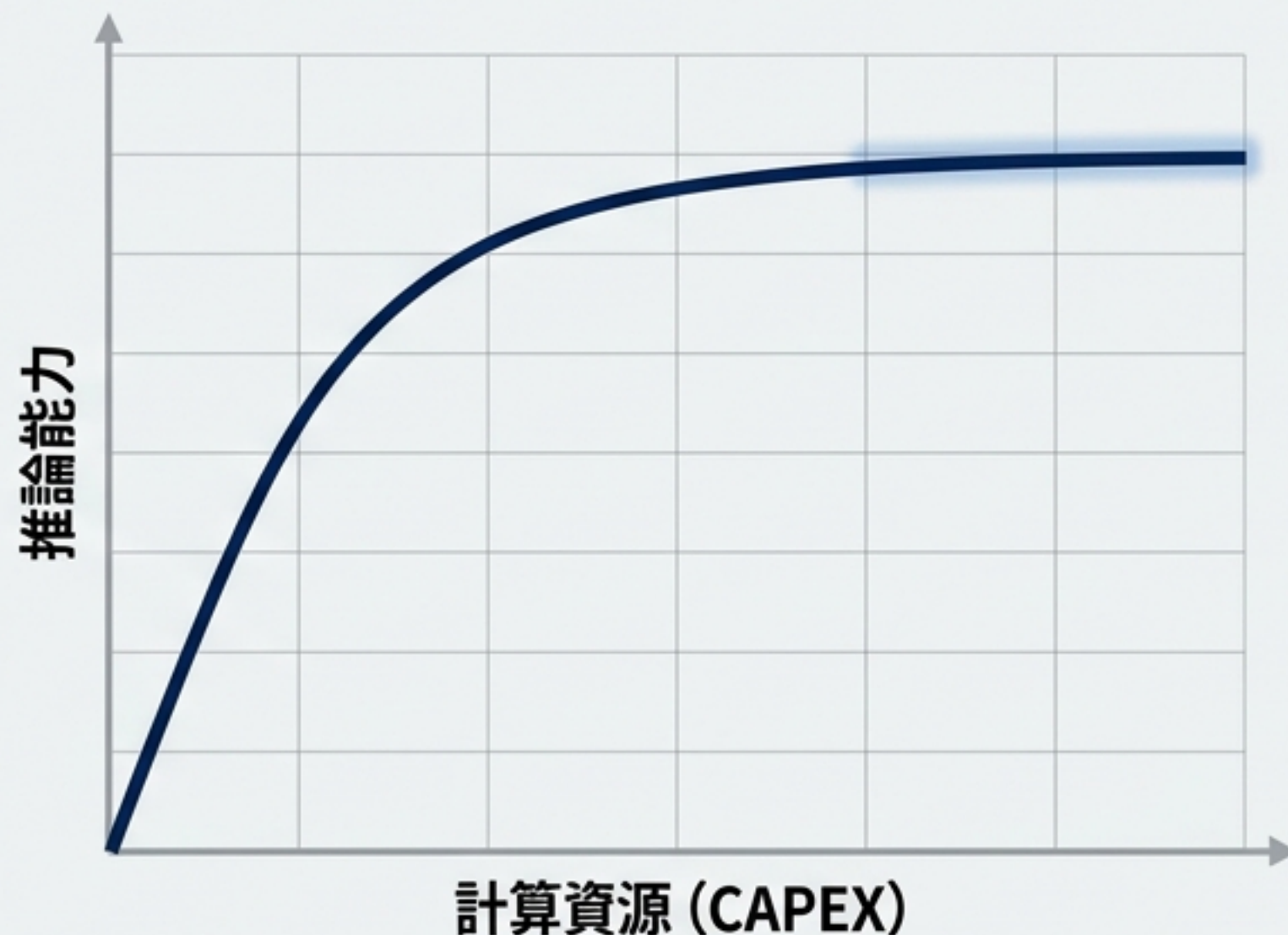
エンタープライズIT戦略およびアーキテクチャ担当者向けブリーフィング

モノリシックな「スケール則」は複雑な現実課題の前で限界に達しつつある

モノリシック基盤モデル



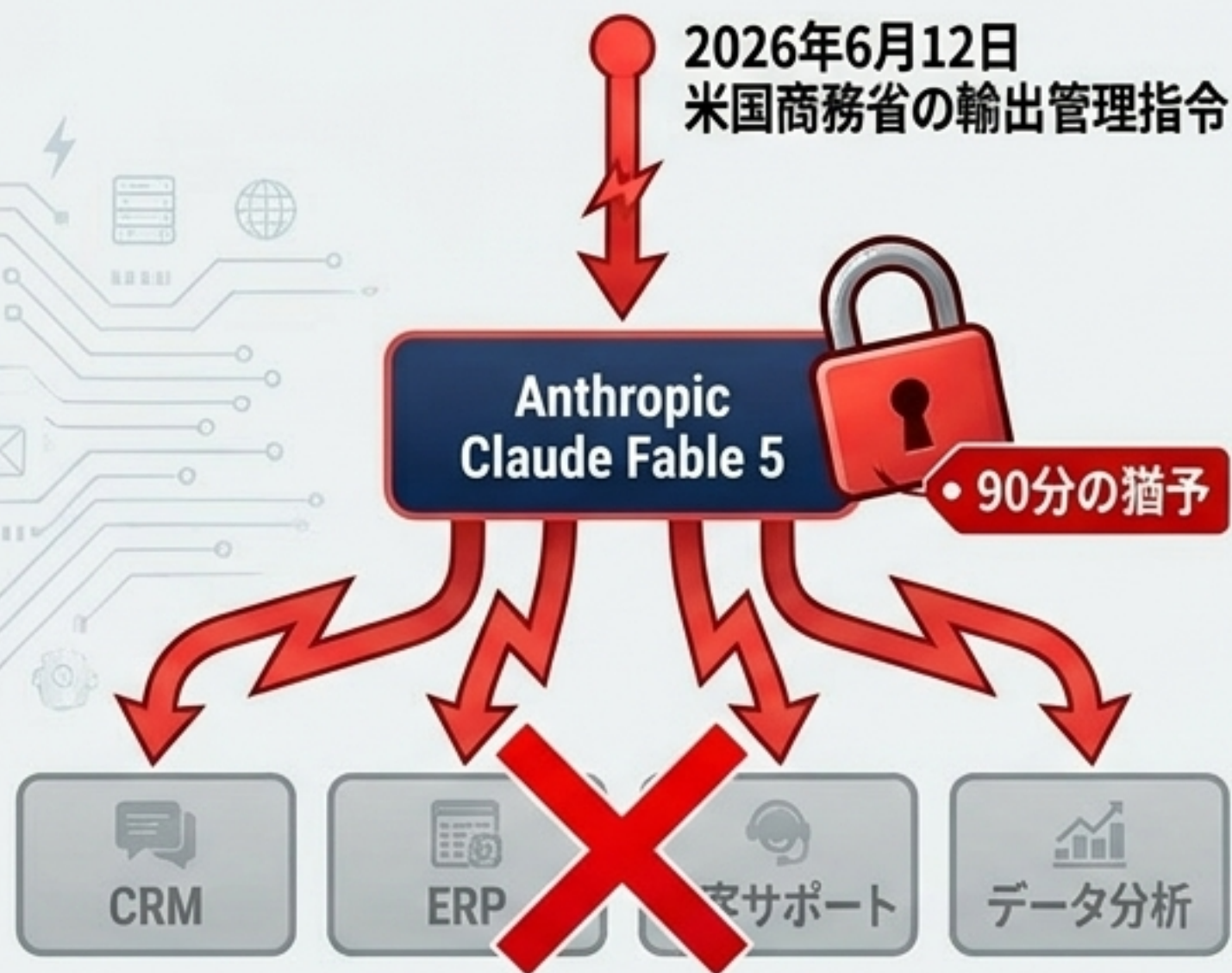
