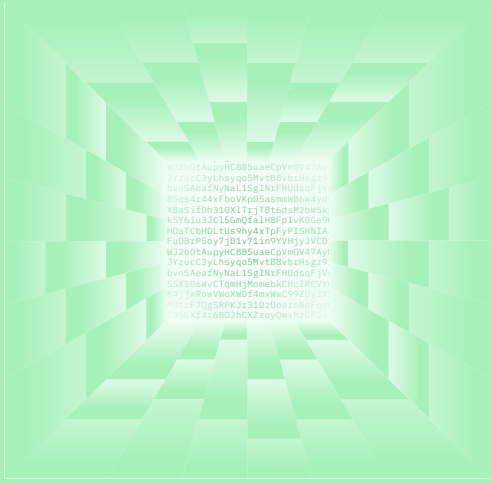
CEO のための生成 AI 活用ガイド No.18
AI モデルの最適化



生成 AI を使い分けて成果を最大化する

生成 AI は過去のどのテクノロジーとも異なっている。
瞬く間にビジネスと社会を揺るがす存在になりつつあり、
リーダーはこれまでの想定や計画、戦略の見直しを迫られている。

こうした変化に CEO が対処するための一助として、IBM Institute for Business Value は生成 AI の調査に基づくガイドをシリーズ化し、テーマごとに公表している。内容はデータ・セキュリティーからテクノロジー投資戦略、顧客体験にまで及ぶ。

今回は第十八弾として「AI モデルの最適化」をお届けする。

その用途に応じた生成 AI モデルがある

ChatGPT は、誰もが AI のエキスパートになったような気分にさせてくれた。しかし、そのシンプルさは「見せかけ」である。CEO が AI モデルのポートフォリオを構築する際に考慮すべき生成 AI にまつわる複雑性を、覆い隠している。

生成 AI モデルは多種多様である。何ができるか、どれだけうまく機能するか、そしてどれほどコストがかかるかは、千差万別である。モデルの所有権、開発方法、トレーニング用データ・セットのサイズは、異なるさまざまなモデルをいつどのように使用するべきかを判断する変数のほんの一部に過ぎない。

