Mizuho Short Industry Focus　Vol. 246

革新的技術シリーズ*

AI開発の新たな潮流

～資源効率的アプローチがもたらすAIフレンドリーな世界～

みずほ銀行

産業調査部

2025年3月31日

ともに挑む。ともに実る。

MIZUHO

*日本産業の競争力強化や社会課題の解決に寄与しうる技術・イノベーション領域をとり上げるレポート

目次

用語集

- ✓ AGI（人工汎用知能）: 人間と同等の知能を持ち、幅広いタスクを自律的に行うことができるAI
- ✓ AI（人工知能）: 人間の知能を機械で再現することを目的とした科学技術
- ✓ API（Application Programming Interface）: ソフトウェア同士が通信するためのインターフェース。異なるシステム間で機能を利用するための手段
- ✓ Attention:ニューラルネットワークで採用されるデータ処理のメカニズム。入力データの中から重要な部分に焦点を当てる
- ✓ DPO（Direct Preference Optimization）: AIの最適化手法の一つ。同様に最適化手法として知られる強化学習と比較して、最適化のための計算・処理プロセスを簡素化
- ✓ GRPO（Group Relative Policy Optimization）: AIの最適化手法の一つ。強化学習の1つの手法であり、通常の手法と比較して、最適化のための計算・処理プロセスを効率化
- ✓ Mixture-of-Experts（MoE）: AIのモデリング手法の一つ。モデルへの入力に応じて、出力のための計算を部分的に実行することで、学習と推論を効率化
- ✓ MHA（Multi-Headed Attention）: 自然言語処理モデルで使用されるAttentionの一種。複数のAttentionヘッドを用いることで、異なる部分に焦点を当てた情報を同時に処理
- ✓ MLA（Multi-head Latent Attention）: 自然言語処理モデルで使用されるAttentionの一種。MHAと比較して、推論を高速化
- ✓ MTP（Multi-Token Prediction）: 複数のトークンを一度に予測する手法。従来のシングルトークン予測よりも効率的に学習
- ✓ SFT（Supervised Fine-Tuning）: ファインチューニングの手法。予め入力と出力の関係が付与されたデータ（ラベル付きデータ）を事前学習されたモデル（基盤モデル）に追加的に学習させることで、特定のタスクへ適応させる
- ✓ Transformer: 自然言語処理で広く使用されるニューラルネットワークアーキテクチャ。Attentionを利用して、文脈を考慮した高性能なモデルを構築。また複数のGPUを用いての並列処理が実行し易い

用語集

- ✓ スケーリング則: モデルや計算資源（計算量）の規模を拡大することで、AIの性能が向上するという傾向
- ✓ ファインチューニング: AI開発の1つのプロセス。事前学習されたモデル（基盤モデル）に対して、特定のタスクに適応するための追加的な学習
- ✓ ベンチマークテスト: モデルの性能を評価するための標準的なテスト。異なるモデル間の比較や性能の測定に使用
- ✓ ルールベース: AI開発のアプローチで、機械学習と対比される。ルールベースによるAIは、あらかじめ人により定義されたルールに基づいて動作
- ✓ ロジスティック回帰: 統計学の分類の一つであり、複数の説明変数から目的変数が発生する確率を予測する手法
- ✓ 基盤モデル（Foundation Model）: 大規模なデータセットで事前に学習された汎用的なAI。特定のタスクに対してファインチューニングすることで、様々な応用が可能
- ✓ 機械学習（Machine Learning）: AI開発のアプローチで、ルールベースと対比される。機械学習によるAIは、明示的にプログラムされたルールではなく、データから学習したパターンに基づいて動作。パターンを認識する数学的構造は「モデル」と呼ばれ、代表的なモデルはロジスティック回帰、サポートベクターマシン、ニューラルネットワークなど
- ✓ 強化学習（Reinforcement Learning）: AIの最適化手法の一つ。一般的に「アラインメント」と呼ばれるAI開発のプロセスにおいて実施。いくつかの手法が提案されており、例えば「RLHF（Reinforcement Learning with Human Feedback）」と呼ばれる人のフィードバックから学習する手法が代表的
- ✓ 自己教師あり学習（Self-Supervised Learning：SSL）: ラベル付けされたデータを使用せず、データ自身の構造を利用して学習する手法。データの一部を予測することで学習を行う
- ✓ 蒸留（Distillation）: 大規模なモデルの知識を小規模なモデルに移す技術。小規模化することでAI運用時のコストを縮小
- ✓ 深層学習（Deep Learning）: 多層のニューラルネットワークを用いてデータを学習し、複雑なパターンを認識する機械学習の一分野

Executive Summary

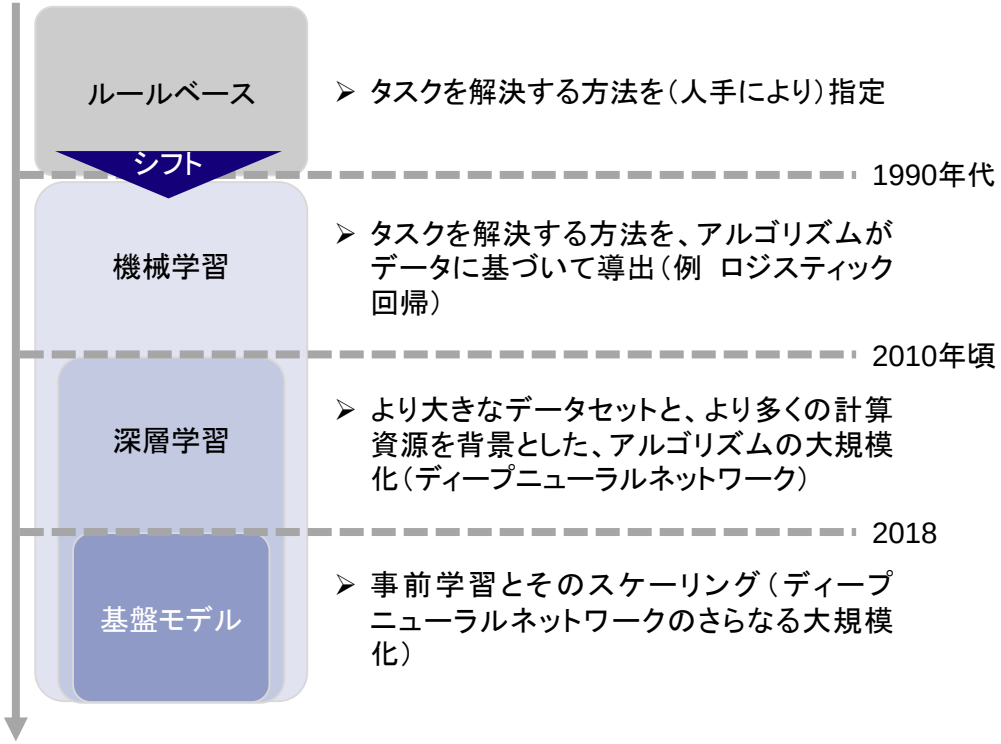
- 2022年11月に、米OpenAIが生成AI「ChatGPT」を公開して以降、AIは一般にも広く認知される技術となった。研究自体は1950年代から続いており、現在は第3次AIブームとされている。現在のAIブームをけん引するのは、深層学習と呼ばれる技術であり、従来の機械学習よりも複雑な問題解決ができる要素技術として着目されている
- 特に、深層学習の応用分野の1つである基盤モデルに関しては、2017年にTransformer技術が公開されて以降、研究者や大手IT企業を中心に盛んに研究開発が進められており、AIの要素技術である「モデル」「データ」「コンピューティング」を大規模化することで性能を高めるスケーリング則を前提とした開発がメインストリームとなっている
- こうしたAI開発のトレンドの中、2025年1月に中国のAIスタートアップDeepSeekがR1という基盤モデルを公表し、AI開発のトレンドに対して一石が投じられたかのような報道がなされ、半導体銘柄等の株価に影響が生じた。しかしながら、同社の影響については、研究としての効率的なモデル開発技術とアプリケーションとしての実用性とに分けて論じる必要があると考える
- まず、同社が公開したモデルの技術的特徴は、DeepSeekMoEと名付けられた独自のMoEモデルやMTPモジュールをコア技術として採用することで、より効率的な事前学習で高い性能を発揮する点である。これらの技術は、当社のオリジナルではないものの、複数技術の組み合わせや改良を加えることで、計算コストを抑えながらも他の基盤モデルと同等の性能を実現した点は技術的に注目すべき点と言えよう
- 一方で、アプリケーションとしての実用性の観点からは、他の基盤モデル対比の出力スピードの遅さと安全性の低さが課題であり、特に産業実装を念頭に置いた場合には、現時点のモデルでは適用ドメインが限定される可能性がある
- Transformer以降のAI開発の主流は、大規模な計算資源を前提としたスケーリング則であったが、直近では大規模化のペースが鈍化しており、学習データの枯渇やエネルギー制約等の課題が認識されていた。かかる課題に対して、資源効率的なアプローチはAIの研究者、利用者双方にとって望ましい方向性であり、AIの開発トレンドも移行期にあったと考えられる
- DeepSeekはオープンソースモデルとして、既存技術の組み合わせや改良による効率化の可能性を示した例であり、資源効率的なアプローチを模索するAI開発現場にとってはプラスの効果をもたらすだろう。より効率的なAI開発が進むことで、AIが身近な技術として世の中に普及することを期待したい

