2024 年 7 月 1 日　全 13 頁

生成 AI 利活用に関する技術・サービスの動向

基盤モデルなどの最新動向、および全体像・自社事例を解説

デジタルソリューション研究開発部　田中誠人

［要約］

- 2024 年も生成 AI は急速な普及が進み、多くの企業や組織がその効果を実感している。特に、技術やインフラ、サービスの領域は動きが速く、基盤モデルの多様化や生成 AI 活用の高度化などが次々と実現している。例えば、基盤モデルの多様化に関するポイントとして、ビッグテックによる基盤モデル開発の進展や、国内企業で扱いやすい日本語化・小型化モデルの増加が挙げられる。

- トピック①：最新の基盤モデルを用いたアプリケーションのデモを紹介した。レポートからラジオを自動で作成するものや、自社の環境で稼働するセルフホスト型 LLM があり、こうしたデモは今後のビジネス活用を考えるうえでヒントになるだろう。

- トピック②：乱立する基盤モデルやサービスを俯瞰するため、生成 AI の技術・サービスを構成するプレイヤーを可視化した。こうした全体像を理解したうえで、企業は自社の現在地を把握し、次にどのような取り組みを行うべきか検討するべきだろう。

1. はじめに

本レポートは、『生成 AI（LLM）のビジネス適用の潮流 ～ 画像系処理と検索拡張生成（RAG）の革新的な可能性』(2024 年 3 月 18 日)の続編である。前レポート以降にあった動きを中心に説明しているが、本レポートから読み始めても問題ないように構成されている。

前レポートでは、生成 AI をめぐる動向として 7 分野の考察を行い、その中でも生成 AI の画像系処理と検索拡張生成（RAG: Retrieval Augmented Generation)について詳しく説明した。画像処理に対応する生成 AI や関連技術が数多く登場し、これらの技術はデジタルトランスフォーメーション(DX)を促進する可能性があることに触れた。また、RAG は、大規模言語モデル(LLM)が事実と異なる不正確な回答を生成してしまうハルシネーション(幻覚)などへの対策として有効な手法であり、この手法を用いた自社データの活用事例が増えてきていることを述べた。

本レポートでは、2 章で生成 AI をめぐる動向のアップデートを行う。前レポートで考察した 7 分野を引き継ぎながら、その動向を追跡して昨今の状況について述べる。3 章では、最新の基盤モデルを用いたアプリケーションのデモを紹介し、ビジネスでの活用について考察する。4 章

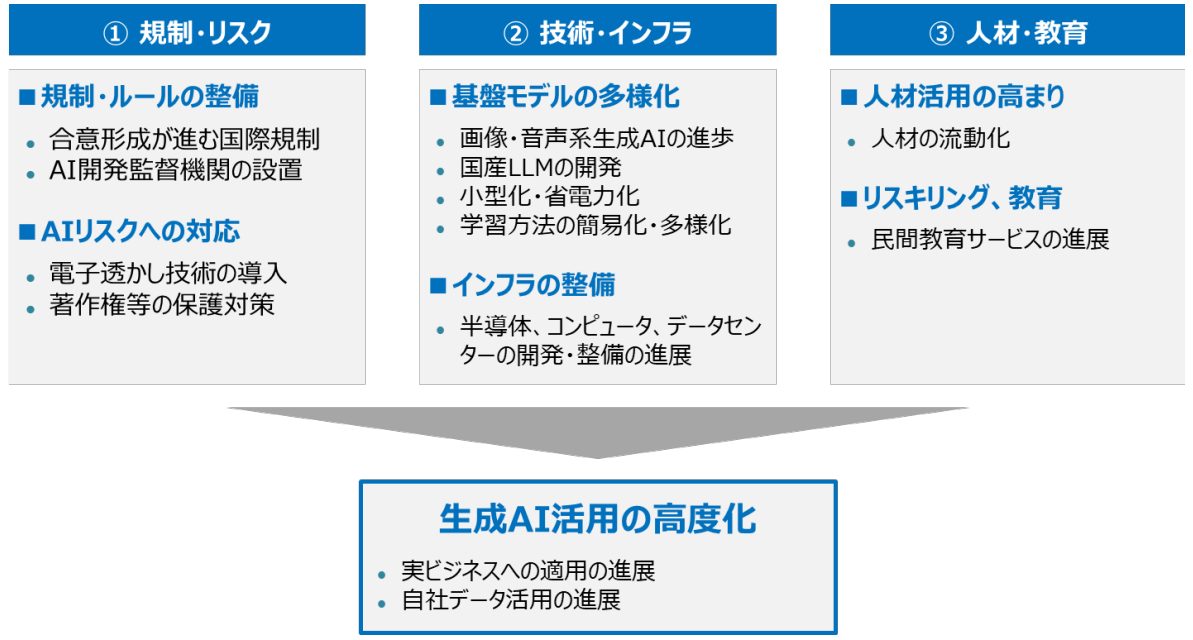
では、乱立する基盤モデルやサービスを俯瞰するため、生成 AI の技術・サービスを構成するプレイヤーを可視化し、企業はどう生成 AI に取り組むべきかを述べる。

なお、生成 AI の基礎知識、金融ビジネスや雇用への影響については、大和総研が公開している別のレポートやウェビナーでも解説しているので、あわせて参照いただければ幸いである [1]。