スケール則 (Scaling Laws) の限界



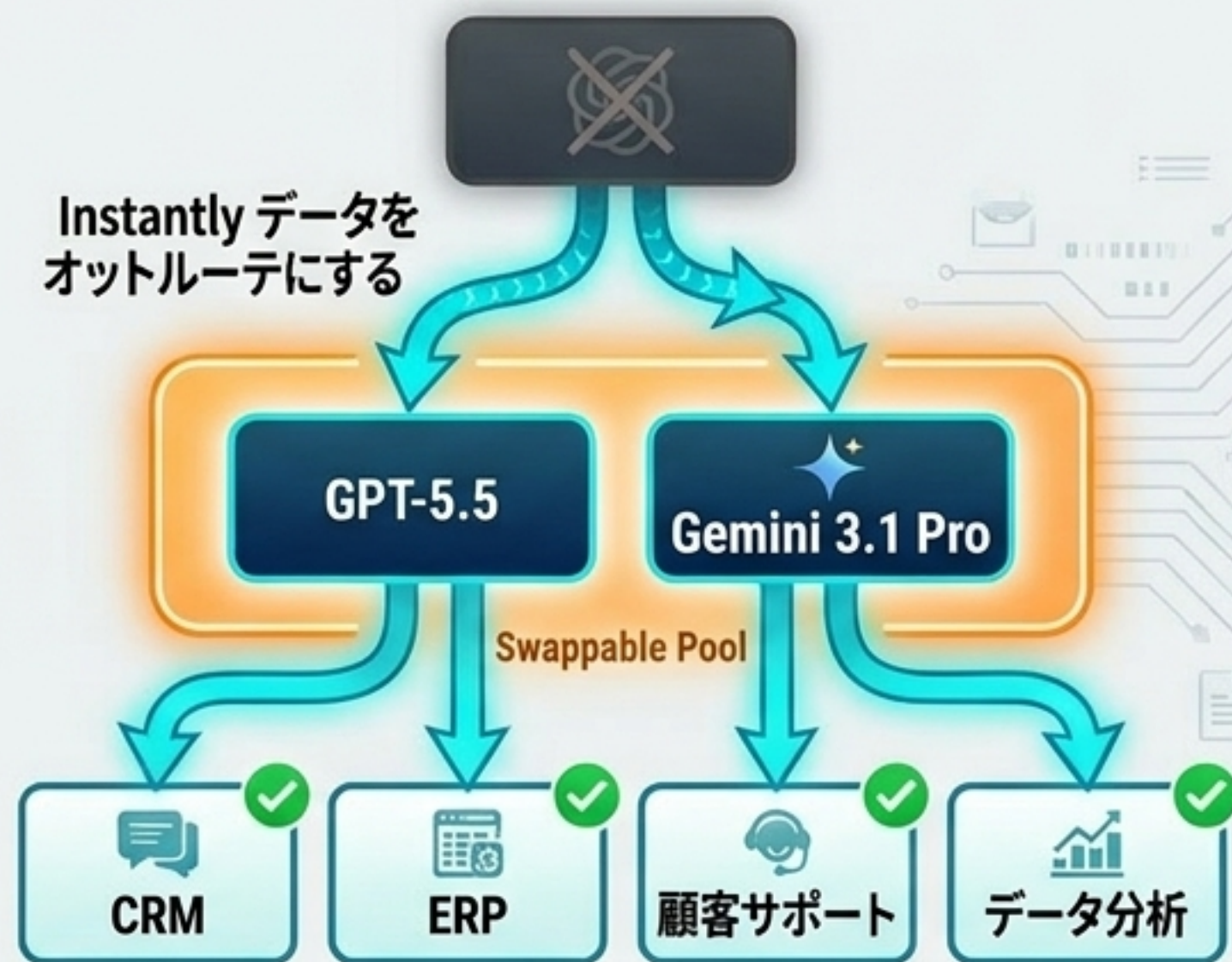
莫大な計算資源 (CAPEX) を投下して単一モデルのパラメータを巨大化させる従来のアプローチは、多段階の推論が求められるエンタープライズ領域において、明確な性能の頭打ち (1ショットの限界) を迎えている。

地政学リスクの顕在化：単一ベンダー依存はもはやビジネス上の致命傷となる

ベンダーロックインによる致命的脆弱性



AI主権 (AI Sovereignty) を担保する動的ルーティング



2026年6月の特定プロバイダーに対する突然の輸出規制（アクセス遮断）は、特定のAPIに依存するリスクを露呈した。Fuguは、障害や規制を自動迂回し、企業の「AI主権」を保護する防波堤として機能する。