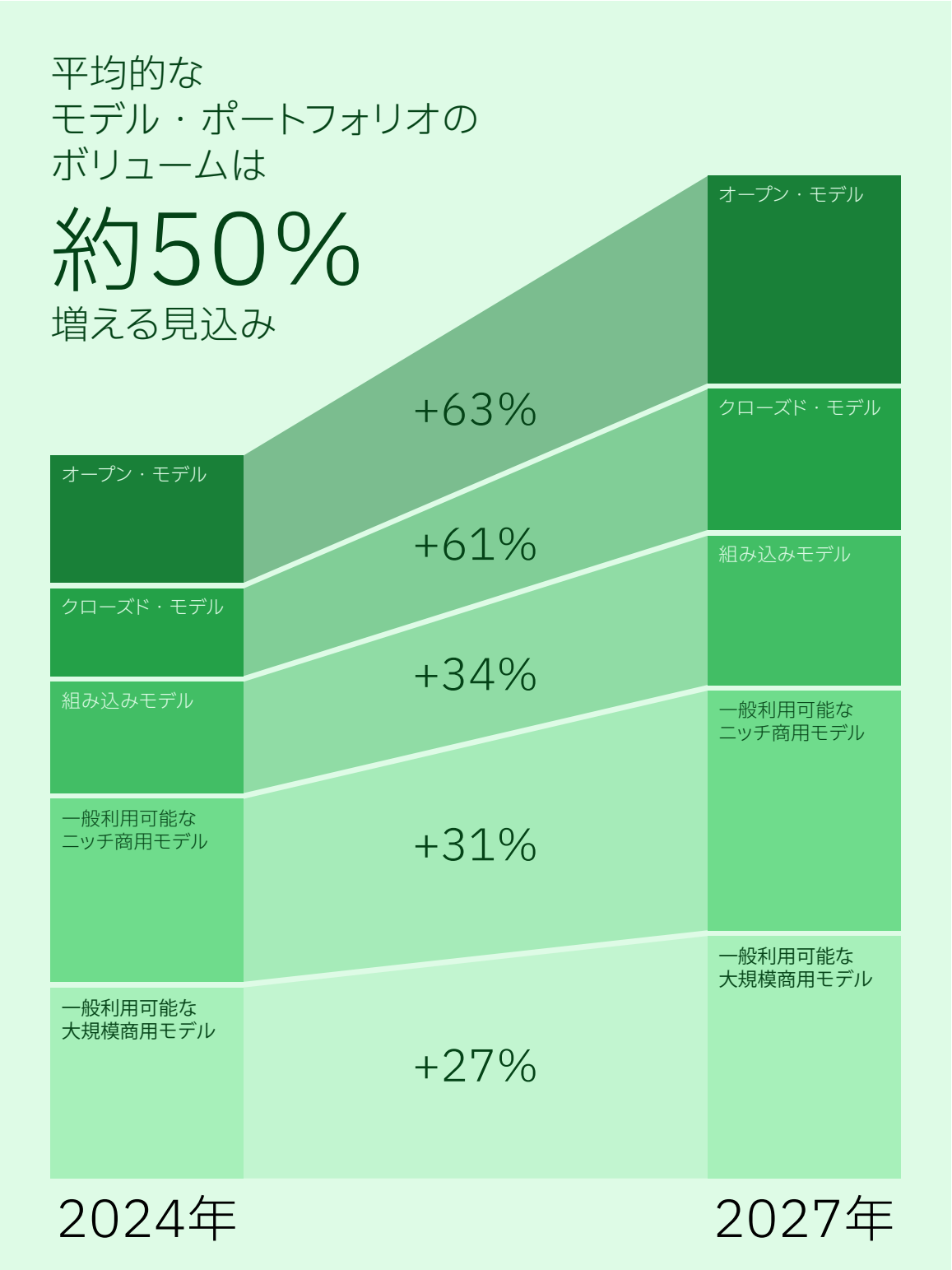
たった 1 つの大規模言語モデル（LLM）のトレーニングに必要なデータおよびリソースの量は膨大で、そのサイズをどうすべきかが、生成 AI に関する多くの議論を占めている。その結果、多くの CEO は、大規模な生成 AI モデルを導入すべきなのか、もしくは、特定の目的に合わせて小規模でニッチなモデルを開発すべきか頭を悩ませている。

答えは、その両方が必要である。

そして、多くの企業はすでにこれを実践している。現在、標準的な組織は 11 の生成 AI モデルを使用しており、今後 3 年以内にそれらのモデルのポートフォリオは約 50% 増えると見込まれている。

なぜそれほど多くなるのだろうか。ユースケースごとに個別の要件と制約があるからだ。そして、ビジネス上の問題が異なれば、必要とされるモデルも異なるためである。

例えば、画像編集やデータ分析といった専門性の高いタスクには、小規模でニッチなデータ・セットによってトレーニングされた生成 AI モデルが必要である。機密情報や専有情報を扱う作業には、機密保持が可能で、情報が外部に漏れない生成 AI モデルが必要となる。また、テキスト生成のような、より一般的なタスクには、できるだけ多くのデータ・セットでトレーニングされた生成 AI モデルが必要となるかもしれない。



CEO はさまざまなモデルの違いを詳細まですべて理解するチームを持つべきである。それと同時に、各タスク、つまり、生成 AI のアプリケーション一つ一つに対して、適切なモデルを選定する重要性を認識することも不可欠だ。コスト、環境負荷、ビジネス価値を増やす要因を知っておけば、AI ポートフォリオのパフォーマンスを最適化したり、競争に打ち勝つために必要なツールを提供しやすくなったりする。