2. 生成 AI をめぐる最新動向

図表 1 は、生成 AI をめぐる動向を構造化したものである。なお、前レポートでの 7 分野を引き継いでいるが、見やすさのために表現の仕方を変更している [2]。まず、①規制・リスクというところでは、国際規制などの強化や、電子透かし技術の導入といったリスク対応が行われている。次に、②技術・インフラは特に動きの大きい分野であり、画像・音声系生成 AI の進歩といった基盤モデルの多様化や、それらを開発するコンピュータなど、インフラ整備の動きが活発化した。さらに、③人材・教育というところでは、生成 AI を使いこなす人材の必要性から、リスキリングや教育も進展している。そして、これらの動きの結果として、実ビジネスへの適用の進展など、生成 AI 活用の高度化が起こっている。

図表 1　生成 AI をめぐる動向の構造イメージ図（更新版）



（出所）大和総研作成

[1] 例えば、以下のようなウェビナーを公開している。
生成 AI が変革する日本の雇用とビジネス（ https://it-solution.dir.co.jp/l/973193/2024-05-07/5wnbt ）

[2] 前レポートにおける (4)規制強化・(5)リスク対応 が本レポートの ①規制・リスク に対応している。同様に、(1)生成 AI の多様化・(2)生成 AI 環境の多様化 が ②技術・インフラ に、(6)人材活用の高まり・(7)リスキリング、教育 が ③人材・教育 に、(3)生成 AI 活用の高度化 が 生成 AI 活用の高度化 に、それぞれ対応している。

これらの領域のなかで、特に動きが速いのは技術・インフラ、および生成 AI 活用の高度化の領域である。そこで、本章では（1）基盤モデルの多様化と（2）インフラの整備、（3）生成 AI 活用の高度化について、前レポート以降にあった主な動きを記載する。

（1）基盤モデルの多様化

基盤モデルの多様化に関する動きを**図表 2** に示した。取り上げた動きのポイントとして、ビッグテックによる基盤モデル開発の進展と、国内企業で扱いやすい日本語化・小型化モデルの増加、の 2 点が挙げられる。

図表 2　基盤モデルの多様化の動き

時期	企業等	内容の概要
2月	日本IBM	日本語性能を向上させたAI基盤モデル「Granite（グラナイト）日本語版モデル」の提供開始を発表。AI＆データプラットフォーム「IBM watsonx」で使用する
	Google	大規模言語モデル「Gemma」を公開。同社のLLM「Gemini」と同じ技術を活用した軽量版LLMで、「オープンなAIモデル」として位置付けている
3月	KDDI	大規模言語モデルの開発を手掛けるELYZA（イライザ）を連結子会社化すると発表。ELYZAは東京大学・松尾豊教授の研究室から発足したAI企業
	楽天グループ	日本語に特化した大規模言語モデル「Rakuten AI 7B」などを公開。フランスのAIスタートアップ・Mistral AIのLLM「Mistral-7B-v0.1」をベースに開発した
	NTT	大規模言語モデル「tsuzumi（つづみ）」の商用サービス提供を始めたと発表。その時点で500以上の企業や自治体から導入相談が寄せられているという
	PKSHA Technology	世界で初めて、Retentive Network（RetNet）を活用した日英対応の大規模言語モデルを開発すると発表
4月	Google	生成AI「Gemini 1.5 Pro」のパブリックプレビューを180カ国以上を対象に開始。対応言語の音声を理解する機能と、新たなファイルAPIを備える
	OpenAI	日本法人である「OpenAI Japan」の設立と営業開始を発表。同時に、日本語に特化したGPT-4も公開された
	Meta	オープンソースのLLMの最新版「Llama 3」を発表。80億パラメータと700億パラメータの2モデルを提供
	Microsoft	小規模言語モデルの「Phi-3」を発表した。3サイズ（mini、small、medium）あり、miniのトレーニングデータは38億パラメータ
	Apple	オープンソースの言語モデル「OpenELM」を公開。パラメータ数の異なる4つのモデルがあり、小さいものから、2億7000万、4億5000万、11億、30億
	NEC	LLM「cotomi」のラインアップ拡充のため、学習データやアーキテクチャを刷新した「cotomi Pro」「cotomi Light」を開発。高速・高性能なモデルだという
5月	東京工業大学など	理化学研究所のスーパーコンピュータ「富岳」を用いて学習した日本語能力に優れた大規模言語モデル「Fugaku-LLM」を公開
	OpenAI	最新モデル「GPT-4o」を発表。パフォーマンス向上やテキスト、音声、視覚にまたがるマルチモーダル機能など、従来のモデルから機能強化が図られている
	Google	テキストから動画・画像・音楽を生成できるAIモデル「Veo」「Imagen 3」「Music AI Sandbox」を発表。たとえば「Veo」は、1分を超えた1080pのビデオを生成可能
	Google	基盤モデル「Gemini 1.5 Pro」のアップデート、新しい軽量モデル「Gemini 1.5 Flash」、次世代オープンモデル「Gemma 2」などを発表