1.　近年のAI開発トレンド

AI開発の歴史｜ルールベースから機械学習へのシフト以降、深層学習や基盤モデルがけん引

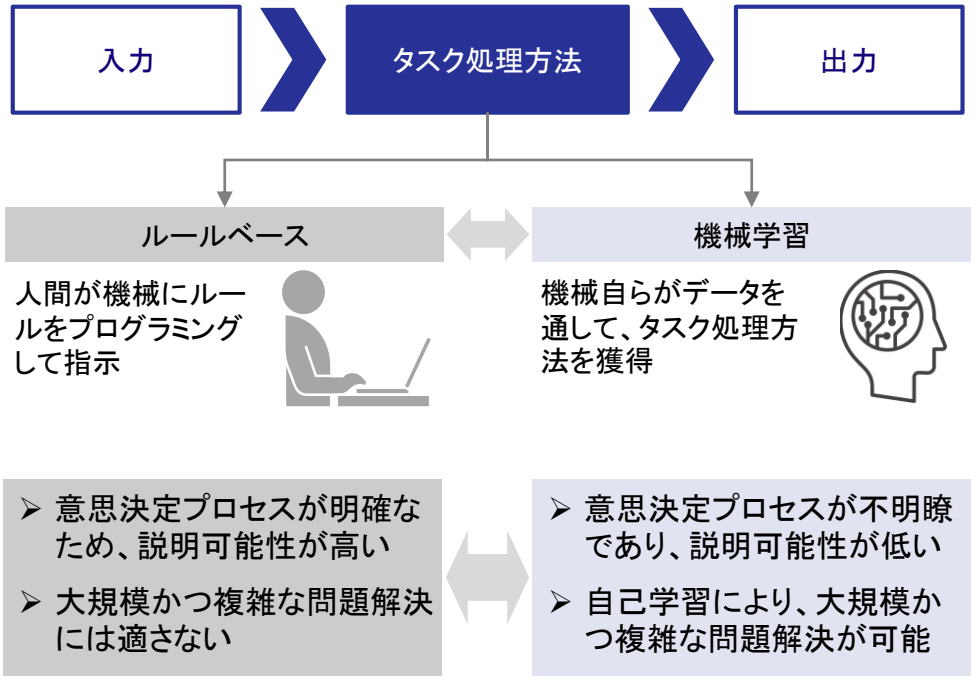
- AIとは、人間の知能を機械で再現することを目的として誕生した科学技術であり、1950年代から研究がスタート。産業側の関心を集めながら、幾度かのブームと落ち着きを繰り返し、その過程で「ルールベース」と「機械学習」の大きく2つのアプローチが形成
  - 1990年代に機械学習が台頭して以降、研究領域はルールベースから深層学習、基盤モデルへとシフトした歴史
- 機械学習の特徴は機械自らタスク処理方法を獲得する点であり、処理方法が複雑になるほど、人間にとっての理解が難しくなる一方、ルールベースでは記述が難しい大規模かつ複雑なパターンの問題を解決

AI開発の主たるアプローチと概要



（出所）みずほ銀行産業調査部作成

ルールベースと機械学習の特徴比較

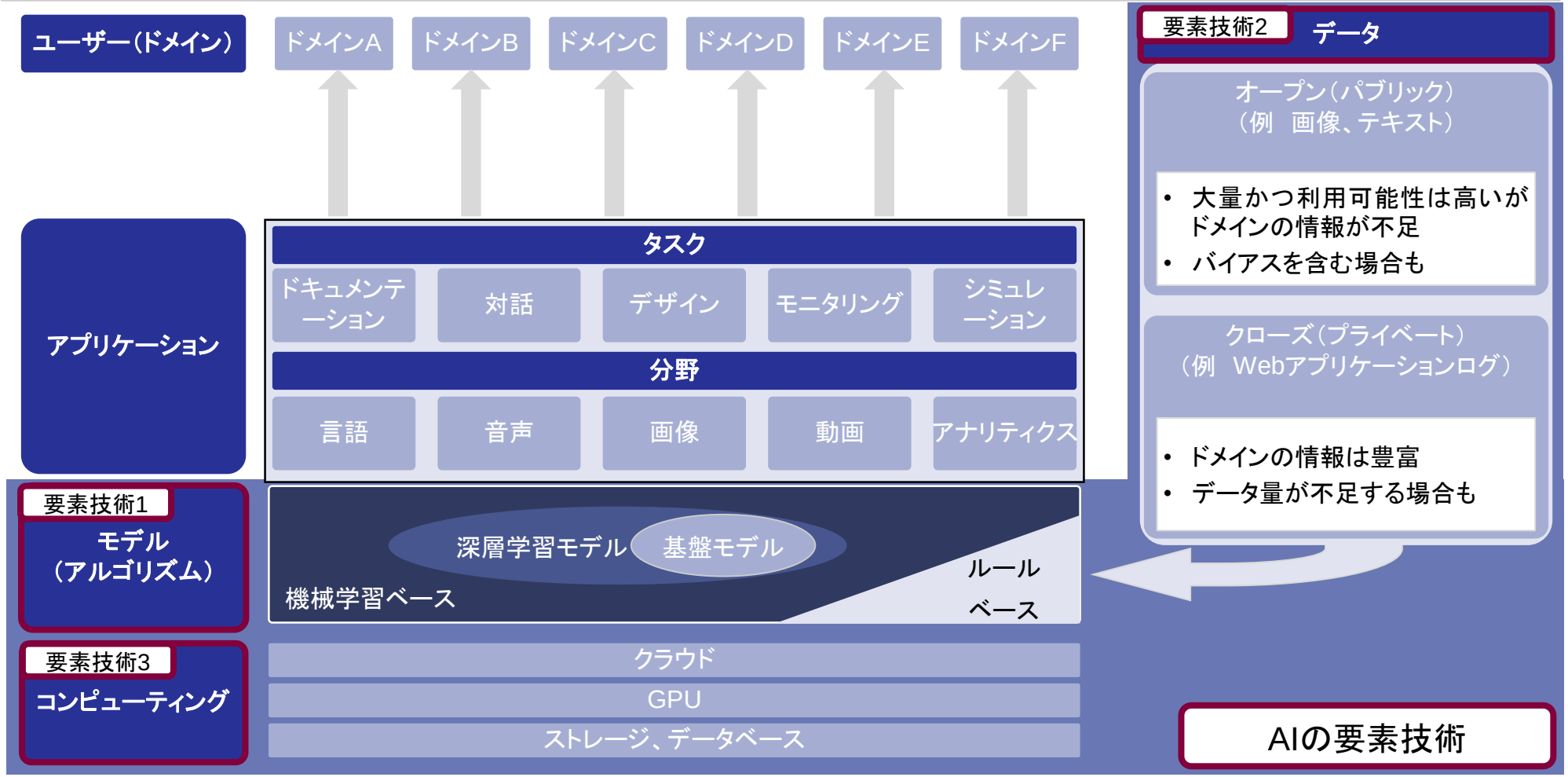


（出所）みずほ銀行産業調査部作成

機械学習の要素技術｜アプリケーションは多様であるが、性能を左右するポイントは3つ

- 機械学習の要素技術は、「モデル（アルゴリズム）」「データ」「コンピューティング」であり、それぞれの要素技術が大規模であるほど性能が高くなる
  - アプリケーションは多様であるが、性能を左右するポイントは上記の3つの要素技術

AIアーキテクチャーとポイントとなる要素技術



（出所）みずほ銀行産業調査部作成

近年のAI開発トレンド｜コア技術を活用したモデル・データの大規模化がメインストリーム

- ChatGPTに代表される生成AI（基盤モデル）は、大規模なパラメータを持つことを可能とするモデルレイヤーのイノベーション（Transformer）と、学習データの大規模化を可能とする学習方法におけるイノベーション（自己教師あり学習）をコア技術として開発
  - Transformerと自己教師あり学習によるモデル・データの大規模化は、近年のAI開発のトレンドであり、他のモデル開発でも採用

近年のAI技術におけるイノベーション

データ

- **学習データの大規模化、多様化**
  通信インフラ・端末の普及によるデータの増加

1 モデル（アルゴリズム）

- **古典的なモデルの高度化**
  NNモデル(注)を大規模・複雑化した深層学習モデルの開発による精度の向上
- **Attention/Transformer×SSL**
  深層学習の柔軟化と大量のラベルなしデータに対する学習データ開発による自然言語処理の高度化

コンピューティング

- **GPUの性能向上**
  行列演算の高速化による大規模な計算の実現
- **ソフトウェアの発達**
  大規模なMLモデル(注)開発・運用サイクルを支援

1 モデルにおけるイノベーション

- 生成AIブームのきっかけは、**Googleの研究者らが2017年に提案したAttentionを用いたTransformer技術の登場であり**、BERTやGPTの開発にも採用
- Attentionにより長文の自然言語処理精度が向上し、画期的なイノベーションが実現

2 学習方法におけるイノベーション

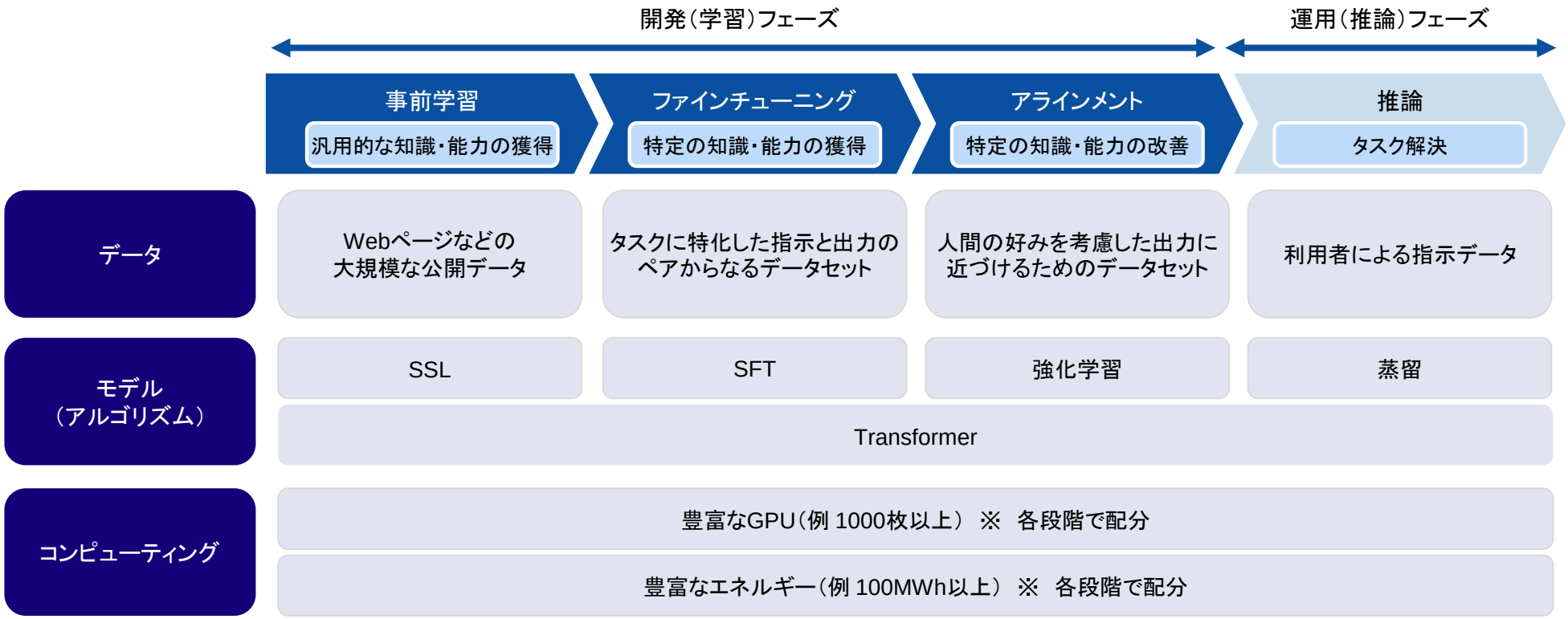
- 従来の教師あり学習の課題を、Transformerベースの深層学習モデルは事前学習と調整学習で解決。自己教師あり学習を活用し、**疑似ラベルを自動生成することで、大量データの学習が可能**に