IBV が考える、すべてのリーダーが知っておくべき 3 つのこと：

1. 万能の生成 AI モデルなど存在しない。
2. 生成 AI のコストはすべてコントロールできる。
3. 生成 AI による優位性はいつか消え去る。

そして、すべてのリーダーが今すぐ実行すべき 3 つのこと：

1. ハンマーとメスを使い分ける。
2. 自分なりの生成 AI のツボを見つける。
3. モデルを最大限に活用する。

1. アジリティー + 生成 AI

リーダーが知るべきこと

万能の生成 AI モデルなど存在しない

適切なモデルを、適切な環境で、適切な目的のために活用すれば、生成 AI は、高い精度とアジリティーを持って、よりスピーディーに組織を動かすのに役立つ。

どこでどの生成 AI モデルを使うかを決めるのはテクノロジー・リーダーが適任だ。一方、さまざまなモデル・タイプの長所と短所、競争環境の方向性を理解することで、CEO はより確実な投資判断ができるようになる。

モデル・タイプ	例および特徴
オープン・モデル	**「Granite」、「Mistral」** – トレーニングは規模や専門性によって異なる – 透明性と説明責任を重視 – 企業やモデルによってオープン性の度合いが異なる – イノベーションのポテンシャルが高い
クローズド・モデル	**カスタム開発の自社モデル** – トレーニングの労力は自社負担 – 対象となるスコープとデータを管理できる – 差別化のポテンシャルが高い
組み込みモデル	**SAP の「Joule」、Salesforce の「Einstein」、Adobe の「Firefly」に搭載されたモデル** – 既存の企業ソフトウェアに組み込まれている – ソフトウェア製品の機能として既存モデルを活用している場合が多い – 通常は単独での使用はできない
一般利用可能なニッチ商用モデル	**Google の「Med-PaLM」** – 大規模な専門的データ・セットでトレーニングされている – 深さと専門性を重視 – 一般的に透明性が低い – 差別化のポテンシャルをある程度持っている
一般利用可能な大規模商用モデル	**「GPT-4」** – 膨大な量のデータ・セットでトレーニングされている – 幅広さと深さを重視 – 一般的に透明性が低い – 差別化のポテンシャルは限定的

リーダーが知るべきこと

万能の生成 AI モデルなど存在しない（続き）

例えば、一般利用可能な大規模商用モデル（例：GPT-4）は、標準的組織で使用されているモデルの約 4 分の 1 を占めるに過ぎないということは、知っておくべきだ。Google の「Med-PaLM」のような一般利用可能なニッチ商用モデルが 23% を占めており、「Granite」や「Mistral」のようなオープン・モデルは 16%、SAP の「Joule」、Salesforce の「Einstein」、Adobe の「Firefly」に搭載されているような組み込みモデルは 14%、組織独自のカスタム開発によるクローズド・モデルは 11% となっている。残りの 12% はその他のモデルである。

モデルのサイズは、どのワークフローにどの生成 AI モデルを利用するかを決める際に、テクノロジー・リーダーが最初に検討する要素の 1 つだ。数千億のパラメーターでトレーニングされた大規模モデルは、能力や知識の幅広さと深さに優れており、より複雑なタスクを扱える。ただし、費用はかさみ、二酸化炭素排出量も多くなる。より小規模でニッチなモデルは、通常数百億のパラメーターでトレーニングされており、コードやコンテンツを特定の言語に翻訳するといった専門的タスクに向けてトレーニングした場合、精度、スピード、効率性が一層向上する。

モデルの所有権も、考慮すべき重要な要素である。公開された商用生成 AI モデルは、標準的な組織の AI ポートフォリオのおよそ半分を占めるほど人気だが、このタイプのモデルには制約がある。どんな組織でも購入やライセンス取得が可能なため、誰もが同一の集積データを用いて作業することになり、競争上の差別化にはあまり役立たないのだ。また、公開モデルは作業のスピードと効率性を上げることには役立つものの、パブリッククラウド上で動作するため、重要なタスクに取り組む際に必要なプライバシー性やコントロール性に欠ける。