（出所）2024 年 5 月 31 日までの各種報道を基に大和総研作成

ビッグテックによる基盤モデル開発の進展

Google や Meta、Microsoft など、ビッグテックと呼ばれる企業による基盤モデルの開発が一層進展している。Microsoft は OpenAI と提携しているが、その OpenAI は 5 月に、新しい基盤モデル「GPT-4o（フォーオー）」を発表した。GPT-4o の「o」は、「すべての」という意味を持つ「omni（オムニ）」から来ており、音声や画像を含めたマルチモーダルな理解力で高い性能を示すとしている。OpenAI によるデモ動画でも、人間同士で話すように自然に会話する様子などが公開されている。

Meta は 4 月に、同社が提供する LLM の最新版「Llama 3」を発表した。80 億パラメータと 700 億パラメータの 2 モデルで、いずれもほぼすべてのクラウドサービスでアクセス可能になるという。同社の LLM はオープンソースであり、他社の LLM 開発にも利用されていることから、このアップデートが波及する効果は大きいだろう。例えば、先代のモデルである「Llama 2」は、KDDI が連結子会社化した ELYZA（イライザ）において、日本語による執筆や情報抽出の性能に優れた LLM を開発するためのベースとして利用されているようだ。

トップレベルの性能を誇る大型モデルの開発が進展する一方で、小規模なモデルが複数発表されていることも注目に値する。Google は 5 月に、新しい軽量モデル「Gemini 1.5 Flash」を公開した。同社の基盤モデル「Gemini 1.5 Pro」よりも高速な応答が可能だという。また、Microsoft は 4 月に、小規模言語モデルとして「Phi-3」を発表した。最も小さいモデルで 38 億パラメータのモデルとなるが、高品質なデータでトレーニングしているため、同じかより大きなサイズの言語モデルよりも優れたパフォーマンスを示すと説明されている。

国内企業で扱いやすい日本語化・小型化モデルの増加

企業で生成 AI の活用を高度化するためには、日本語や日本文化に強い基盤モデルや、運用コストなどに優れた軽量なモデルが役に立つと言われている。このような日本語化・小型化モデルが増加してきたことも、最近の動きで注目すべき点だろう。

現在、多くの国内企業が国産 LLM の開発に取り組んでいる。例えば楽天グループは 3 月に、日本語に特化した LLM「Rakuten AI 7B」を公開した。フランスのスタートアップ企業が開発した LLM をベースに開発されており、70 億パラメータの LLM であるようだ。また、NTT は 3 月に、独自開発の LLM「tsuzumi（つづみ）」のサービス提供を始めたと発表した。その時点で 500 以上の企業や自治体から導入相談が寄せられていたという。さらに、NEC は 4 月に、同社が提供する LLM「cotomi」のラインアップ拡充のため、「cotomi Pro」「cotomi Light」を開発した。グローバルの LLM と同等の高い性能を、十数倍の速度で実現する高速・高性能なモデルだという。

日本語化・小型化の動きはグローバルの基盤モデルでもみられる。小型化については前述のとおり、Google や Microsoft などが軽量なモデルを相次いで発表している。日本語化は依然として日本企業による動きが中心ではあるが、OpenAI が 4 月に、日本法人である「OpenAI Japan」を設立し、同時に、日本語に特化した GPT-4 を公開した、という動きもみられた。

(2) インフラの整備

インフラの整備に関する動きを**図表 3** に示した。取り上げた動きのポイントとして、生成 AI の急速な普及などでデータセンターや半導体の需要は高い状況が続いており、それに対応する

図表 3　インフラの整備の動き

時期	企業等	内容の概要
2月	Qualcomm	AIや5Gに向けた半導体のラインアップ強化を発表。オンデバイスでマルチモーダルAIモデルを実行できる機能や、様々な生成AIを利用できる環境を提供
3月	NVIDIA	次世代のAI・データセンター向けGPUプラットフォーム「Blackwell」を発表。現世代の「Hopper」と比べて高いパフォーマンス、コスト効率や電力効率を実現するという
	NVIDIA	最新のGPU「H200」の出荷を始めたと発表。最新のAI半導体を相次ぎ投入して高い市場シェアの維持を狙う
4月	Microsoft	日本の大規模クラウドコンピューティングおよびAIインフラの拡充のため、今後2年間で29億ドル (約4,400億円) を投資すると発表
	Google Cloud	分散マネージドインフラサービスの「Google Distributed Cloud」を活用し、ローカルな環境でも生成AIアプリケーションを構築・運用できることを説明
	Meta	独自で手掛けるAI半導体の新型を開発したと発表。AIが回答を導く「推論」の計算処理を速める
	KDDI など	経済産業省は、KDDIやさくらインターネットなど5社のAIスパコンの整備に最大725億円を補助する。国産のAI開発が経済安全保障の観点で重要だと判断した
5月	NEC	生成AIの急速な普及などでデータセンターの需要が拡大していることを受け、神奈川と神戸にデータセンターを開設したことを発表。100%再生可能エネルギーを用いる

（出所）2024 年 5 月 31 日までの各種報道を基に大和総研作成

データセンターや半導体の需要は高い状況が続く

NVIDIA は 3 月に、次世代の AI・データセンター向け GPU プラットフォーム「Blackwell」を発表した。現世代の「Hopper」と比べて高いパフォーマンス、コスト効率や電力効率を実現するといい、基盤モデルの開発や運用にも貢献することが期待される。また、同社は最新の GPU「H200」の出荷を開始したことも発表しており、AI 半導体の高い市場シェアを維持するよう動いていることが分かる。