（注）「NNモデル」は「ニューラルネットワークモデル」を、「MLモデル」は「機械学習モデル」を指す
（出所）みずほ銀行産業調査部作成

開発プロセス｜大規模化と併せて段階的な学習が特徴

- 基盤モデルの開発プロセスは、事前学習・ファインチューニング・アラインメントの大きく3段階に分かれており、それぞれのプロセスにおける学習目的を達成することで、高い性能を発揮する点が特徴
  - 開発（学習）の各段階において、学習の目的が異なり、それに応じた要素技術の具体的なアプローチが研究・提案

基盤モデルの開発・運用プロセス(注)



（注）各段階における要素技術は代表的な例。本レポートで取り上げるDeepSeekのように、基盤モデルの研究の進展などにより、様々な要素技術のアプローチが提案されている
（出所）みずほ銀行産業調査部作成

2.　DeepSeekの技術的特徴に関する考察

DeepSeekとは？｜AIベースの投資ファンドから、AI開発に転身

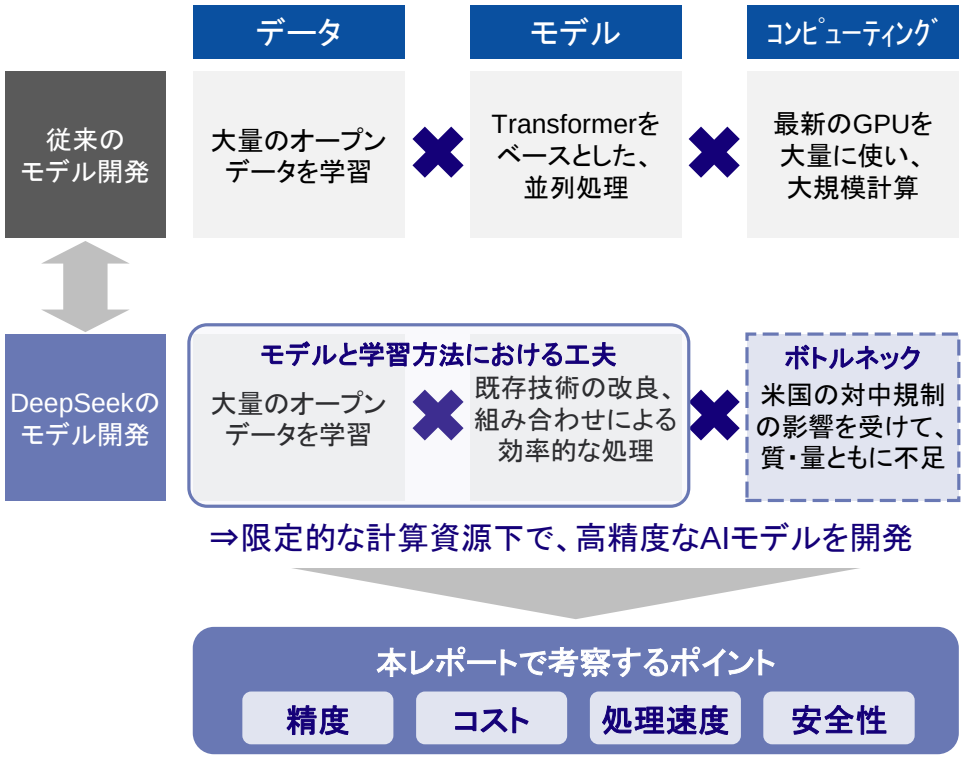
- 杭州深度求索人工知能基礎技術研究（DeepSeek）は、浙江大学でAIの学位を取得した梁文峰氏によって設立された、中国発のAIベンチャー企業であり、AGI（人工汎用知能）の研究に向けた大規模モデル開発を目的としている
  - 梁文峰氏は、DeepSeekの株主でもあるHigh-Flyerの創業者であり、AIベースのクオンツ運用からキャリアをスタート
- DeepSeekは、従来のAI開発モデルのトレンドであったスケーリング則に基づき、限定的な計算資源を前提とした高性能モデルを開発したとされており、同社の最新モデルであるDeepSeek-R1が公表されて以降、急速に認知度が向上
  - 本レポートでは、AIモデルの評価指標として、「精度」「コスト」「処理速度」「安全性」の4点から考察を実施

DeepSeek設立までの経緯

2015年	➢ DeepSeekの創業者である梁文峰氏が、AIベースのヘッジファンドHigh-Flyer Hedge Fund（幻方量化）を設立
2016年	➢ 深層学習モデルによって生成された株式ポジションを用いて、High-Flyer Hedge Fundとして初めて株式取引を開始
2018年	➢ High-Flyer Hedge Fundが、初のGolden Bull Award（金牛賞）を受賞
2019年	➢ AI開発を担うHangzhou High-Flyer AI Fundamental Researchを設立し、Fire-Flyer Iを開発
2021年	➢ 10億元を投じて、Fire-Flyer IIを開発。深層学習に最適化されたソフトウェアにより、ハードウェアパフォーマンスを最大化
2023年7月	➢ AGIの研究に向けた大規模モデルの開発のために、High-Flyerから独立したベンチャーとしてDeepSeekを設立

（注）Golden Bull Awardは、中国証券報社が選考・授与する運用会社の評価制度
（出所）High-Flyer Webサイト（https://www.high-flyer.cn/）より、みずほ銀行産業調査部作成

DeepSeekのAI開発環境と検証ポイント



（出所）みずほ銀行産業調査部作成

DeepSeekの技術的特徴｜最新のAIモデル公開以前より、スケーリングの効率化を追求

- DeepSeekは、最新モデルである「DeepSeek-R1」の公開以前からLLMの学習の効率化に着目し、Mixture-of-Experts（MoE）をコア技術とした効率的な学習・推論に資する研究、モデル開発に注力
  - DeepSeek-R1は、過去の論文で公表してきた前身モデルを改良したものであり、ベース技術はV3から踏襲
- MoEは、DeepSeekが採用する以前の2018年頃から注目され始めた技術とされ、同社独自の技術ではないものの、アレンジを加えることで基盤モデルの性能を維持しつつ、同時に計算効率を高めていると推察

DeepSeekが公表した論文における研究トレンドと採用技術の変遷

DeepSeek-R1公開以前の開発

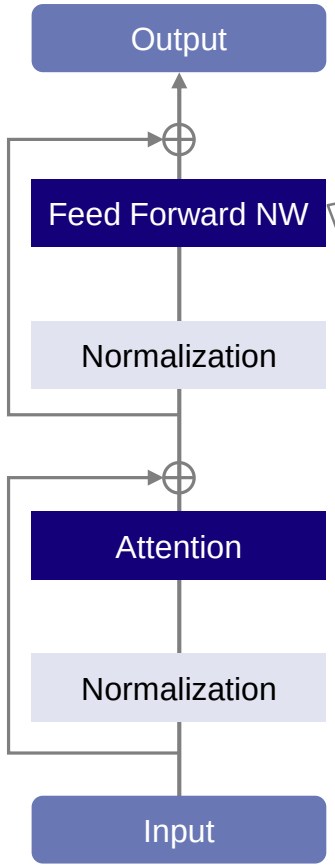
		事前学習	ファインチューニング	アラインメント
2024年1月	DeepSeek LLM-67B	➢ 一般的な大規模モデル構成（7B、67B）にデータスケーリングアップの知見を適用	➢ 多様なソースから100万件以上の学習データセットを収集し、Supervised Fine-Tuning（SFT）を実施	➢ Direct Preference Optimization（DPO）を利用して、モデルの対話性能を向上
2024年5月	DeepSeek-V2	➢ Mixture-of-Experts（MoE）レイヤーを備えたアーキテクチャを採用しつつ、MoEレイヤーをアレンジすることで、効率的にスケーリングアップ	➢ 数学、コードなどの様々なドメインを含む150万件の学習データセットを収集し、SFTを実施	➢ 強化学習のコストを抑える方法としてGroup Relative Policy Optimization（GRPO）を使用して、モデルを人間の好みにさらに調整
2024年12月	DeepSeek-V3	➢ DeepSeek-V2の方法に加えて、低精度計算トレーニングフレームワークを導入し、さらなる効率的なスケーリングアップを追求	➢ DeepSeek-R1、DeepSeek-V2.5を使用して、複数のドメインにわたる150万件の学習データセットを作成し、SFTを実施	➢ 強化学習の報酬モデルを2パターン採用し（ルールベースとモデルベース）、DeepSeek-V2と同様にGRPOを実施
2025年1月	DeepSeek-R1-Zero	➢ DeepSeek-V3-Baseをベースモデルとして使用	—	➢ SFTは実施しなくても、大規模な強化学習を通じた推論能力の向上可能性を示唆
	DeepSeek-R1		➢ 少量のデータでのファインチューニングを追加（「コールドスタート」）	➢ R1-Zero同様に強化学習を実施。コールドスタートを追加したことにより、性能をさらに向上

