

Sakana Fugu：複数のLLMを統合する次世代AIオーケストレーションの全貌



2026年6月22日
正式提供開始

4月のβテストを経てGA（一般提供）へ移行。日本発のグローバル雇員を見据えたマネージドAPIサービスです。



マルチモデル協調フロー

タスク分析とルーティング

ユーザー
閲覧



タスク分析と
ルーティング

入力されたタスクの難易度や
領域を判定し、最適なワーカ
ーモデルを割り当てます。



Trinity
(軽量ルーター)



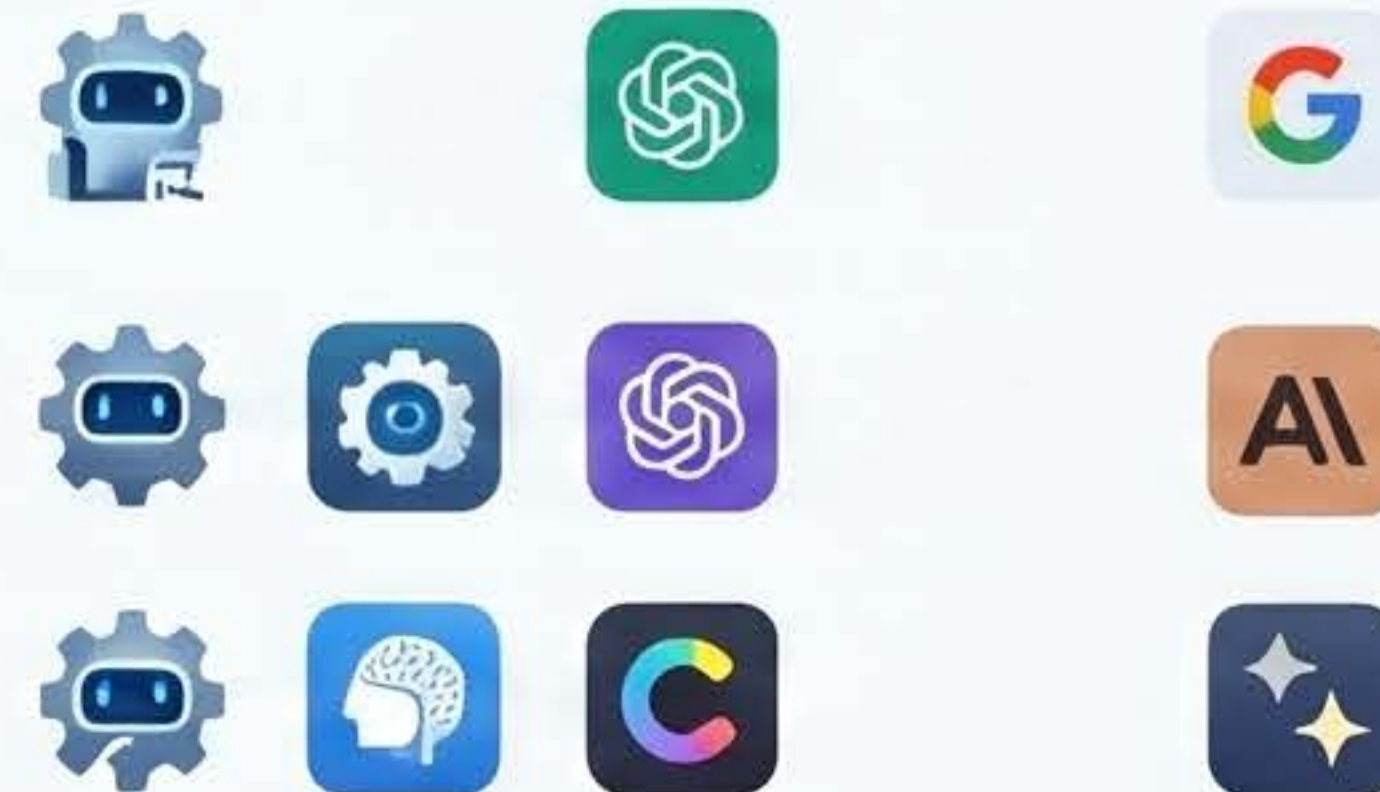
Conductor
(ワークフロー設計)



Trinity（軽量ルーター）や
Conductor（ワークフロー
設計）を用いて最適
なワーカーモデルを割り
当てます。

ワークフローの構造的定し、
最適的なタスクを指定

並列・逐次実行による解決



複数のエージェントがサブタスクを並列処理、あるいは
段階的に処理を異にし、複雑な問題を分解します。

検証と回答合成 (Synthesis)



各ワーカーの出力を検
査・統合し、最終的な
回答を生成してユーザ
ーに返します。

料金プランと利用コスト

Standard:
\$20/月
(約3,226円)