データセンターの整備については、各社が多額の資金を投入して推し進めている状況である。例えば、Microsoft は 4 月に、日本の大規模クラウドコンピューティングおよび AI インフラの拡充のため、今後 2 年間で 29 億ドル（約 4,400 億円）を投資すると発表した。また、NEC は 5 月に、神奈川と神戸にデータセンターを開設したことを発表しており、さらに利用する電力を 100%再生可能エネルギーで賄うとしている。

国は、国産の AI 開発が経済安全保障の観点で重要だとの判断のもと、国内企業のインフラ整備を支援している。そのひとつとして、経済産業省は 4 月に、KDDI やさくらインターネットなど 5 社の AI スパコンの整備に最大 725 億円を補助することを発表している。このような動きは、国内で生成 AI を開発・運用したい企業にとってポジティブなものとなるだろう。

(3) 生成 AI 活用の高度化

生成 AI 活用の高度化に関する動きを**図表 4** に示した。取り上げた動きのポイントとして、さまざまな分野で活用が広がっていること、自社データの活用など高度化が進んでいること、の 2 点が挙げられる。

図表 4　生成 AI 活用の高度化の動き

時期	企業等	内容の概要
2月	楽天グループ	OpenAIと協業し、通信事業者向けのAIツールを共同開発すると発表。AIを活用して、ネットワークの不具合を事前に察知し故障を防ぐ機能などの提供を目指す
3月	東芝テック	生成AIを活用し、販促クーポンを配信できるシステムを開発。顧客データが足りなくても生成AIが架空の消費者像を浮かび上がらせ、好みを推定してクーポンを配信する
	Salesforce	営業や問い合わせ対応のメール文面案を顧客に合わせて自動で作り分ける技術を日本でも導入。過去の購入歴や関心といった顧客情報を反映できる
	みずほFG	2024年内にもシステム運用業務に生成AIを本番導入する。勘定系システム「MINORI」を中心とした重要システムが対象だという
	オルツ	独自LLMを基盤にした新サービスを発表。自動オペレーションシステムや採用人事やM＆Aに関するサービスを展開する
4月	Fintertech など	生成AIとAIアバターを活用し、クラウド型応援金サービス「KASSAI（カッサイ）」の問い合わせに応対するAIオペレーター「KOTO（コト）」を開発
	Google Cloud	開発者、Google Cloudサービス、アプリケーション向けの新世代AIアシスタント「Gemini for Google Cloud」を発表
	大和総研	生成AIと高度な分析技術を用いて、上場企業の人的資本情報開示や施策実行をサポートするウェブサービス「KOSMO-ウェルビーイングナビ」を開発
	Adobe	画像生成AI「Firefly」による「画像を生成」機能をPhotoshopのβ版に搭載。その他、Fireflyを使った生成塗りつぶし機能なども搭載される
	DeepL	初のLLM搭載製品「DeepL Write Pro」を発表。ビジネス用にカスタマイズされた文章作成支援サービスで、独自のLLMを採用し、企業のナレッジワーカー向けに展開する
5月	太陽生命　など	生成AIを活用したアバターによる生命保険募集において実証実験を実施。その結果、将来の商用利用を想定した、より具体的な検討に進めるとの結論に至った
	AWS	ソフトウェア開発の迅速化や社内データ活用を支援する生成AIアシスタント「Amazon Q」の一般提供開始を発表
	Google	脅威インテリジェンスソリューション「Google Threat Intelligence」を発表。生成AI「Gemini」の搭載によって脅威情報の検索と洞察の迅速化を実現する
	OpenAI	「ChatGPT」のデータ分析機能の強化を発表。表やグラフを操作したり、「Googleドライブ」や「Microsoft OneDrive」からファイルを直接追加したりできるようになる
	MILIZE	複数の大規模言語モデルを用いるマルチLLMを自律型のエージェントモデルで提供するフレームワーク「MILIZE Financial AGENT」の開発を発表
	NEC	Celonisと共同で、生成AI「cotomi」とCelonisのプロセスインテリジェンスプラットフォームを連携させたサービスの実証を開始したことを発表

（出所）2024 年 5 月 31 日までの各種報道を基に大和総研作成

さまざまな分野で活用が広がる

生成 AI を活用したサービスは多岐に亘るため、その動きを網羅的に取り上げることは難しいが、ここで取り上げた動きだけを見ても、さまざまな分野で活用が進んでいることが分かる。例

えば、楽天グループは2月に、OpenAIと協業し、通信事業者向けのAIツールを共同開発すると発表している。AIを活用して、ネットワークの不具合を事前に察知し故障を防ぐ機能などの提供を目指すという。また、マーケティングの分野では、東芝テックは3月に、生成AIを活用し、販促クーポンを配信できるシステムを開発した。顧客データが足りなくても生成AIが架空の消費者像を浮かび上がらせ、好みを推定してクーポンを配信するようだ。さらに、システム開発・運用の分野では、みずほフィナンシャルグループは、2024年内にもシステム運用業務に生成AIを本番導入するという。勘定系システム「MINORI」を中心とした重要システムが対象だとされている。その他、セキュリティに関して、Googleは5月に、脅威インテリジェンスソリューション「Google Threat Intelligence」を発表した。生成AI「Gemini」の搭載によって脅威情報の検索と洞察の迅速化を実現するという。