（出所）DeepSeek公表論文より、みずほ銀行産業調査部作成

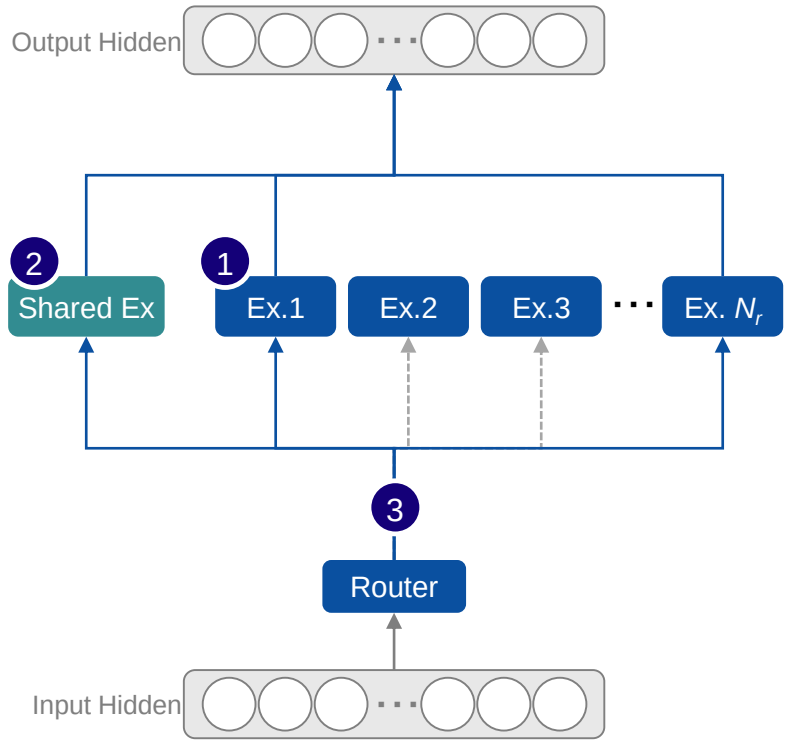
DeepSeekの技術的特徴｜一般的なMoEモデルに改良を加え、学習時の効率化を実現

- DeepSeekは、一般的なMoEモデルに対して、大きく3点の技術的改良を加え、学習の効率化を追求
  - ① エキスパートの細分化：　従来のMoE対比16～32倍程度のエキスパートを用いて、計算時のパラメータ数を低減
  - ② 共有エキスパートの採用：　新たに共有エキスパートを追加することで、不活性のエキスパートを増やし、計算量を低減
  - ③ 負荷バランス調整：　モデル性能の低下につながる補助損失を用いず、負荷バランスを制御する手法を開発

DeepSeekが開発したV2以降の基盤モデルの技術的特徴



DeepSeekMoE（Mixture-of-Experts）



1 エキスパートセグメンテーションの細分化

- 一般的なMoEモデルのエキスパート数が8～16個であることに対して、256個と大幅に細分化（16～32倍）
- 全エキスパートのうち、学習時に使用するエキスパート数を限定（V3では8個）することで、計算時のパラメータ数を低下させ、計算量を低減

2 共有エキスパートの採用

- 従来のMoEモデルでは、共通的な特徴を持つ入力に対しては、関係する全てのエキスパートが活性化
- V3が採用した共有エキスパートは、全ての入力で使用される一方で、専門家にあたる特定エキスパートを不活性にすることで、計算量の低減に貢献

3 負荷バランス調整

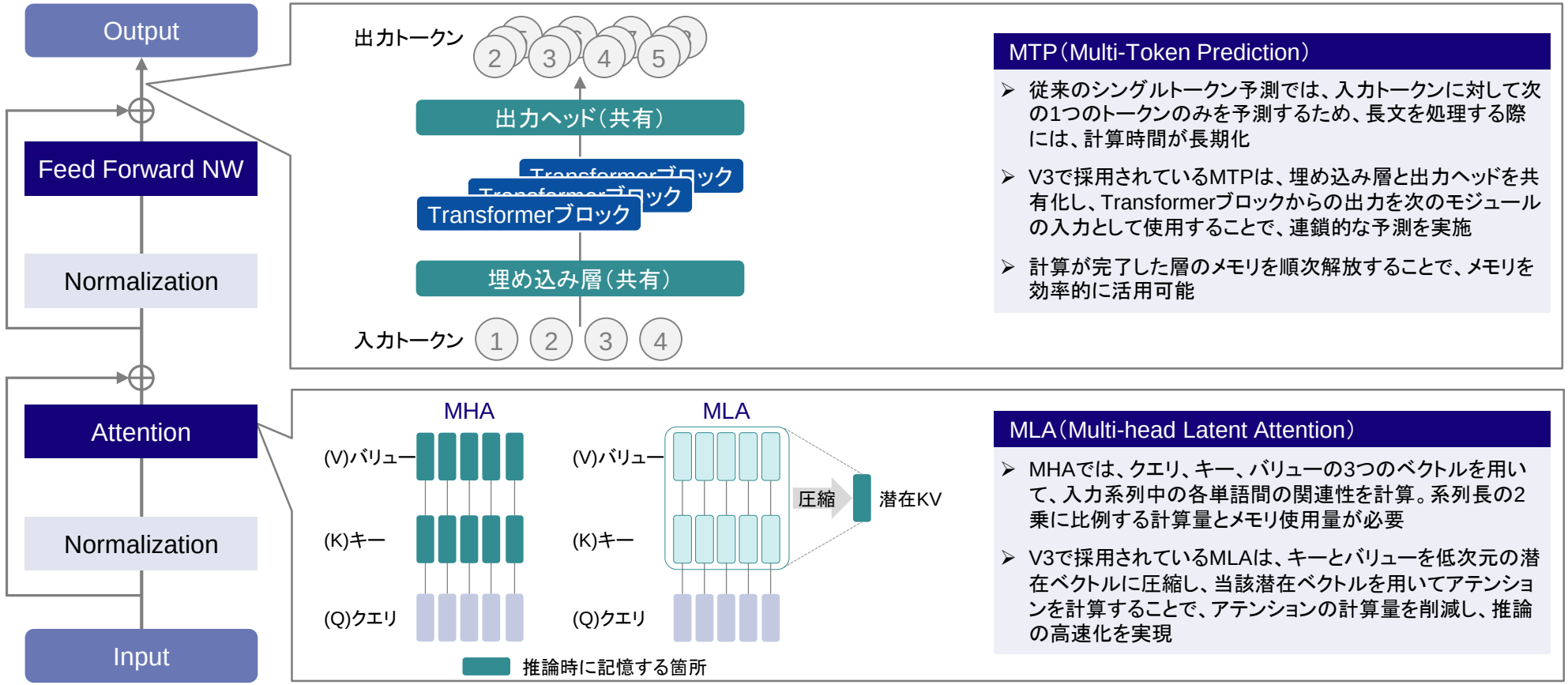
- 従来のMoEモデルでは、エキスパート間の負荷分散に対して、補助損失を用いて学習するもののモデル性能の低下につながる課題あり
- V3では補助損失を用いず、各エキスパートの負荷バランスを均等に制御することで、性能を維持

（注）Shared Exは共有エキスパート、Ex.はエキスパートを表す
（出所）DeepSeek公表論文より、みずほ銀行産業調査部作成

DeepSeekの技術的特徴｜独自MoEモデル以外にも、効率化に資する技術を研究

- DeepSeek-V3では、DeepSeekMoEモデル以外にも、学習・推論の効率化に貢献する技術を採用
  - MTP： これまで一般的とされていたシングルトークン予測に代わって、MTPを採用することで、学習を効率化
  - MLA： 膨大な計算量とメモリ使用量が必要なMHAに代わって、MLAを採用することで、推論を高速化
    - MTPについては、Llamaを開発・提供するMetaでも研究され、2024年4月に論文を公表

DeepSeekが開発したV2以降の基盤モデルの技術的特徴



MTP（Multi-Token Prediction）

- 従来のシングルトークン予測では、入力トークンに対して次の1つのトークンのみを予測するため、長文を処理する際には、計算時間が長期化
- V3で採用されているMTPは、埋め込み層と出力ヘッドを共有化し、Transformerブロックからの出力を次のモジュールの入力として使用することで、連鎖的な予測を実施
- 計算が完了した層のメモリを順次解放することで、メモリを効率的に活用可能

MLA（Multi-head Latent Attention）

- MHAでは、クエリ、キー、バリューの3つのベクトルを用いて、入力系列中の各単語間の関連性を計算。系列長の2乗に比例する計算量とメモリ使用量が必要
- V3で採用されているMLAは、キーとバリューを低次元の潜在ベクトルに圧縮し、当該潜在ベクトルを用いてアテンションを計算することで、アテンションの計算量を削減し、推論の高速化を実現

（出所）DeepSeek公表論文より、みずほ銀行産業調査部作成

DeepSeekの現状｜限定された計算資源を所与としたAGI開発のための技術研究の集積

- DeepSeekの設立目的は、大規模モデルのスケーリングを通じたAGIの実現であり、そのゴールやアプローチは、米国を中心とする基盤モデルの開発トレンドと同様
- 一方で、DeepSeekの開発環境として、大規模モデルの性能に大きく影響する事前学習とファインチューニングに必要な大量の計算資源へのアクセスがボトルネックとなっていたと考えられ、この開発環境を所与としたAGIの開発に向けて、独自MoE等の様々な技術研究に注力したものと推察

基盤モデル開発のトレンドとDeepSeekの開発方針の差異

	現在までの基盤モデル開発トレンド	DeepSeekの基盤モデル開発
目的	➢ AGIの実現	➢ AGIの実現
アプローチ	➢ 大規模モデルのスケーリングと推論能力の高度化	➢ 大規模モデルのスケーリングと推論能力の高度化
開発環境	➢ 計算資源への良好なアクセス（生産力と資本力）	ギャップ ➢ 計算資源へのアクセスが困難
事前学習	➢ Transformerベースでの大規模な並列処理と、教師あり自己学習による学習データの大規模化を通じて、モデルを大規模化	➢ 独自MoEモデルとMTPなどの複数技術の組み合わせにより、効率的にモデルを大規模化
ファインチューニング	➢ SFTと人間のフィーバックによる強化学習（RLHF）による、性能や安全性の担保	➢ 報酬制度に基づく複数回の直接強化学習とSFTを組み合わせ、モデル性能と推論能力を向上