そこで登場するのが、企業のクローズド・モデルの生成 AI モデルだ。これらのモデルは使用する組織が自ら開発、所有、管理するため、リーダーがアウトプットの元となるデータを自由に決められる。これにより、モデルやモデルが生み出す成果物が質の悪い情報によって汚染される可能性が低くなるのだ。また、クローズド・モデルを採用することで、ローカル環境とクラウドのどちらでモデルを運用するか、モデルのパフォーマンスのファイン・チューニングのためにユーザーからの提供情報をどのように保管・利用するかを、テクノロジー・リーダーがより柔軟に決められるようになる。その結果、プライベート・データや機密データが不適切に利用・共有されるリスクが軽減される。これは極めて重要な利点である。というのも、不正利用、プライバシー、精度は、経営層が生成 AI モデルを選ぶ際の主たる懸念事項だからだ。

オープン型生成 AI モデルも、こうした懸念事項について対処可能だ。オープンソース開発者コミュニティーの支援によって透明性高く、大きくも小さくも構築できるからだ。オープンに構築されるため、モデルのトレーニングに使用されたデータを知ることができる。また、入念に精査されているため、アウトプットの知的財産権や著作権法の抵触の有無をめぐる問題やリスクに対して、速やかな特定および対処が可能だ。そして、企業がベースのモデルを修正し、カスタマイズすることで、イノベーションの加速、パフォーマンスの改善、生成 AI に対する信頼構築が可能となる。

組み込み型生成 AI モデルは、ソースが多様であり、SAP、Adobe、Salesforce などのプラットフォームやソフトウェアに完全に組み込まれ、ソフトウェアの機能範囲に関わる特定のニーズを満たす。このタイプの生成 AI モデルは、それを組み込んでいる製品に付加価値をもたらすが、単独で使用することはできない。

生成 AI モデルの導入は、今後 3 年間で急速に進む見込みだが、それをけん引するのはオープン・モデルだろう。平均すると、経営層は自社の AI モデル・ポートフォリオに含まれるオープン・モデルが現在より 63% 多くなると見込んでいる。柔軟性、透明性、カスタマイズに対するニーズが追い風となる。経営層はまた、信頼性と拡張性に優れる大規模商用モデルの使用が 27%、専門性を高められるニッチ商用モデルの使用が 31% 増加すると見込んでいる。同期間にクローズド・モデルの使用は 61%、組み込みモデルの使用は 34% 増えると、経営層は予測している。

1. アジリティー + 生成 AI

リーダーが
実行すべきこと

ハンマーとメスを使い分ける

基盤モデルのポートフォリオを評価し、戦略的ワークフローと整合しているか判断する。大規模な生成 AI モデルに投資することで生産性を高めつつ、専門的なタスクにはニッチ・モデルを採用する。

多種多様な生成 AI に目を向ける。LLM、企業のカスタム開発クローズド・モデル、オープン・モデルなどのさまざまな生成 AI モデルの違いを理解する。異なる目的のために異なるモデルに投資する準備を整える。

AI に関する現状を把握する。AI の責任者に、組織で使用中のすべての生成 AI モデルについて、各モデルの目的、機能、パフォーマンス指標をまとめた包括的なカタログを作成させる。AI に関する状況の変化が反映されるよう、リストを定期的に更新する。

最適な組み合わせを見つける。各生成 AI モデルがその強みや、弱み、特徴に基づいて、適切なワークフローと組み合わされていることを確認する。組み合わせのずれを特定する。1 つの辞書で事足りるにもかかわらず、何冊もの百科事典を用いるようなことがあってはならない。