このように、生成AIを活用したサービスはさまざまな分野で展開されており、業務の効率化や高度化に活かすことが可能である。企業は、自社の目的に合ったサービスを選択し、活用していくことが求められる。

自社データの活用など高度化が進展

前レポートでも指摘したように、自社データの活用などでより専門的なタスクを生成AIに実行させる事例が増加している。これを実現する手法として主に用いられているのが検索拡張生成（RAG）である。改めて説明すると、RAGは、自社に蓄積した情報や外部の最新情報を活用する手段として、信頼できるデータを検索して情報を抽出し、それに基づいてLLMに回答させる方法である。LLMが事実と異なる不正確な回答を生成してしまうハルシネーション（幻覚）への対策として有効とされている。このようなカスタマイズ・チューニングの利用が広がることで、サービス活用の高度化が進んでいる。

同時に、生成AIが既存のサービスに組み込まれる事例が多くみられることも、注目すべきだろう。例えば、Salesforceは3月に、営業や問い合わせ対応のメール文面案を顧客に合わせて自動で作り分ける技術を日本でも導入した。生成AIを活用し、過去の購入歴や関心といった顧客情報を反映できるという。また、Adobeは4月に、画像生成AI「Firefly」による「画像を生成」機能をPhotoshopのβ版に搭載した。その他、Fireflyを使った生成塗りつぶし機能なども搭載されるようだ。さらに、DeepLは4月に、初のLLM搭載製品「DeepL Write Pro」を発表した。ビジネス用にカスタマイズされた文章作成支援サービスで、独自のLLMを採用し、企業のナレッジワーカー向けに展開するという。

これらのサービスは、以前から存在していたものか、あるいは存在していたサービスを拡張したものであり、（追加のライセンスが必要な場合もあるが）従来のユーザーがこれまでの延長として利用できる。今後も、普段使っているサービスに生成AIが搭載される、という事例は増えていくとみられ、生成AIのさらなる普及に寄与すると考えられる。

3. トピック①：最新の基盤モデルを用いたアプリケーションのデモ

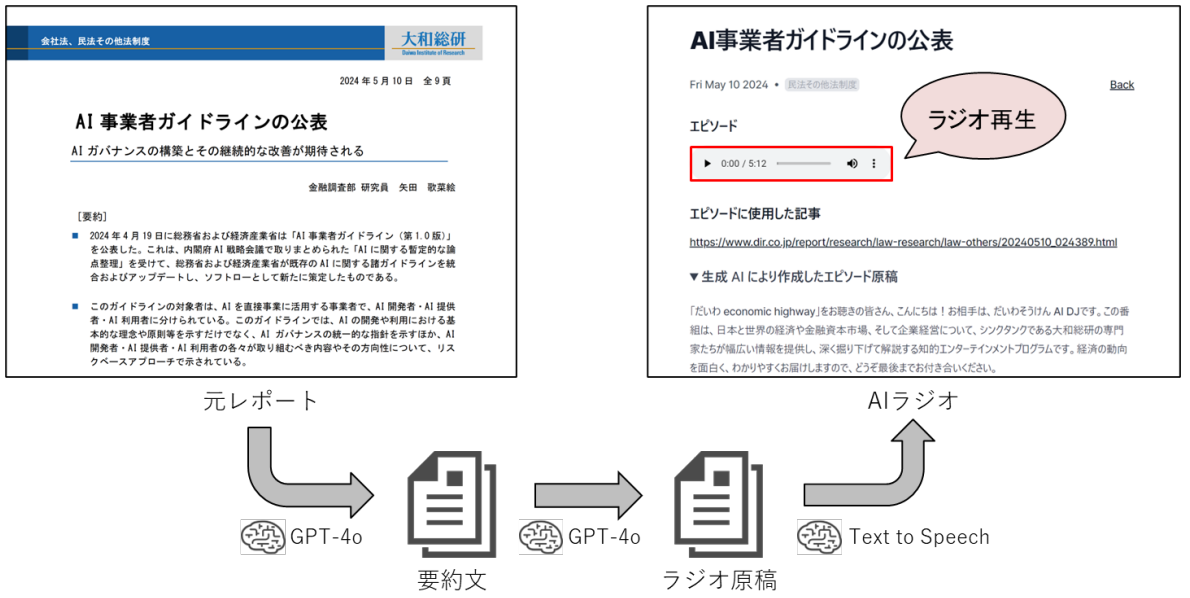
大和総研では、最新の基盤モデルの検証を随時行っており、それらを用いたアプリケーションを開発して社内で公開している。こうしたアプリケーションは、最新技術を体感することに役立つほか、実ビジネスへの適用を検討するうえでアイデアのヒントとなる。そこで、本章では、最新の基盤モデルを用いたアプリケーションのデモを二つ紹介し、ビジネスでの活用について考察する。

レポートの AI ラジオ

AI ラジオは、大和総研が公開しているレポートを、生成 AI と Text to Speech の技術を用いて自動でラジオ化したものである。単なる音声読み上げと異なり、生成 AI がラジオ用に作成した原稿を AI が読み上げるため、レポートの要点をわかりやすく聞くことができる。