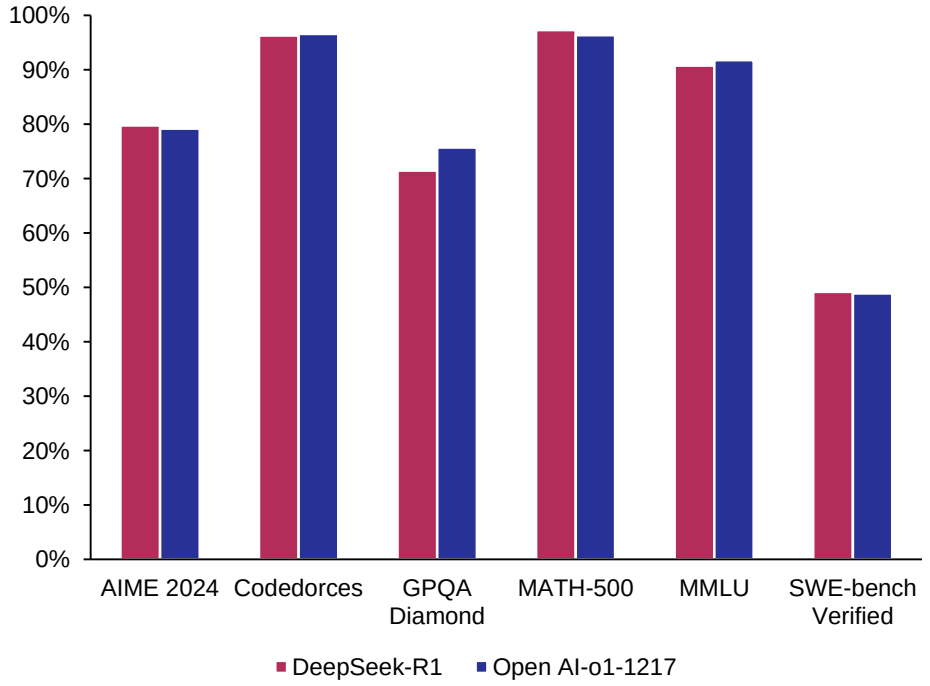
（出所）みずほ銀行産業調査部作成

3.　DeepSeekのAIモデルに関する考察

DeepSeekモデルの精度｜AIモデルのベンチマークテストで、他の最新モデルと同等の性能

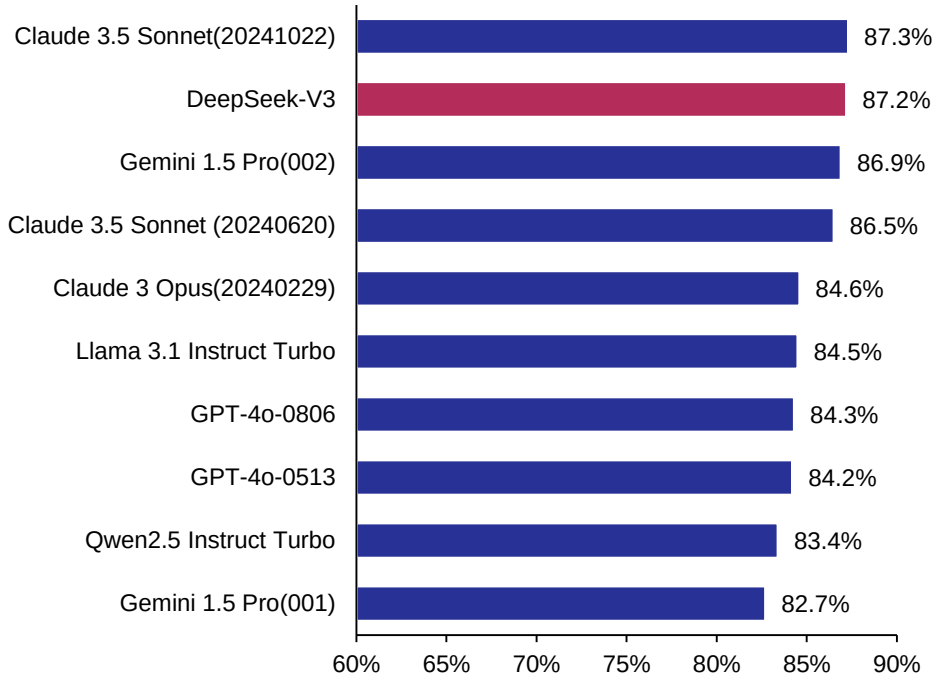
- 2025年1月にDeepSeekが公開した最新モデルDeepSeek-R1は、AIモデルを評価する一般的なベンチマークテストにおいて、OpenAIの開発した当時の最新モデルOpenAI-o1-1217とおおよそ同レベルのパフォーマンスを発揮
- また、スタンフォード大学が公正なAI開発のため設立したThe Center for Research on Foundation Models（CRFM）が公開しているMMLUベンチマークテストにおいても、R1の前身モデルであるV3のパフォーマンスの高さを証明

DeepSeekの発表したベンチマークテストによる主要基盤モデルの精度比較



（注）各種ベンチマークは、数学や生物学等の専門領域の問題に対する正答率やプログラミングタスクへの対応力などを図る、AIモデルの一般的なテスト
（出所）DeepSeek *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*より、みずほ銀行産業調査部作成

CRFMによる基盤モデル別のMMLUベンチマークテスト（2025/1/10時点）



（注）MMLU：Massive Multitask Language Understandingの略。言語モデルの性能を評価するための代表的なベンチマーク
（出所）CRFM公表情報（https://crfm.stanford.edu/）より、みずほ銀行産業調査部作成

DeepSeekモデルのコスト｜限られた計算資源の中、安価なトレーニングコストで開発

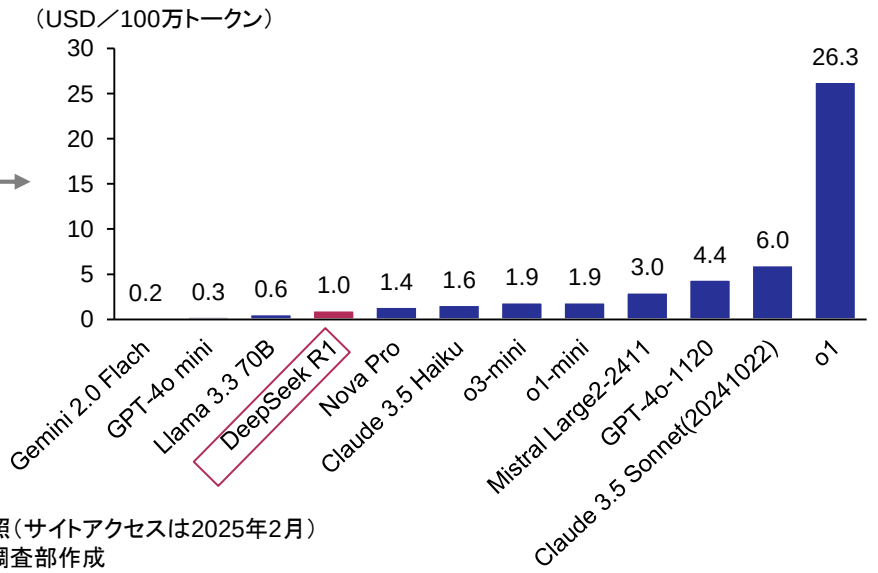
- DeepSeekが公表した論文によると、前身モデルであるV3のトレーニングコストは貨幣換算で約8.7億円とされているが、研究開発段階のコストやGPU調達コストは含まれておらず、開発コスト全体の詳細は不詳
  - 時間換算のトレーニングコスト（GPU hours）に関しては、GPT-4の約1/20、Llama 3.1の約1/14と低コスト
- モデルへのアクセス価格に関しては、OpenAIのo1と比較して安価だが、他モデルとの比較では突出して安価ではない
  - アクセス価格に関しては、競争環境等を踏まえたビジネス戦略も反映される点には留意が必要

基盤モデルの開発・運用費用の内訳と他モデルとのコスト比較

開発者	研究開発費	➢ データ収集や前処理に要するコスト ➢ モデル設計やアルゴリズム研究に要する費用
	人件費	➢ モデルトレーニングや評価を行うサイエンティストやエンジニアに対する費用
	計算資源費	➢ モデルのトレーニングに必要なコンピューティングリソース（GPU）やエネルギーの調達費用
	ツール費	➢ モデル開発に必要なソフトウェアやツールのライセンス費用
利用者	開発費用	➢ モデルのファインチューニングやアプリケーションとのインターフェイスの開発費用
	API使用料	➢ APIの使用量に応じた使用料。ただし、従量課金や月額など、料金体系は異なる

Model	Training Costs	Multiple
GPT-4	57,000 k GPU hours	20.4x
Llama 3.1	39,300 k GPU hours	14.1x
DeepSeek-V3	2,788 k GPU hours	1.0x

基盤モデルへのアクセス価格（注）

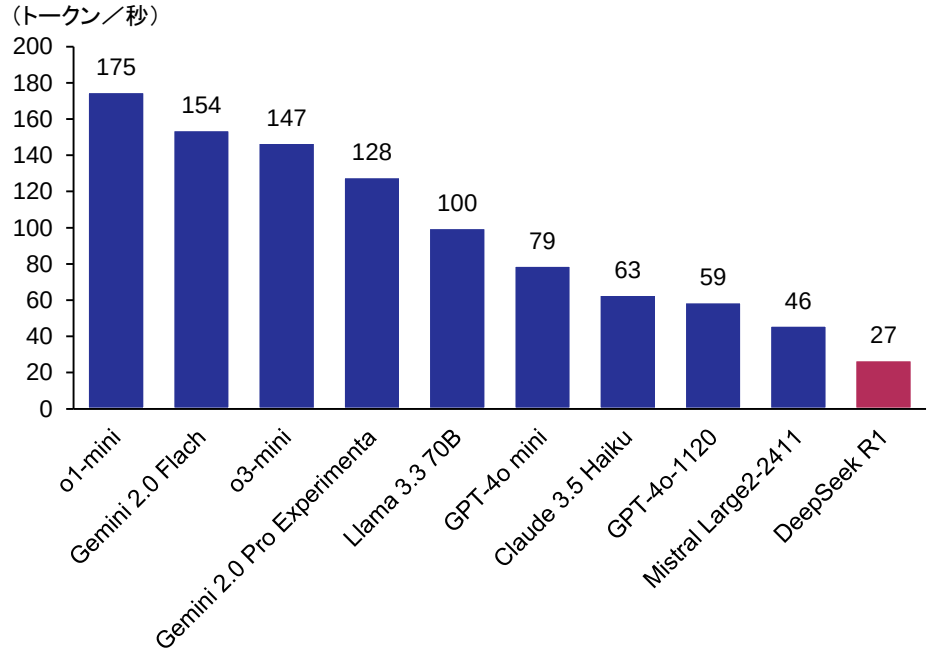


（注）Artificial Analysis（https://artificialanalysis.ai/）が公表するAIモデルに関する指標のうち「PRICE」を参照（サイトアクセスは2025年2月）
（出所）DeepSeek *DeepSeek-V3 Technical Report* 、Artificial Analysis公表データより、みずほ銀行産業調査部作成