2. コスト + 生成 AI

リーダーが
知るべきこと

生成 AI のコストはすべてコントロールできる

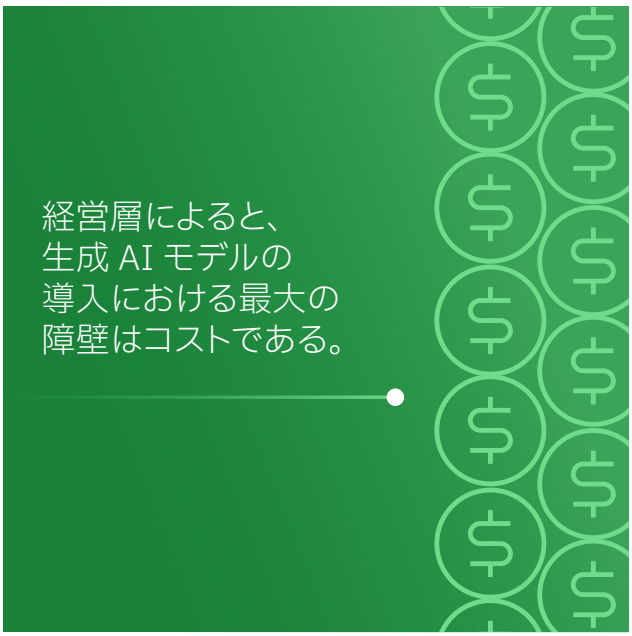
CEO は生成 AI の必要性を認識している。しかし、どれほどのコストをかけるつもりがあるだろうか。生成 AI が組織のあらゆる領域に浸透しようとする中で、ビジネス・リーダーはそれぞれの状況に適したモデルを選定しようとしている。そこでまず検討するのは、コスト効率をいかに大規模に実現できるかという点だ。生成 AI モデルの導入障壁に関する質問において、63% の経営層がモデルのコスト、58% がモデルの複雑性を主な懸念事項として挙げている。

なぜコストがそれほど重大な問題となるのか。それは、使用するモデルによって大きな幅があるからだ。例えば、大規模モデルではデータ・ストレージやコンピューティング・コストが増えるので、クラウド関連費用がかさむ可能性がある。また、大規模モデルは、頻繁なアップデート、ファイン・チューニング、メンテナンスが必要なため、人件費もかかる。一方、ニッチ・モデルはコンピューティング、データ・ストレージ、エネルギーのコストが低い上に、組織の AI ポートフォリオの環境負荷も低減できる。また、迅速な展開が可能で、メンテナンスも少なく済むため、人件費が抑えられる。

適切なタスクに対して適切なサイズのモデルを選択することが、生成 AI のコスト抑制に大きく貢献する。例えば、長文作成、重要な意思決定、研究仮説の検証といった、多くのスキルセットを用い、高い精度が要求される複雑なタスクには、より大規模で高価なモデルが必要となり得る。コスト効率に優れたニッチ・モデルは、より専門的なタスク、中でもスピードと効率性が重要なリアルタイムのチャット・アシスタントや、スパム検出、データ拡張、プロトタイピングなどに適している。推論連鎖などの高度な手法を用いて、複雑な作業をニッチ・モデルが扱える小さなタスクに分けることで、最もコストがかさむ LLM への依存を減らすことができる。

テクノロジーが成熟するにつれ、ニッチ・モデルはより広範なタスクをよりうまく処理できるようになっていくと見込まれる。そのため、組織はコスト管理をより精細に行えるようになるチャンスがある。「目的に合った」モデル、つまり、特定の要件や目標に合致するよう設計、トレーニングされ、有効性が確認されたモデルを用いることで、各タスクに用いるリソースを必要最小限に抑制できる。そして、より専門的なニッチ・モデルのトレーニングに大規模モデルを活用すれば、モデル開発のコスト効率を高めることが可能となる。

近い将来、リーダーはエンタープライズ生成 AI 管理センターを使用して、どのタスクにどのモデルを用いるかの判断を効率化することにより、コスト管理を一層向上させることができるようになるだろう。ユーザーフレンドリーなエクスペリエンス・レイヤーによって、ポートフォリオ全体にわたる各種のモデルや、アシスタント、プロンプトをつなげれば、コスト管理を徹底できる上に、セキュリティー、プライバシー、コンプライアンスのための制御策を組み込むこともできる。これにより、誰がいつ使うかにかかわらず、モデルを適切かつ効率的に使用できるようになる。



2. コスト + 生成 AI

リーダーが
実行すべきこと

自分なりの生成 AI のツボを見つける

モデルを的確に使い分ける価値を認識する。タスクごとに適正なサイズの生成 AI モデルを用いることで、コストを抑え、AI 全体の ROI を向上させる。

モデルに依存しないマインドセットを育む。価格とパフォーマンスの観点から最適なモデルを状況に応じて導入し、精度、必要となるリソース、スピードの間で適切なバランスを取る。