図表 5 はレポートから AI ラジオを作成するイメージを示している。まず、元レポートを GPT-4o に入力し、要約文を作成する。元レポートの長さはものによって異なるが、この例では 9 ページのレポートから 800 文字程度の要約文を作成している。その後、ラジオ構成の指示などとともに要約文を GPT-4o に入力し、ラジオ原稿を作成する。最後に、こうして作成したラジオ原稿を Text to Speech で読み上げ、音源を作成する。

図表 5　レポートの AI ラジオ



（出所）大和総研作成

このアプリケーションのポイントは、自動化によってラジオ作成の手間を大幅に削減できることである。ラジオの形式でレポートの内容を聞けることは、忙しいビジネスパーソンにとって価値があると考えられるが、人力でこれを作り上げるには一定程度の手間が必要である。また、一般に、レポートの書き手は原稿の読み上げが得意でない場合もある。上記の AI ラジオでは、

このような課題を一挙に解決できる。

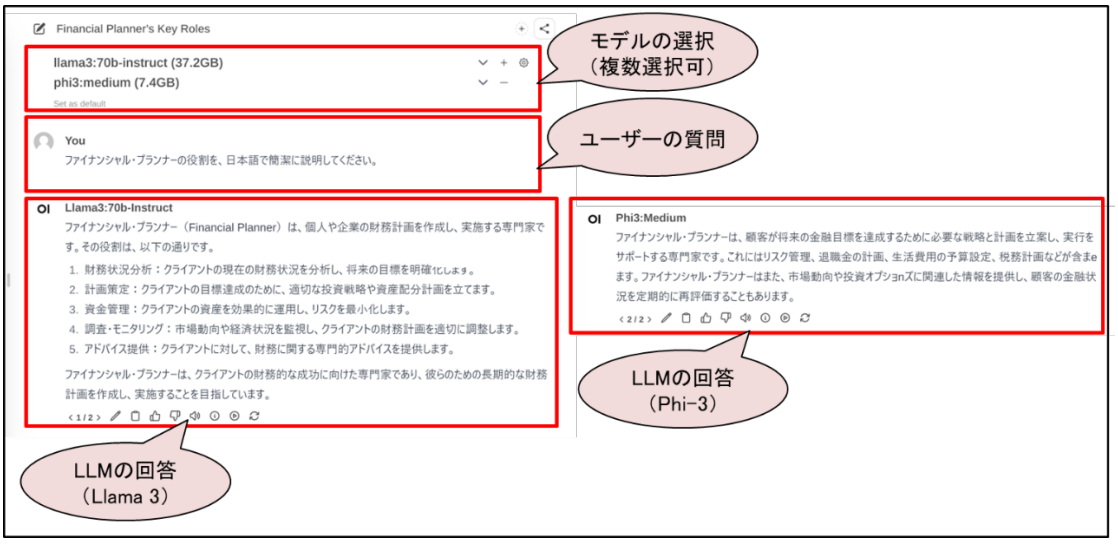
もちろん、AI より人が優れている点もある。生成 AI を用いている以上、ハルシネーションのリスクは考えなければならないし、原稿の内容自体も人が作成した方がよい場合もあるだろう。しかし、自動化で手間が削減できることは大きな利点であり、作成時間の削減や成果物の増加を実現しうる。例えば、より高い品質が求められるものは人が作成し、量やスピード感が求められるものは AI が作成する、というような使い分けによって、質と量の両取りを実現できる可能性があるだろう。

セルフホスト型 LLM

セルフホスト型 LLM とは、オンプレミスなど自社の環境でホスティングできる LLM のことである。例えば、Meta の Llama 3 や Microsoft の Phi-3 などが公開されており、これらの LLM は自社の環境で動かせる。一方、OpenAI の GPT-4o などは（2024 年 6 月時点で）クローズドの LLM であり、Web サービスや API 経由でないと利用できない。

図表 6 は、大和総研社内のサーバーで稼働しているセルフホスト型 LLM の利用イメージである。ここで利用している LLM は、従来の LLM よりも格段にパラメータが少ない軽量なモデルでありながら、ベンチマークによってはそれらに匹敵する性能があるとされている。また、下図の例では、社内でホスティングしている複数の LLM に同時に質問を行い、回答を比較することを

図表 6　セルフホスト型 LLM の利用イメージ



（出所）大和総研作成

このアプリケーションのポイントは、インターネットを介さないプライベートな環境でも、一般的な LLM に遜色ない性能の LLM が実装できることである。例えば、機微情報を扱う、外部と通信ができない環境にある、大量データを処理したいが API の費用が負担になる、などの場合に、セルフホスト型 LLM は有効な選択肢になると考えられる。