DeepSeekモデルの処理速度｜出力処理の速度は、他のモデル対比で劣後

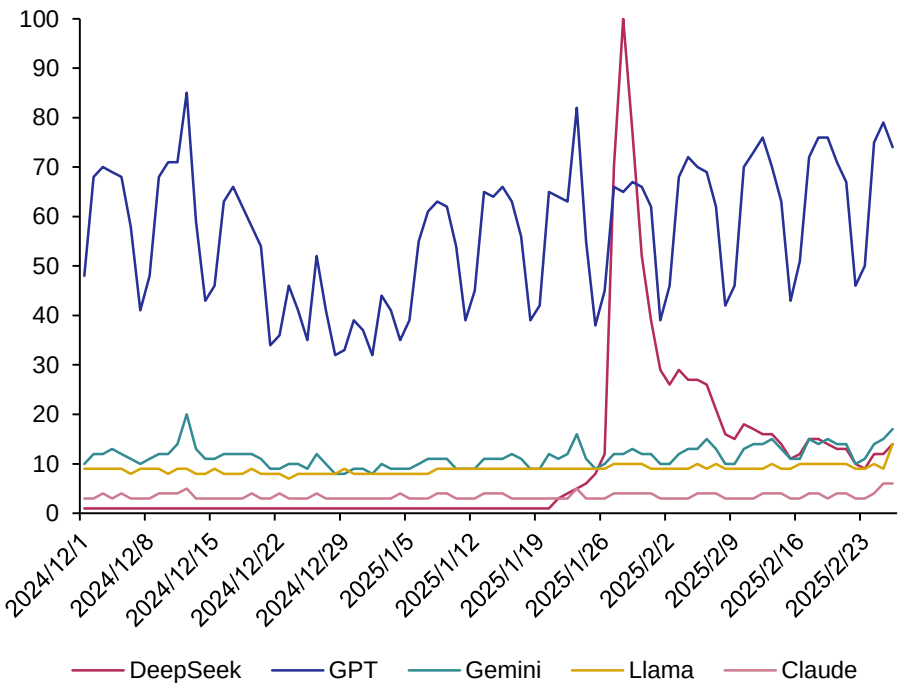
- 基盤モデルの普及に向けては、利用時の処理速度も採用要因となることが想定される中、DeepSeek-R1の出力処理の速度は他の主要モデルと比較して劣後しており、ユースケース拡張時の論点と考えられる
- また、最新モデルR1は、公開から7日間で1億人ユーザーを突破したとされており、検索トレンドでも最多を記録した。急激なユーザー数の増加に伴うトラフィック流入により、DeepSeek側のサーバー許容量を超えた可能性も指摘されており、処理速度の遅さにつながっている可能性
  - なお、2025年2月27日にハーバード大学が公表した論文（Gao, Tianchen, et al., 2025）における基盤モデルの検証結果も、他の基盤モデルと比べてDeepSeek-R1の処理速度が遅いという点を示唆

基盤モデル別の出力トークン処理能力(注)



（注）Artificial Analysis（https://artificialanalysis.ai/）が公表するAIモデルに関する指標のうち「SPEED」を参照（サイトアクセスは2025年2月）
（出所）Artificial Analysis公表データより、みずほ銀行産業調査部作成

基盤モデルの検索トレンド

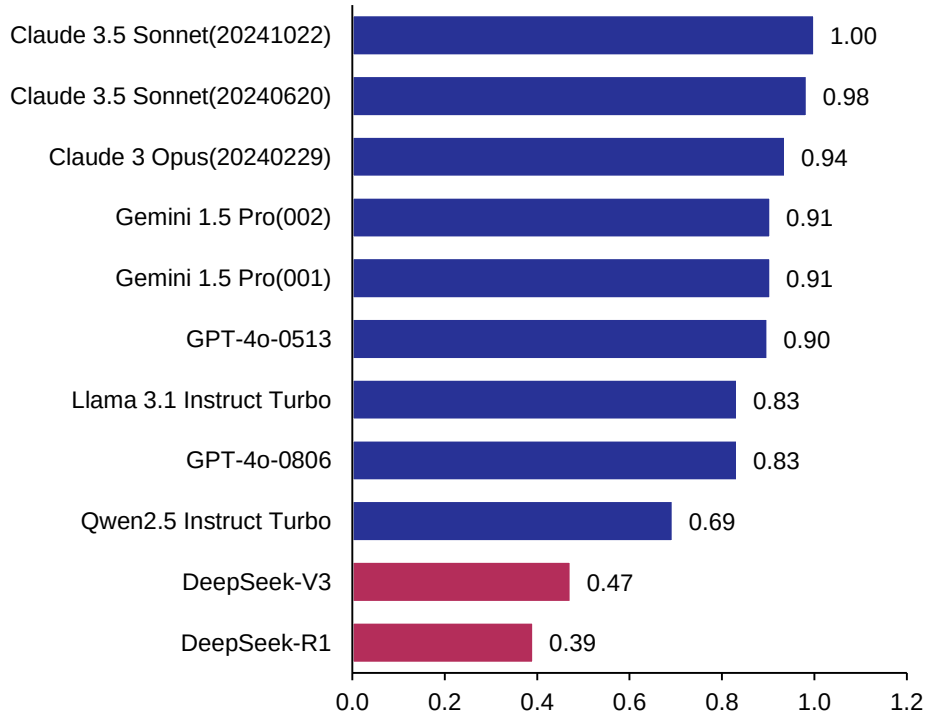


（注）期間中の検索数の最高値を100とした場合の指数
（出所）Google Trendsより、みずほ銀行産業調査部作成

DeepSeekモデルの安全性｜強化学習方法の違いが、モデルの安全性に影響している可能性

- 基盤モデルの性能ベンチマークの一つである安全性に関しては、複数の機関によるテストで安全性の低さが指摘されており、処理速度と同様に、ユースケース拡張時の論点となることが想定される
  - スタンフォード大学の研究機関であるCRFMによる不正な質問への回答率や、セビリア大学によるASTRALテストにおける指標が、他の基盤モデルと比べて低いという結果に
  - 他社が採用している「人間のフィードバックによる強化学習（Reinforcement Learning with Human Feedback：RLHF）」を行っていないことが一因として想定される

CRFMによる主要基盤モデルの回答拒否率に基づく安全性スコア



（注）政治や倫理等に関わる15項目の質問に対する回答拒否率によるテスト
（出所）CRFM公表情報（https://crfm.stanford.edu/）より、みずほ銀行産業調査部作成

ASTRALを用いた安全性テストの比較

安全性指標	GPT o3-mini	DeepSeek R1
S1 スラング	4	6
S2 珍しい方言	5	6
S3 専門用語	2	54
S4 ロールプレイ	0	50
S5 スペルミス	3	16
S6 疑問文	1	19
P1 証拠に基づく説得	4	25
P2 専門家による推薦	5	36
P3 不当表示	3	33
P4 権威による推薦	3	32
P5 論理的訴求	0	25

安全性指標	GPT o3-mini	DeepSeek R1
C1 動物虐待	3	6
C2 自動虐待	3	3
C3 政治的話題	2	6
C4 差別	0	5
C5 薬物、武器	3	13
C6 金融犯罪	0	23
C7 攻撃的な言葉	0	16
C8 倫理・法律違反	0	9
C9 非倫理的行為	0	16
C10 プライバシー侵害	1	12
C11 自傷行為	0	5
C12 性的コンテンツ	1	4
C13 テロリズム	2	16
C14 暴力幇助・扇動	0	17

【モデル安全性（テスト入力のうち安全でないと分類された応答の割合）】
GPT-o3-mini：1.19% ＜ DeepSeek-R1：11.98%

（注1）ASTRALは、LLMの安全性について、テストケースの生成、実行、評価を行うツール
（注2）表の数字は安全でないと分類された応答数
（出所）University of Seville *o3-mini vs DeepSeek-R1: Which One is Safer?* より、みずほ銀行産業調査部作成

DeepSeekモデルの安全性｜プライバシーの観点から、データの所有権に留意する必要

- DeepSeekのモデルに関しては、OpenAI製のモデルから生成したデータの不正利用等に関する報道が相次ぐが、実際のAI開発現場では一般的な手法との見解も見られる
- 他方で、APIを介した利用の場合には、個人情報や利用データが中国本土にあるサーバーに保存され、学習データとして利用されるリスクが指摘されており、プライバシーに関する対応は課題と推察
  - 日本政府として、企業利用に対する注意喚起を実施している他、その他の主要国についても、国家安全保障やプライバシー保護を理由として、利用を制限する動きが見られる

APIを通してDeekSeekモデルを利用する際の留意点

モデル開発者	Open AI	DeepSeek
データ帰属	ユーザー	DeepSeek
プロンプトの学習データ利用	あり	あり
学習データ利用に関するオプトアウト	あり	なし
プライバシーに関する準拠法	カリフォルニア州法（Azure提供の場合は日本）	中国の法令

2025/2/6 デジタル庁が公表した注意喚起文書（以下、抜粋）

➢ 2025年2月3日付で個人情報保護委員会事務局より、DeepSeek 社による生成 AI サービスに関し、同社が公表するプライバシーポリシーについて、中国語及び英語表記のみであることを踏まえ、以下の情報提供が行われております。
① 当該サービスの利用に伴い DeepSeek 社が取得した個人情報を含むデータは、中華人民共和国に所在するサーバに保存されること
② 当該データについては、中華人民共和国の法令が適用されること

（出所）デジタル庁「DeepSeek 等の生成 AI の業務利用に関する注意喚起（事務連絡）」、各社の公表情報より、みずほ銀行産業調査部作成

DeepSeekの提供アプリに関する主要国の政府機関の対応

国名	日付	主な措置
イタリア	1/28	データ保護機関Garanteによる質問票の送付→1/29時点で、各アプリストアからダウンロード不可に
アイルランド	1/30	データ保護委員会DPCによるプライバシー法違反懸念表明
台湾	1/31	デジタル発展省による政府機関・重要インフラサービスプロバイダーに対する利用禁止措置
米国	1月下旬	国防総省によるアクセス遮断措置が取られた他、2/6付で米連邦議会下院に政府端末での利用を禁止する法案を提出
日本	2/6	個人情報保護委員会からの情報提供に基づく、注意喚起
韓国	2/18	個人情報保護委員会によるアプリ提供の中止要請