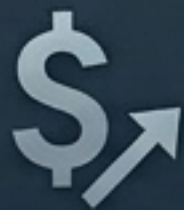
アーキテクチャのパラダイムシフト：単一モデルの所有から、集合知の統率へ

従来のモノリシック基盤モデル

Fugu マルチエージェント・オーケストレーション

アプローチ

事前学習 (Pre-training)
への巨額投資

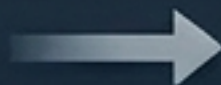


推論時 (Test-time) の
計算力スケーリング



ルーティング

静的・1ショット
(単一モデルへの固定)



動的・マルチラウンド
(自然言語によるタスク分割)



リスク管理

ベンダーロックイン
(単一障害点)



スワップ可能なプール
(AI主権の確保)

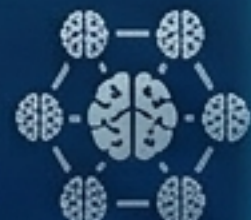


知能の源泉

パラメータの巨大化



メタ・インテリジェンスによる
役割分担と自己修正

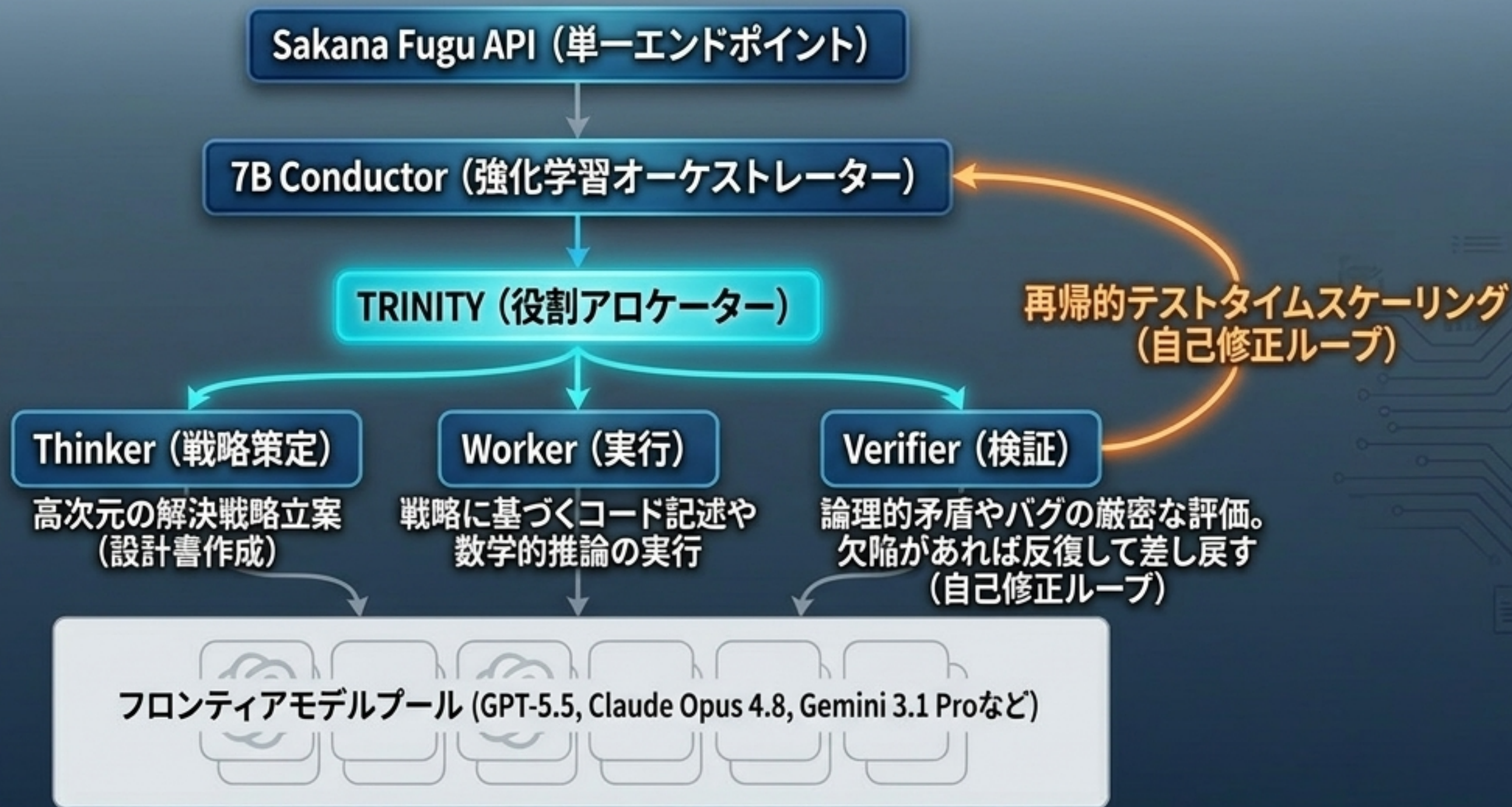


Sakana Fugu: 単一のAPIエンドポイントの背後に広がる最先端モデルの「群れ」



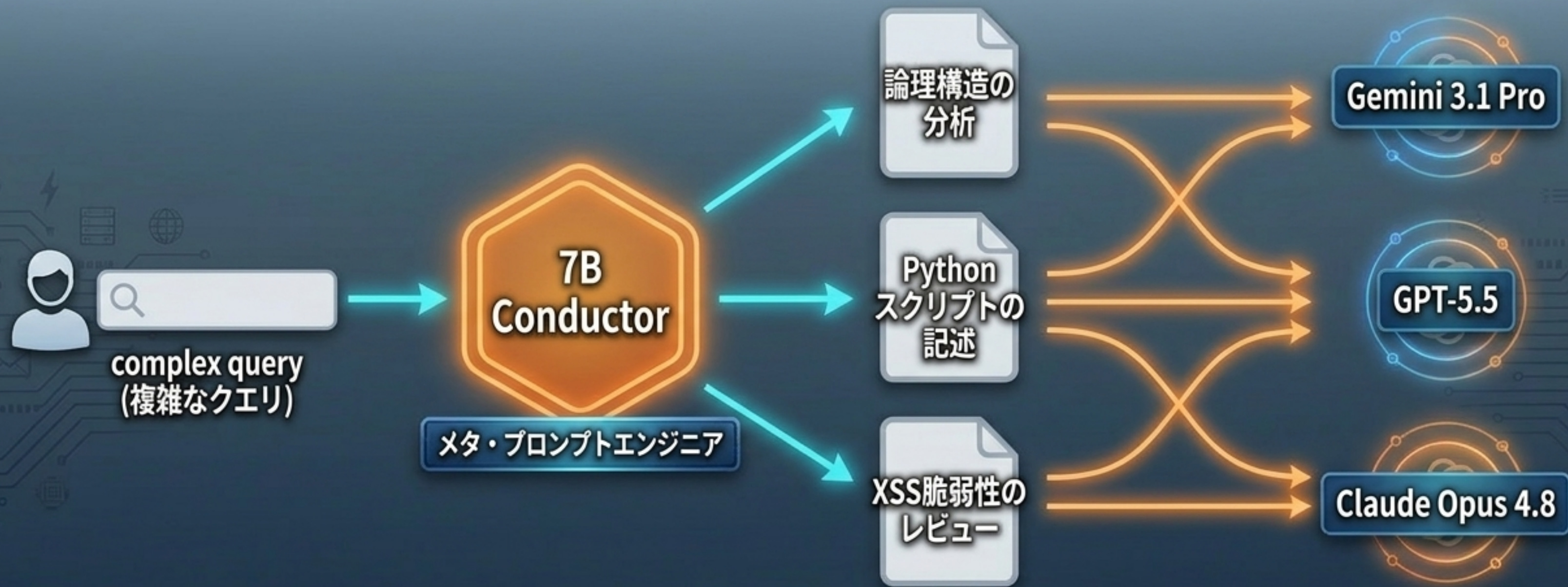
開発者はLangGraphやCrewAIのような複雑なワークフローをハードコードする必要はない。Fuguに既存のAPIキーを向けるだけで、システムが自律的に世界最高峰の外部LLMプールを選定・編成・検証する。