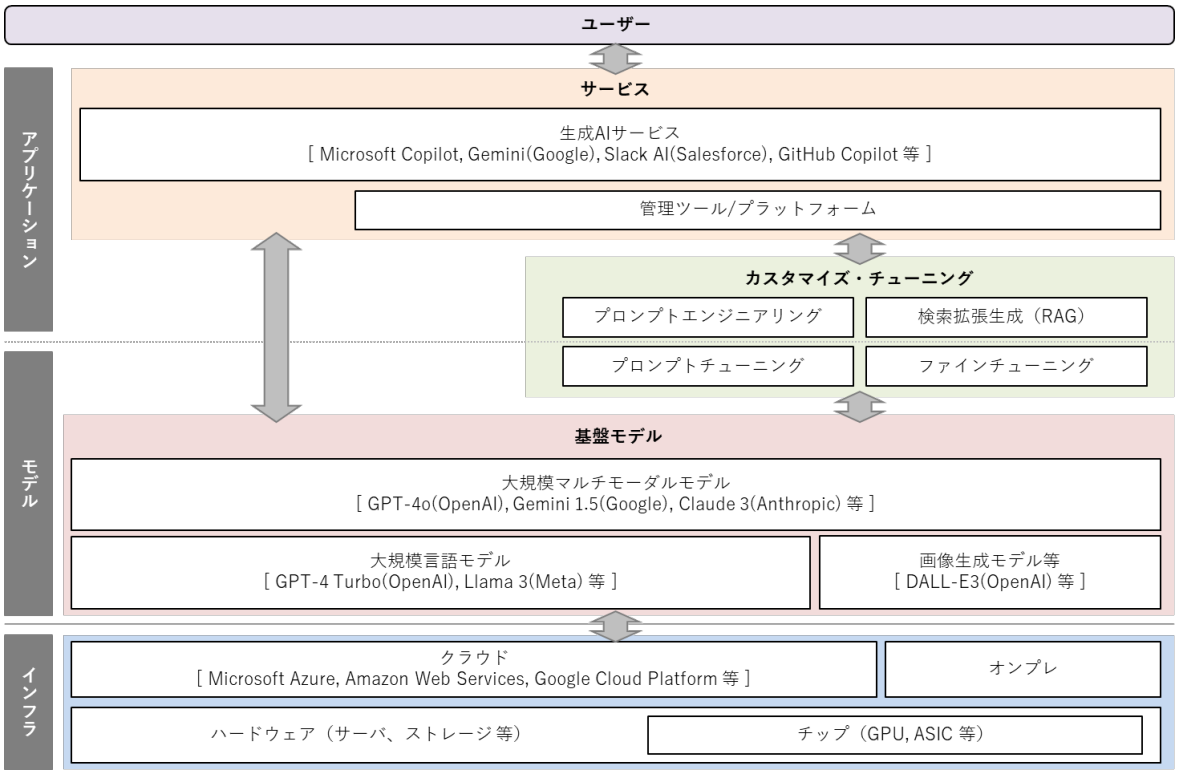
4. トピック②：生成 AI 利活用に関する技術・サービスの全体像

生成 AI に関連する技術やサービスは日に日に拡大しているため、その動向を追うのは苦労が伴う。特に、最近になって生成 AI に取り組みはじめたユーザーにとっては、全貌を理解するのは容易ではないだろう。そこで、本レポートでは、生成 AI の技術・サービスを構成するプレイヤーを可視化し、これに取り組む企業の一助となることを目指す。具体的には、**図表 1** で紹介した各領域のうち、技術・インフラ、および生成 AI 活用の高度化（サービス等）の領域の解像度を高め、これらを構成するプレイヤーの様相を捉える。

生成 AI の技術・サービスを構成するプレイヤー

生成 AI の技術・サービスを構成するプレイヤーを可視化したものが**図表 7** である。プレイヤーは、アプリケーション、モデル、インフラと三つの領域に分類できる。まず、アプリケーションは、ユーザーが直接触れることになるサービスを中心とした領域である。生成 AI を活用したサービスとして、Microsoft Copilot や Gemini、Slack AI などが挙げられる。その他、生成 AI の導入を支援するための管理ツールやプラットフォームもこの領域に含まれる。これらは、前述のサービスよりは IT 部門寄りのアプリケーションとなるだろう。

図表 7　生成 AI の技術・サービスを構成するプレイヤー



（注）記載内容は作成時点の情報に基づく。
（出所）大和総研作成

続いて、モデルは、生成 AI の知能に相当する基盤モデルを中心とした領域である。周知のと

おり、基盤モデルとは、ユーザーの質問や指示（プロンプト）に応じて回答を生成するAIモデルのことである。特に、テキストデータに特化して訓練されたものは、大規模言語モデル（LLM）と呼ばれる。一方、画像や音声に対応する基盤モデルや、その関連技術も数多く登場している。画像解析に対応したOpenAIのGPT-4oもそのひとつである。言語も含めて複数の種類のデータを一度に処理できる能力を持つAIは、マルチモーダルAIと呼ばれ、このような能力を備えた基盤モデルを大規模マルチモーダルモデル（LMM）と呼ぶことがある。大規模マルチモーダルモデルは、生成AIの能力を画像や音声の処理にまで拡張させるものとして注目されており、最新の基盤モデルのスタンダードになりつつある。

アプリケーションの領域とモデルの領域にまたがって存在しているのが、カスタマイズ・チューニングである。前述の生成AIサービスでは、基盤モデルを直接用いることもあるが、より専門的なタスクではカスタマイズ・チューニングを行ったうえで利用することが多い。カスタマイズ・チューニングには複数の手法があり、手法によってその仕組みも異なるため、本レポートではアプリケーションとモデルにまたがるものとして定義した。例えば、前レポートで詳しく解説した検索拡張生成（RAG）もひとつの手法で、社内情報の活用や、ハルシネーション対策などに有効である。最近は、RAGを容易に実現するためのサービスも増加しており、以前よりも導入のハードルが低下している。