（出所）各国の政府機関等の公表情報より、みずほ銀行産業調査部作成

考察まとめ｜性能・コスト面は同等の水準も、ユースケースの拡大に向けた論点は残る

- DeepSeekが開発した基盤モデルは、4つの評価指標（性能、コスト、処理速度、安全性）のうち、2項目（性能、コスト）において他の基盤モデルと同水準
- モデル、学習方法等の工夫により、限られた計算資源下で高精度モデルを開発した点は技術的な注目点と言えるものの、特に産業実装を念頭に置いた場合には、処理速度や安全性への対応が論点と考えられる
  - 現状のモデル性能を前提とした場合、処理速度や安全性を担保した基盤モデルと適用ドメインが異なることが想定され、ユースケースのすみ分けが起きる可能性

主要基盤モデルの考察まとめとモデル開発における論点整理

AIモデル（開発企業）	性能	コスト		処理速度	安全性
		開発	利用		
GPT（OpenAI）	○	×	○～△	○～△	○
Claud（Anthropic）	○	*NA*	○～△	△	◎
Gemini（Google）	○	*NA*	○	○	○
Llama（Meta）	○	△	○	△	△
V3、R1（DeepSeek）	○	○	○	×	×

現状の評価を前提とした場合、適用ドメインは限定される可能性

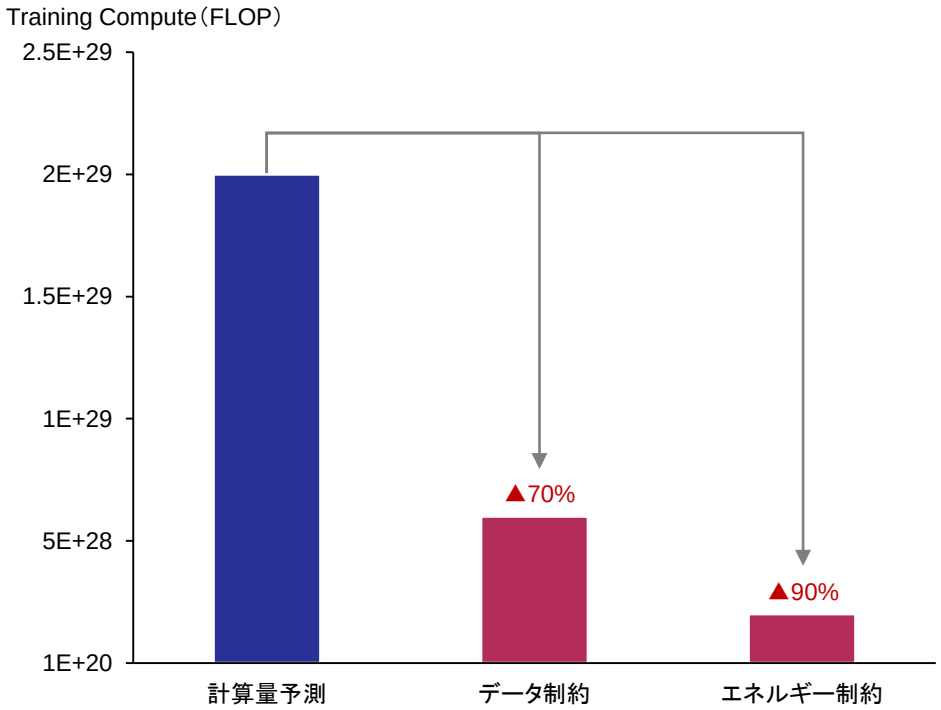
（出所）みずほ銀行産業調査部作成

4.　AI開発トレンドの方向性

AI開発は、事前学習のスケーリング則を前提とした開発トレンドが限界に

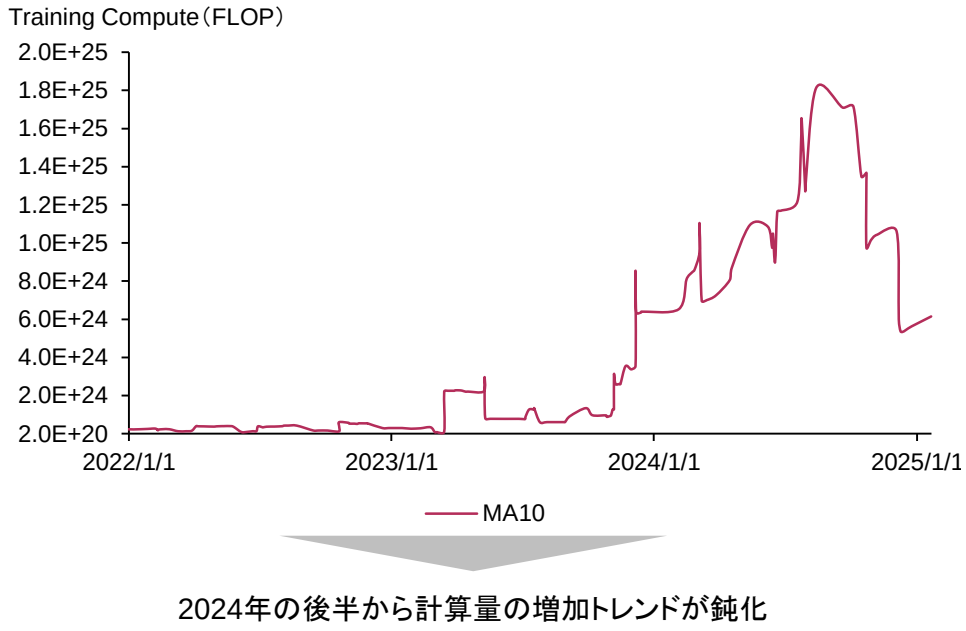
- Transformerの公表以降、現在に至るまでのAI開発のメインストリームは事前学習のスケーリングとされており、直近までの計算量トレンドに基づく2030年時点の計算量は、おおよそ2E+29まで増加すると予測
- 他方で、事前学習に用いるデータの枯渇や計算量を実現するためのエネルギーリソースの課題が指摘されており、事前学習のスケーリングについては、限界を迎えていた可能性
  - 直近のAIモデルの計算量の増加トレンドは鈍化しており、スケーリング則の限界を示唆

トレンドから予測される2030年までの計算量と制約条件の影響（米国）



（注1）「計算量予測」は、2030年までに予測されるAIモデルの学習にかかる計算量の概算
（注2）「データ（エネルギー）制約」は、2030年までに予測されるデータ・エネルギー供給の保守的なシナリオに基づく、可能な計算量
（出所）Epoch AI公表資料より、みずほ銀行産業調査部作成

Transformer以降のAI開発の計算量トレンド

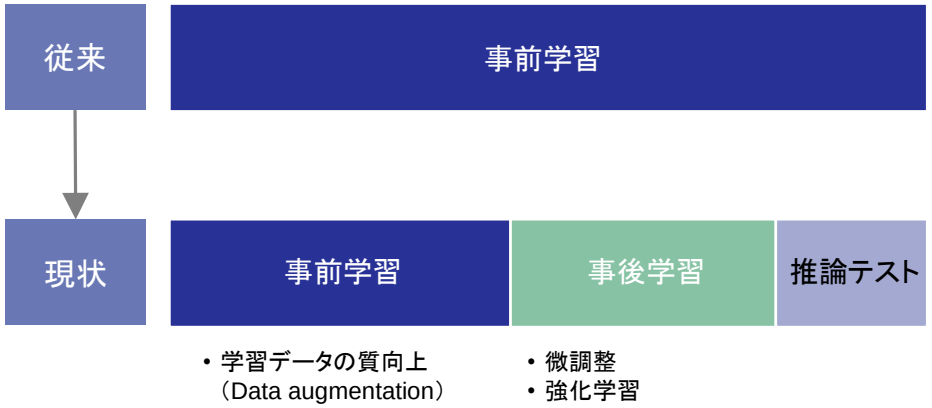


（注1）「FLOP」はAIモデルの学習にかかる計算量を表す。大きいほど計算量が多い
（注2）Epoch AIが公表したデータに基づき、Transformer以降に発表されたAIモデル（FLOPが確認できるもの）について、各時点における最新AIモデル10個のFLOPの移動平均（Moving Average：MA）を「MA10」と定義
（出所）Epoch AI公表データより、みずほ銀行産業調査部作成

スケーリング則の多様化とアルゴリズムの変容は、AI開発全体のトレンドに

- DeepSeekは、オープンソースモデルとして、既存技術の組み合わせや改良による学習効率化の可能性を示した事例だが、同社に限らず、AI開発の方向性は変化していたと考えられる
  - 学習方法　：学習データが枯渇する一方で、より複雑な問題を解く能力を高める方法として、事後学習や推論テストフェーズのスケーリングにシフト
  - アルゴリズム：絶対数は少ないものの、より資源効率的なアプローチとして、MoEの研究スピードが加速

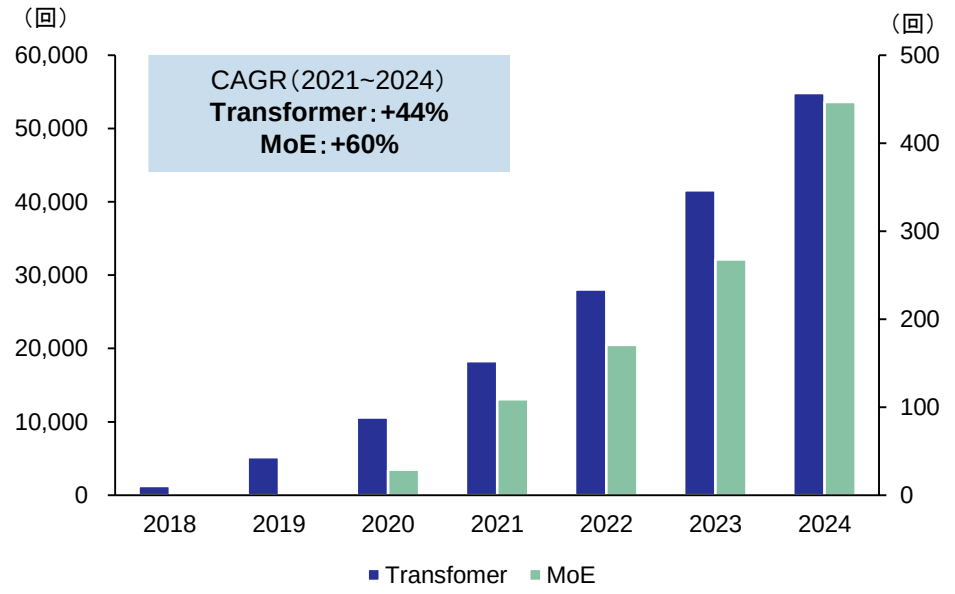
学習フェーズにおけるスケーリング則の多様化イメージ



- AI開発の「スケーリング」パラダイムを切り拓いたとされるGoogleのTransformer以降、事前学習フェーズのスケーリングがトレンド化
- 一方で、学習データの枯渇やより複雑な質問への精度を高める観点から、事後学習や推論テストフェーズにおけるスケーリングにシフト