Pro:
\$100/月

Max:
\$200/月

ProはStandardの10倍、Maxは
20倍の利用枠をもちます。

Fugu Ultraの従量課金（1M tokensあたり）

入力

\$5

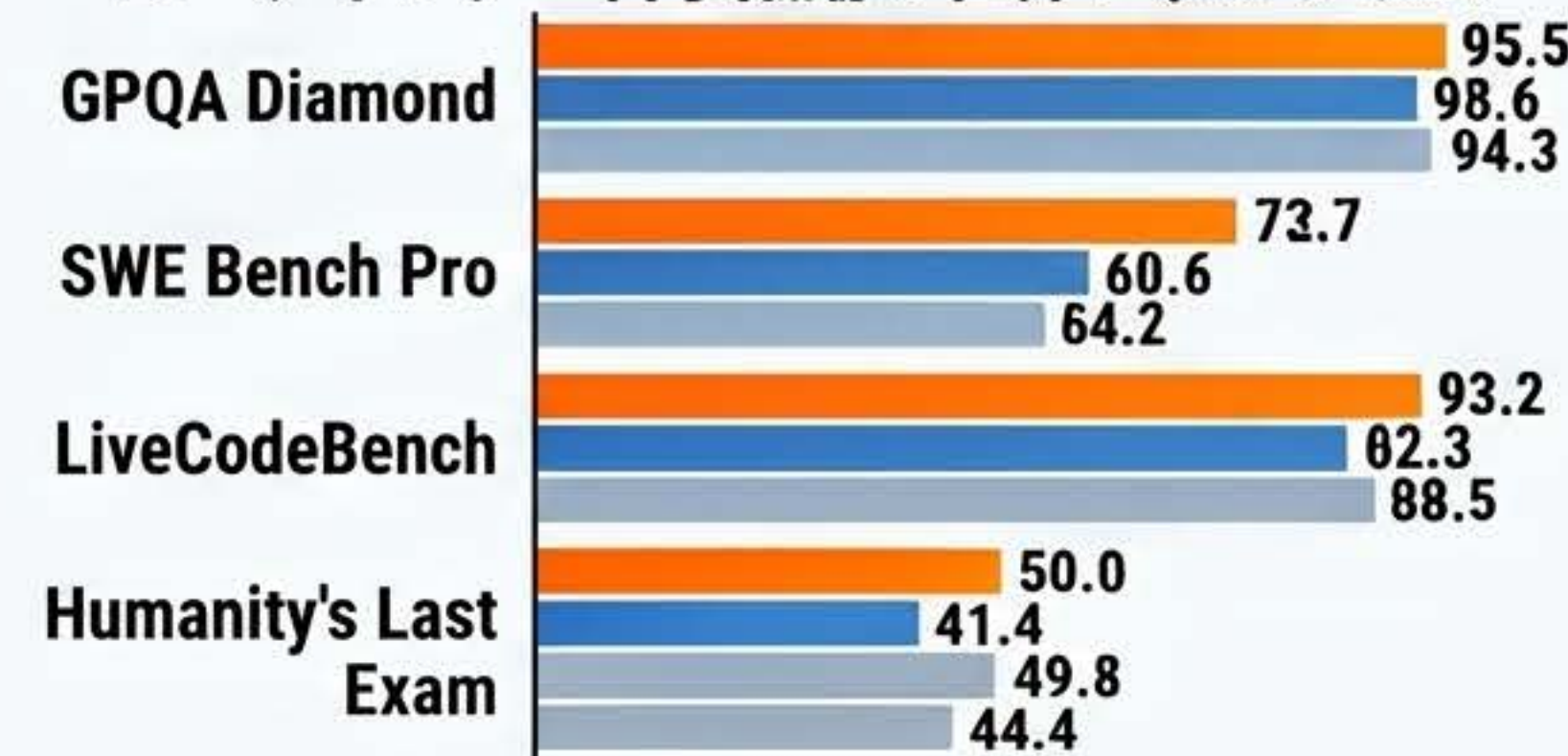
出力

\$30

- ・オーケストレーション用のトークンも課金対象となるため、誤差が生まれる可能性があります。
- ・Standardは約3,226円、Ultraの標準入力100万トークンあたり約806円（1ドル=161.289円換算）。

圧倒的なベンチマーク性能

コーディング・科学論文で世界トップクラス



！ 導入時の重要チェック事項 ！



データ保護と学習利用の明示

入出力コンテンツは自社モデルの学習・改善や人手レビューに使
われる可能性があり、デフォルトで「機密情報入力不可」の検討が。



外部プロバイダーへのデータ開示

処理経路でOpenAIやGoogle等の外部ベンダーへデータが送信さ
れます。データ主権を登録する企業はopt-out規定が必須です。



推奨される用途と対策

信頼性や用途、質の高いコーディングに古く一方、機密情報の投入制
限、利用基準の策定、人間による出力確認が推奨されます。