効率性を追求する。モデルの規模を展開先の環境に合わせる。モバイルやリアルタイムのアプリケーションには小規模でスピードに優れたニッチ・モデルを、高い精度を要する複雑なタスクには大規模モデルを優先的に採用する。

無駄をなくす。生成 AI を展開する度に、明確なパフォーマンス指標とベンチマークを設定する。データに基づくインサイト（洞察）を活用することで、生成 AI が意図した価値をもたらしている箇所と、コストを抑制すべき箇所を見極める。

3. 競争力 + 生成 AI

リーダーが
知るべきこと

生成 AI による優位性はいつか消え去る

今日、生成 AI がもたらす競争上の優位性は、近い将来には当たり前のものとなる。人々の生成 AI に関する経験値が上がり、モデル自体もより高性能になる中で、CEO は継続的な改善に注力せねばならない。

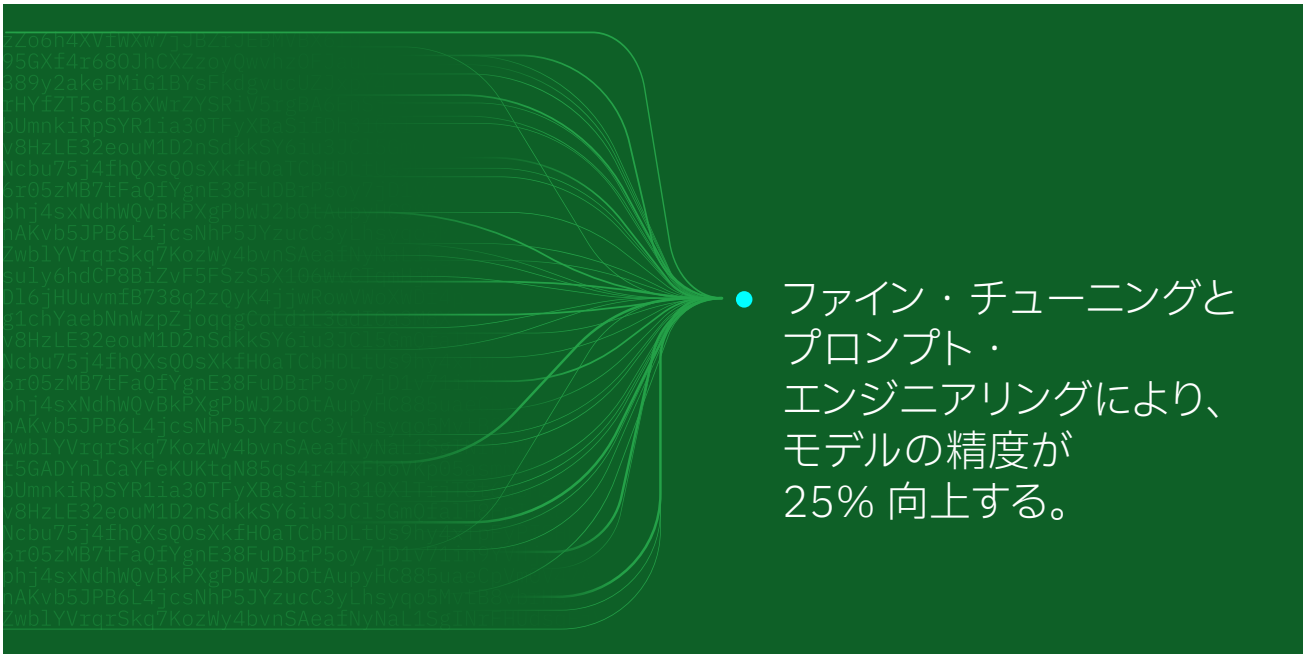
継続的な最適化に力を注ぐ組織は、顕著なパフォーマンス向上を見込める。IBM Institute for Business Value の調査によると、ファイン・チューニングやプロンプト・エンジニアリングを行う組織は、そうでない組織に比べて、モデルのアウトプットの精度がおよそ 25% 高く報告されている。精度の向上は、予測や、リソース配分、パーソナライゼーションの改善につながる。そして、これらすべてが収益増加に寄与する。