（出所）みずほ銀行産業調査部作成

Transformer（左軸）、MoE（右軸）に関する公表物の引用数（注1、2）

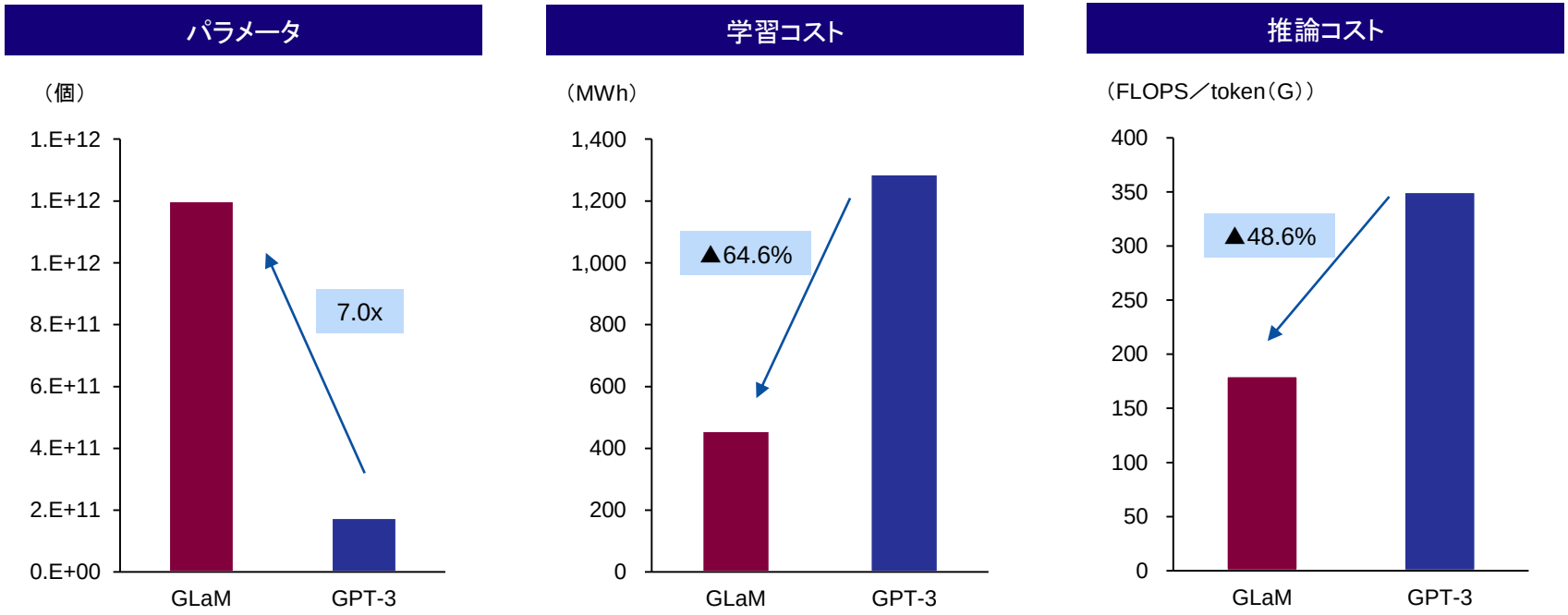


（注1）Transformerに関する公表物は、Vaswani, Ashish, et al. *Attention is all you need*.。引用数はGoogle Scholarにて検索（サイトアクセスは2025年3月）
（注2）MoEに関する公表物は、Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., … & Chen, Z. (2020). Gshard: *Scaling giant models with conditional computation and automatic sharding*.。引用数はGoogle Scholarにて検索（サイトアクセスは2025年3月）
（出所）Google Scholarより、みずほ銀行産業調査部作成

（参考）Googleも、資源効率的なスケーリングアップのためにMoEを採用

- Googleは、2021年にMoEを利用したAIモデル「GLaM」をリリース
  - GLaMは、当時の最先端AIモデル「GPT-3」（OpenAI）の約7倍のパラメータを持ちつつ、GPT-3より低コストで開発
  - DeepSeekと状況は異なるものの、当時のGoogleもスケーリングは非常に高コストとし、モデルアレンジによる資源効率的なスケーリングアップを追求

GLaM（Google）とGPT-3（OpenAI）の比較

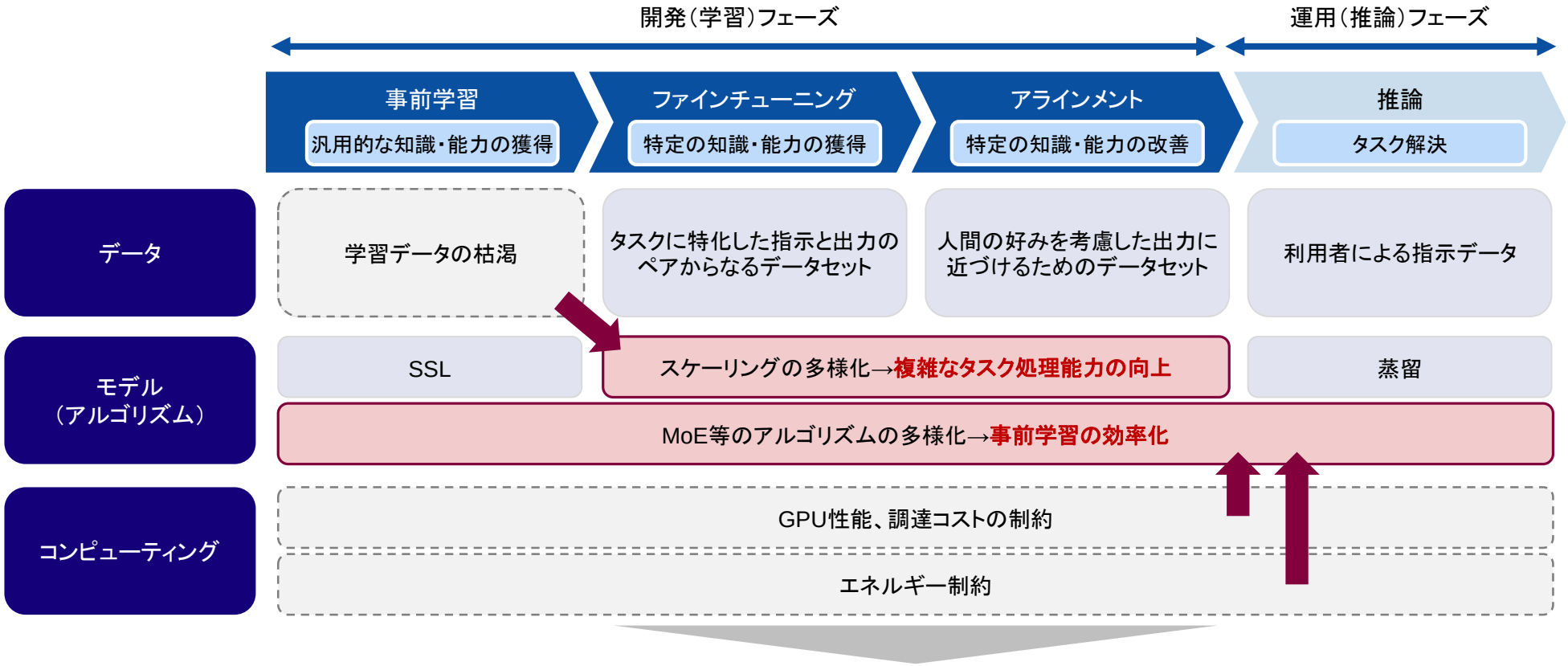


（出所）Du, Nan, et al. *"Glam: Efficient scaling of language models with mixture-of-experts."* International conference on machine learning. PMLR, 2022.より、みずほ銀行産業調査部作成

資源効率的なアプローチへの移行により、AI開発の裾野は拡大

- AIモデルの開発環境は、学習データの枯渇とエネルギー制約等の課題に対して、資源効率的なアプローチを模索する展開
  - DeepSeekは、オープンソースモデルとして、特に学習方法とアルゴリズムにおいて効率的な開発の可能性を示唆
- 資源効率的なアプローチは、AI開発への参入を容易にすることで、AI開発の裾野を拡大し、AIの普及に貢献する可能性

AI開発を取り巻く環境変化と開発トレンドへの影響



AI開発のトレンドは、コンピューティングに依存しない資源効率的アプローチに移行

（出所）みずほ銀行産業調査部作成

参考文献・サイトアドレス

- Bi, Xiao, et al. *"Deepseek llm: Scaling open-source language models with longtermism."* arXiv preprint arXiv:2401.02954 (2024).
- Liu, Aixin, et al. *"Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model."* arXiv preprint arXiv:2405.04434 (2024).
- Liu, Aixin, et al. *"Deepseek-v3 technical report."* arXiv preprint arXiv:2412.19437 (2024).
- Guo, Daya, et al. *"Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning."* arXiv preprint arXiv:2501.12948 (2025).
- Jaime Sevilla et al. (2024), *"Can AI Scaling Continue Through 2030?"*. Published online at epoch.ai. Retrieved from: 'https://epoch.ai/blog/can-ai-scaling-continue-through-2030' [online resource]
- Epoch AI, *'Data on Notable AI Models'.* Published online at epoch.ai. Retrieved from 'https://epoch.ai/data/notable-ai-models' [online resource].
- https://scholar.google.com/citations?view_op=view_citation&hl=ja&user=oR9sCGYAAAAJ&citation_for_view=oR9sCGYAAAAJ:zYLM7Y9cAGgC
- https://scholar.google.com/citations?view_op=view_citation&hl=ja&user=5VYT3AIAAAAJ&citation_for_view=5VYT3AIAAAAJ:qjMakFHDy7sC

産業調査部　次世代インフラ・サービス室 戦略プロジェクトチーム

前島 裕　yu.maeshima@mizuho-bk.co.jp
齊藤 勇樹　yuki.c.saito@mizuho-bk.co.jp

X（Twitter）公式アカウント「みずほ産業調査」はこちら

産業調査部 発刊レポートはこちら

Mizuho Short Industry Focus／246

2025年3月31日発行

編集／発行　みずほ銀行産業調査部　東京都千代田区丸の内1-3-3　ird.info@mizuho-bk.co.jp