

企業 AI 支出、重点 5 領域の管理で 12～18 カ月に 20～40%の削減余地【A.T. カーニー】

推論費用からデータ保存・監視まで、AI 支出の継続管理を提言

A.T. カーニー株式会社（東京都港区、日本代表：針ヶ谷 武文）は、論考『AI を実験から企業価値へ：全社規模で AI のコスト曲線を管理する』（英題：From AI experimentation to enterprise value: managing the cost curve of AI at enterprise scale）を公開しました。本論考は、企業の AI 活用が実験段階から全社展開へ進む中、構築後の運用費用をどのように管理するかを論じています。

本論考は、データストレージ、MLOps・可観測性ツール、オーケストレーションの乱立、外部データ、推論費用の 5 領域を重点対象に挙げています。これらに集中的に取り組むことで、12～18 カ月で経常的な AI 支出の 20～40% を削減できるとしています。削減効果を維持するには、用途ごとの費用を把握し、FinOps（利用量に応じて変動する IT 費用を継続的に管理・最適化する運営手法）と全社共通の運用基盤を整える必要があると提言しています。

AI の全社展開で、費用の中心は「構築」から「継続運用」へ

本論考は、企業の AI 関連費用を人件費、データ、推論・コンピューティング、ソフトウェア・プラットフォームの 4 つに整理しています。人件費は開発段階、推論費用は本番運用に集中し、ユースケースが増えるにつれて費用構成も変わります。費用の負荷が高い領域として、モデル利用料に加え、検索を支えるデータストレージやエージェントを監視するツールを挙げています。

新世代のエージェントは、単純なチャットボットに比べてタスク当たり 5～30 倍のトークンを使います。問い合わせ 1 件当たりの単価が下がっても、利用量の拡大によって請求総額が増え得るため、本論考は現在の費用と増加傾向の両方を見て優先順位を決めるよう促しています。

図表1 AI支出の重点管理領域は5つ

本論考が規模と成長傾向を踏まえて優先対象に挙げる費目

01	データストレージ・検索	検索を支える保存・処理費用
02	MLOps・可観測性ツール	エージェント群の監視・評価費用
03	オーケストレーションの乱立	重複する基盤と保守工数
04	外部データ	チームごとに重複するライセンス
05	推論費用	本番運用で継続するモデル利用料

出所：Kearney論考「From AI experimentation to enterprise value: managing the cost curve of AI at enterprise scale」

重点領域ごとに異なる削減余地を、自社の支出に照らして評価

本論考が示す削減幅の目安は、データストレージ・検索でデータ費用の15～40%、オーケストレーションの統合でプラットフォーム費用の20～35%、外部データの調達一元化でデータライセンス費用の10～25%です。一般的なワークロードでは、推論費用についても30～60%の削減余地を挙げています。MLOps・可観測性・FinOps ツールの整備には、単独でも5～15%の削減効果があるとしています。これらは公開ベンチマークやベンダー事例などに基づく目安であり、数値を合算せず、実行前に自社の支出構造に照らして再評価する必要があります。

推論費用ではキャッシュとモデルの振り分けが有力な施策ですが、別々に最適化すると逆効果になる場合があります。会話の途中でモデルを切り替えると、それまで蓄積したキャッシュが失われることがあるためです。本論考は、モデル変更による節約額と失うキャッシュの価値を比較できるよう、両者を一体で設計することを提案しています。

図表2 重点5領域で見込む削減余地

削減幅は施策ごとの目安。対象費用・実施期間はそれぞれ異なる

データストレージ・検索 15～40% 対象：データ費用 6～12カ月	MLOps・可観測性 5～15% 単独施策の効果 他施策の測定基盤	オーケストレーション 20～35% 対象：プラットフォーム費用 12～18カ月
外部データ 10～25% 対象：データライセンス費用 目安	推論 30～60% 対象：推論費用 一般的な業務処理	

注：数値は合算できません。施策前に自社の支出構造に照らした再評価が必要です。

出所：Kearney論考「From AI experimentation to enterprise value: managing the cost curve of AI at enterprise scale」

FinOps の運営規律と共通基盤で、削減効果を継続

本論考は、AI コスト管理を一度限りの施策で終わらず、測定、統制、予測、最適化の 4 段階を継続するよう提言しています。用途・チーム別に費用を配賦し、予算や利用上限を設け、将来の利用量と事業成果当たりの費用を見通すことが重要です。最適化の対象には、キャッシュ、モデルの振り分け、バッチ処理などを挙げています。

これらを全社規模で運用するため、論考は統合された推論ゲートウェイ、コストの計測・配賦、予測・予算管理ツール、共有のデータ・検索基盤、エージェントの共有、評価・可観測性、ガバナンスの仕組みを示しています。着手点として、今四半期に支出上位 3 つの AI ユースケースを特定し、それぞれの費用配賦を監査することを提案しています。

図表3 AI向けFinOpsは4段階で継続運用

費用を用途別に把握し、共通の仕組みで統制と改善を続ける



出所: Kearney論考「From AI experimentation to enterprise value: managing the cost curve of AI at enterprise scale」

論考について

論考名: 『AI の実験から企業価値へ: 全社規模で AI のコスト曲線を管理する』（英題: From AI experimentation to enterprise value: managing the cost curve of AI at enterprise scale）

URL: <https://www.jp.kearney.com/issue-papers-perspectives/from-ai-experimentation-to-enterprise-value-managing-the-cost-curve-of-ai-at-enterprise-scale>

監修者

大塩 崇 シニアパートナー

早稲田大学理工学研究科卒。NTT データ、米系戦略コンサルティングファームを経て、A.T. カーニーに入社。ハイテク、通信・ICT、エレクトロニクス、商社を中心に、シナリオプランニング、新規事業構築、Go-to-market、事業ポートフォリオの再構築、中長期経営計画、M&A 戦略、収益改善、オペレーション改革などのコンサルティングに従事。

川崎 健史 パートナー

東京大学理学部卒、Duke University, Fuqua School of Business (MBA) 修了。NTT 東日本、米系戦略コンサルティングファーム、Big4 コンサルティング部門等を経て、A.T. カーニーに入社。日系及び東南アジアの通信・ハイテク企業を中心に、中期経営計画・グローバル経営マネジメント・テクノロジー戦略・事業モデル変革・業務パフォーマンス変革等の幅広い CXO アジェンダに従事。近年では、戦略コンセプトの立案から実行サポートまで一貫したデジタルトランスフォーメーションのサポートも数多く手掛ける。

A.T. カーニーについて

A.T. カーニー（グローバル・ブランド名：Kearney）は、100 年にわたり世界有数の経営コンサルティングファームとして、Fortune Global 500 企業の 4 分の 3 以上をはじめ、世界各国の政府機関に信頼されるパートナーであり続けてきました。世界 40 カ国以上に拠点を展開し、私たちの最大の強みは「人」にあります。インパクト・ファーストを掲げ、独創的な発想と実行力をもって、顧客企業が直面する最も困難な課題に挑み、変革の実現をともに推進します。日本には 1972 年に進出し、主要産業のリーディングカンパニーに、戦略策定から変革の実行まで一貫した支援を提供しています。 <https://www.jp.kearney.com/>

本件に関するお問い合わせ先

部署名：A.T. カーニー株式会社 広報

メール：Japan.PR@kearney.com