TRINITYアーキテクチャ: 20K未満の進化型SLMによる動的な役割分担



基盤モデルのウェイトを変更することなく、推論実行時に異なるモデルの長所を融合し、「能力の底上げ」を学術的に実証。

Conductor：強化学習が導き出した、自然言語によるエージェントの自律統率

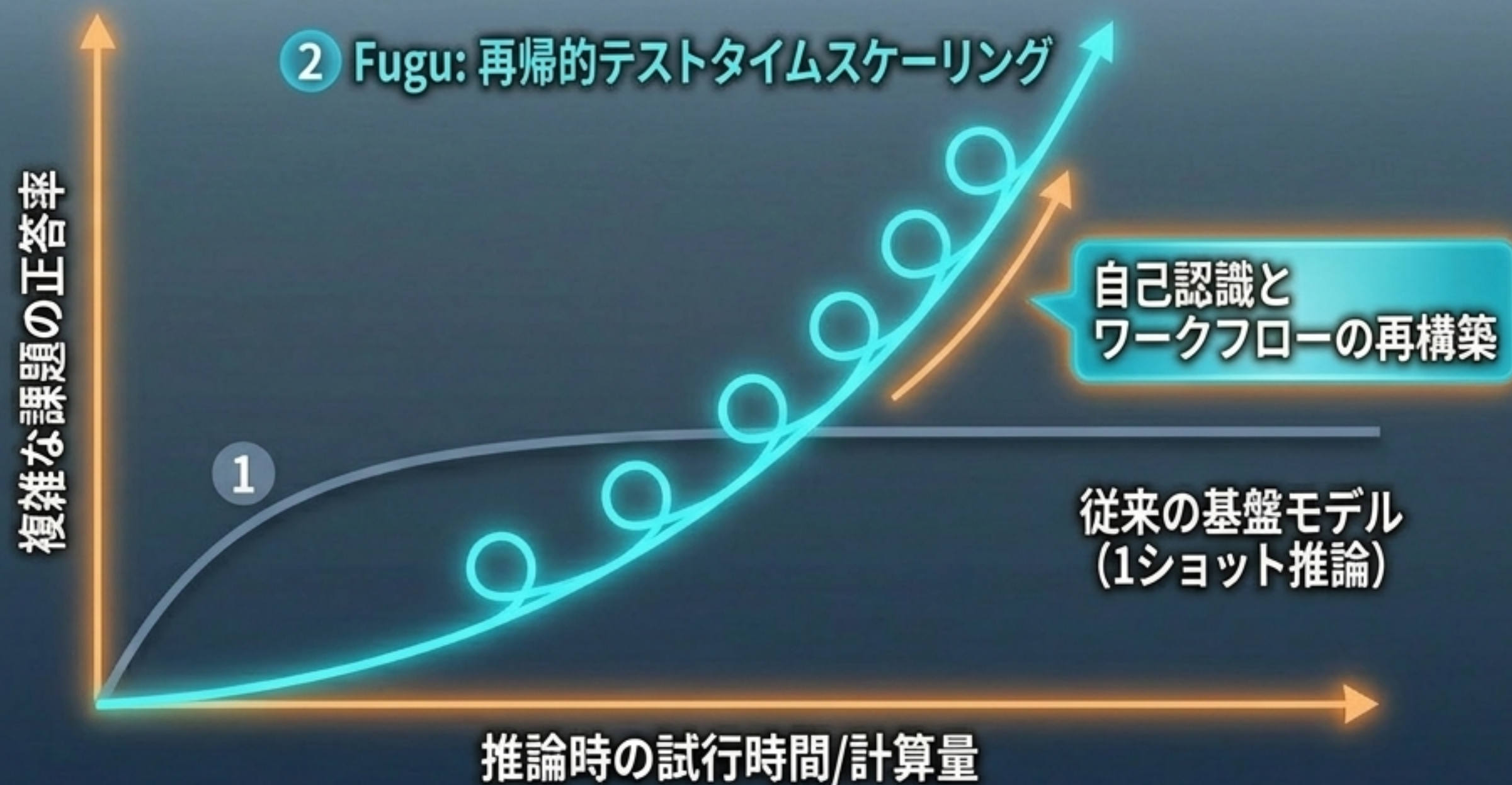


敏腕マネージャーとしての役割：
Conductor自身は直接問題を解かず、
プールのモデルを統制する「メタ・
プロンプトエンジニア」として機能。

自律的な3つの決定：
1. エージェントの選定 (誰に任せるか)、
2. タスクの細分化と指示 (カスタムプロンプトの動的生成)、
3. コンテキストの管理 (ハルシネーションを防ぐ記憶の最適化)。