ところが、モデルの精度向上のために、プロンプト・エンジニアリング（求めるアウトプットにつながるインプットを設計するプロセス）を常に実践していると回答した経営層は 42% に過ぎない。

ただし、モデルの最適化は対策の 1 つでしかない。ポートフォリオの進化に伴い、モデルのガバナンスも進化させる必要がある。例えば、組織がモデルのリストを運用・管理する方法や、モデルの開発、トレーニング、ファイン・チューニングに関わる権限の付与対象者について、定期的なアップデートを行うべきだ。また、モデルのパフォーマンス指標の追跡や、ドリフト（時を追ってモデルの精度が低下する現象）への対処、モデルのアウトプットに含まれるバイアスの修正のための明確なプロセスも確立しなくてはならない。加えて、急速に変化する規制を順守するための取り組みも必要となる。

さらに、自社の AI インフラの改善、ひいてはハイブリッドクラウド戦略の改善も、継続的に行わなくてはならない。より強力な AI モデルが開発されたら導入するためだ。データの量とモデルの複雑性はいずれも増していくので、テクノロジー・インフラもより多くの処理量に対応できるようにしていく必要がある。拡張性の問題もある。ありとあらゆる形式の生成 AI を使用するチームが増えていくに従い、組織は需要の増加に応えられるよう、インフラやクラウド環境を進化させなくてはならない。

実際のところはどうだろうか。現在、少なくとも半数の組織は、ネットワーク・インフラの最適化、データ処理の高速化、分散コンピューティングに注力している。全体で見ると、経営層の 63% が 1 つ以上のインフラ最適化手法を用いていると回答している。



3. 競争力 + 生成 AI

リーダーが
実行すべきこと

モデルを最大限に活用する

初期の成功に満足してはならない。最新の AI 手法とインフラを用いて、モデルのパフォーマンスの積極的な改善と、競争力の強化を継続的に後押しするべきだ。

生成 AI の水準を引き上げる。プライベートクラウドもしくはオンプレミス環境で、既存の生成 AI モデルに自社データを追加し、独自の価値を創出する。ファイン・チューニングやプロンプト・エンジニアリング、その他の最適化手法を駆使して、競合他社の 3 歩先を行く。

将来を見据えた AI インフラを構築する。クラウドベースのサービスや専用ハードウェア、そしてオープンなフレームワークに投資することで、AI 主導の持続的な大変革を利益につなげられるようにする。

歩みを止めない。明確なガバナンス・フレームワークを構築することで、他社よりも速く生成 AI を進化させる。規制に対する準備状況に抜かりはないか自問し、厳格な基準で自社を評価する。

IBM Institute for Business Value

CEO のための生成 AI 活用ガイド

AI モデルの最適化

本レポートに記載されているインサイトは、IBM Institute for Business Value がオックスフォード・エコノミクス（Oxford Economics）社の協力を得て実施した独自調査に基づいている。調査は 2024 年 6 月に実施され、米国を拠点とする企業の経営層 200 人に AI モデルの最適化に関して質問を行った。

IBM Institute for Business Value

IBM Institute for Business Value（IBV）は、20 年以上にわたって IBM のソート・リーダーシップ・シンクタンクとしての役割を担い、ビジネス・リーダーの意思決定を支援するため、研究と技術に裏付けられた戦略的洞察を提供しています。

IBV は、ビジネスやテクノロジー、社会が交差する特異な立ち位置にあり、毎年、何千もの経営層、消費者、専門家を対象に調査、インタビューおよび意見交換を行い、そこから信頼性が高く、刺激的で実行可能な知見をまとめています。

IBV が発行するニュースレターは、ibm.com/ibv よりお申し込みいただけます。また、LinkedIn（ibm.co/ibv-linkedin）をフォローいただくと、定期的に情報を入手することができます。

ibm.co/ceo-generative-ai-model-optimization

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America | July 2024

本書は英語版「The CEO's guide to generative AI: AI model optimization - Tailor-made gen AI delivers precision power」の日本語訳として提供されるものです。

WPXGRV7D-JPJA-01