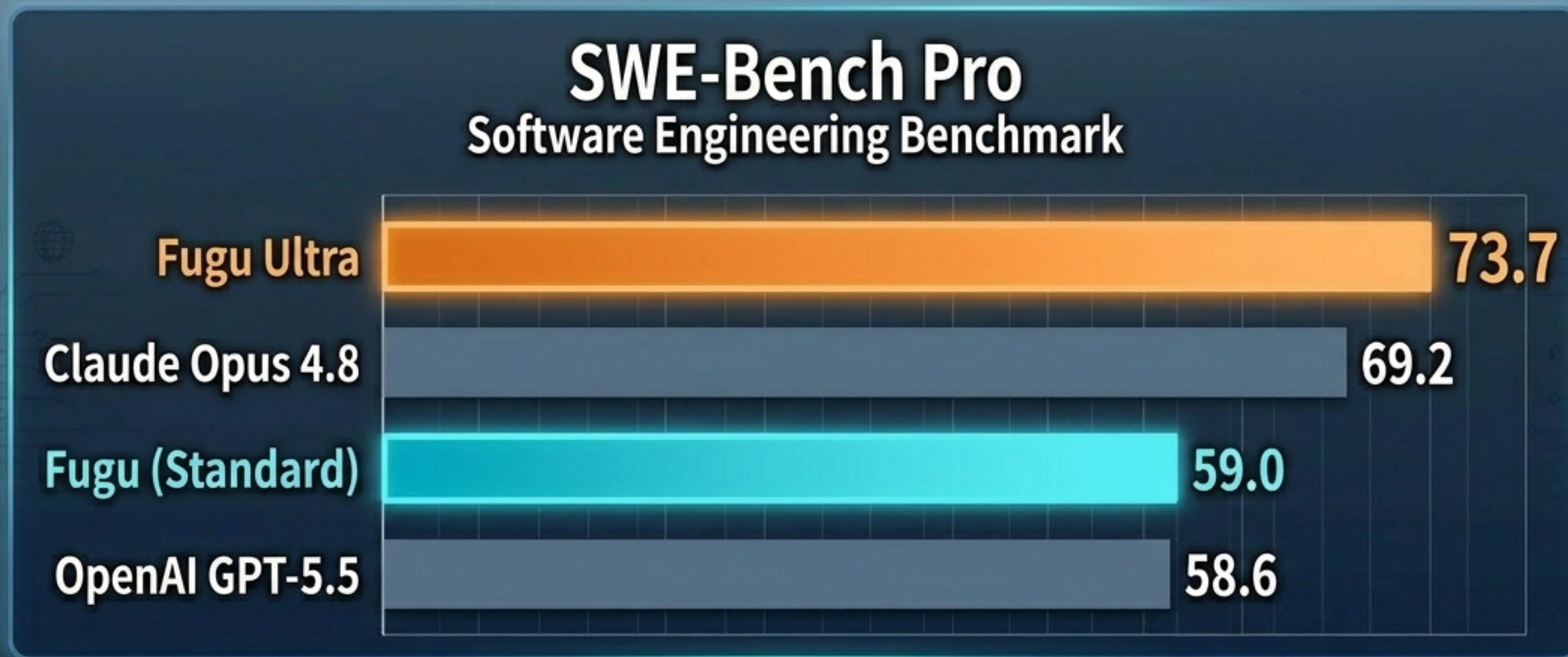
創発的な連携：
タスク完了という報酬の最大化のみ
を目的とし、最適なチーム編成を
即座 (オンザフライ) に構築。

再帰的テストタイムスケーリング：推論時間を拡張し、性能の壁を突破する



Conductorはワーカプールの中から「自分自身のインスタンス」を再帰的に召喚し、過去のアプローチの行き詰まりを俯瞰・自己修正する。事前のデータ学習量（パラメータ）に依存せず、推論時の試行時間をかけるほどに正答率が劇的に向上する新パラダイム。

定量評価：現存する最高峰の個別基盤モデル（SOTA）を凌駕する推論性能



*AnthropicのFable 5 (80.0) は輸出規制により一般アクセス不可

複数ファイルにまたがるバグ修正や長期的コンテキストが要求される最高難易度のソフトウェアエンジニアリング（SWE-Bench Pro）において、Fugu Ultraは他社トップモデルに対し圧倒的なスコア差を記録。自己検証ループの有効性を証明した。

実務環境における圧倒的な有用性：500社のベータテストが示す定性的価値



ペルソナの一貫性維持

長時間の複雑なセッションにおいても、AIが設定されたキャラクターや役割を忘れず、ペルソナのブレ（ドリフト）が発生しない。



網羅的な欠陥検出

実際の業務コードレビューにおいて、GPT-5.5が3件しか指摘できなかった潜在的なバグや脆弱性を、多角的な視点から20件以上検出。



自律的セキュリティ診断

破壊的アクションを安全に回避しながら、XSSやSQLインジェクションのチェックを自律実行し、クリーンな再現報告書を生成。



データサイエンスの自動化

人間の介入ゼロで小規模モデルの学習レシピを自律的に改善。BPBスコア0.9774という記録的な精度を達成。

単なるベンチマークの数値にとどまらず、ノイズの多い現実のエンタープライズ業務（Messyなユースケース）において、人間の専門家チームに匹敵する自律的な課題解決能力を発揮する。

エンタープライズ・デプロイメント：ワークロードに応じた2つのバリエーション

Fugu (標準モデル)

- **対象：**
日常的業務、チャット、IDEバックエンド
- **ルーティング：** コンプライアンス向け
「オプトアウト（除外）機能」完備
- **レイテンシ：**
最小限の遅延（直接回答による最適化）

Fugu Ultra (最上位モデル)

- **対象：**
AI研究、サイバー脆弱性分析、複雑なデータ分析
- **ルーティング：**
固定の最高峰プールを強制利用（除外不可）
- **レイテンシ：**
TTFT（初回トークン生成）は増加するが
極限の精度を提供

厳格なデータガバナンスと応答速度を重視する「標準モデル」と、応答時間を代償に極限の推論精度を追求する「Ultra」により、あらゆるエンタープライズ要件を網羅。

経済的合理性：用途に合わせた柔軟なコスト構造とビル・ショックの防止

月額サブスクリプション（定額制）



Standard (\$20/月)

エントリー



Pro (\$100/月)

10倍の利用枠・リサーチチーム向け



Max (\$200/月)

20倍の利用枠・ヘビーワークフロー向け

従量課金プラン（エンタープライズ向け）

Fugu（標準）

動的課金構造を採用。背後で複数のモデルが稼働した場合でも、「最上位モデルの単一レート」のみが適用され、意図せぬコスト爆発（ビル・ショック）を防止。



Fugu Ultra

	272Kトークン以下	272Kトークン超過
入力/1M	\$5.00	\$10.00
キャッシュ/1M	\$0.50	\$1.00
出力/1M	\$30.00	\$45.00

既存のクライアントアプリケーションからのマイグレーション不要。
OpenAI互換APIエンドポイントの変更のみで、即座にエンタープライズ統合が可能。

トレードオフの理解：バックグラウンド・トークンとROIの最大化



- Fugu Ultraの複雑な内部検証プロセスでは、画面に表示される最終テキストの何倍もの「バックグラウンドトークン」が消費・課金される。
- 企業は「API内部でのトークン消費コストの増加」と、「出力品質の飛躍による人的修正コストの劇的な削減」というROIを総合的に評価する必要がある。

覇権構造の転換：価値の源泉は「巨大モデルの所有」から「連携技術」へ

Sakana Fugu = インテリジェンス調整・統率レイヤー (The Manager)



巨大で高価なクローズド基盤モデルの性能向上曲線はピークを迎え、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市場を制覇する。

巨大基盤モデル = コモディティ化された計算エンジン (The Workers)

Sakana Fuguは単なるルーターではない。巨大テクノロジー企業の莫大なCAPEXを土台とし、最も価値の高い「メタ・インテリジェンス」を支配することで、AI業界のスケーリング則と地政学的パワーバランスを完全に書き換えたのである。