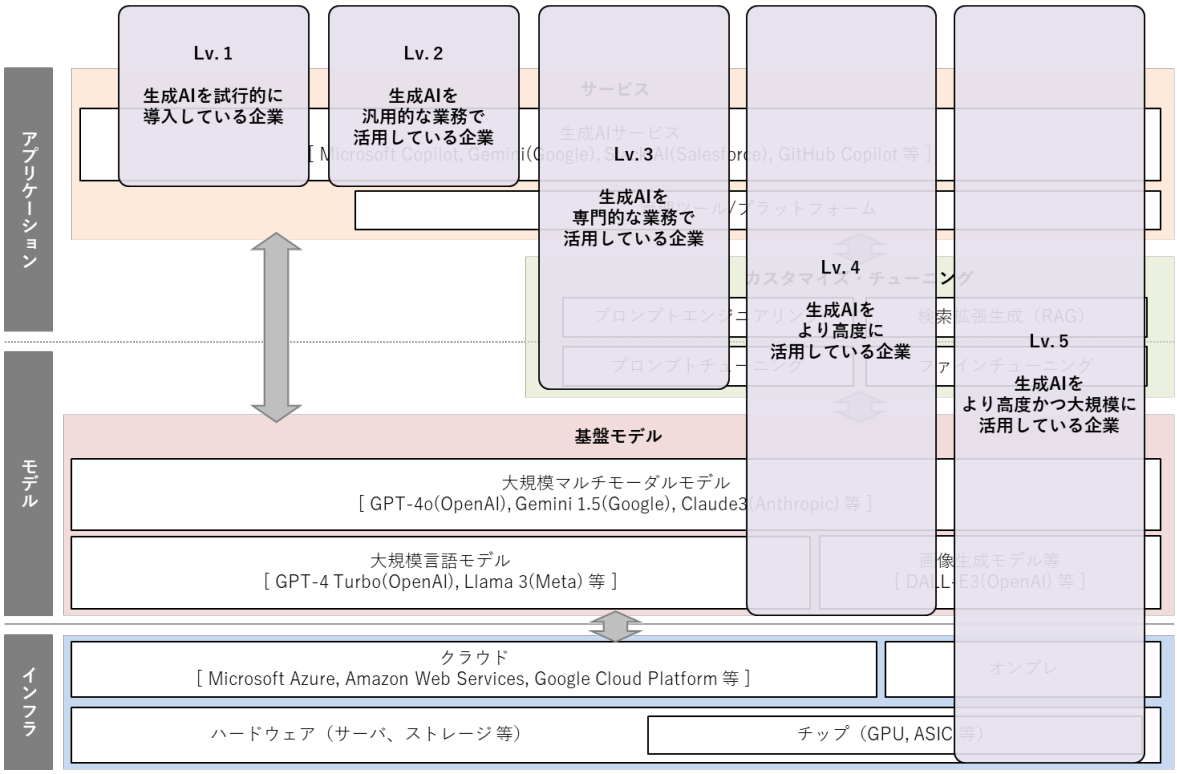
最後に、一番下に位置する領域がインフラである。インフラは、主に基盤モデルを開発・運用するために必要なもので、特に基盤モデルを開発（訓練）する際には膨大な計算資源が必要となる。従来、生成AIのインフラとして用いられているのはクラウド環境であり、Microsoft Azure、Amazon Web Services、Google Cloud Platformなどのメガクラウドが主に利用されている。ただし、最近ではセキュリティ対策のニーズの高まりや、軽量モデルの登場などで環境が整いつつあることから、オンプレ環境を選択する取り組みもある。さらに、これらの環境を構築するためにハードウェアが必要となるが、特に重要なのがGPUなどのチップである。この分野で圧倒的なシェアを誇っているのはNVIDIAだが、MicrosoftやMetaなどの大手企業でチップの独自開発に取り組む事例もみられる。また、国内では、国産のAI開発が経済安全保障の観点で重要だとの判断のもと、国内企業のインフラ整備を支援する動きもある。

生成AIに関連する技術やサービスは、これらの領域が中心となって構成されていると言ってよい。ただし、実際には技術・サービスの領域で完結することはなく、**図表1**で示したような規制・リスクの領域や人材・教育の領域とも相互に関係しており、これらの動向もあわせて考慮することが求められる。

企業は生成AIにどう取り組むべきか

それでは、前述のような全体像を理解したうえで、企業は生成AIにどう取り組むべきだろうか。その答えは、企業が現時点でどの程度生成AIを活用しているかによって異なる。企業のレベル別に見た取り組み方のイメージを**図表8**に示した。

図表 8　企業のレベル別に見た取り組み方（イメージ）



（注 1）Lv. 1～5 は生成 AI の活用度合いで定義したユーザー企業のレベルで、その領域は当該レベルの企業が考慮すべき技術・サービスの要素の範囲を示している。

（注 2）各レベルの領域はその広さを視覚的に表現したものであり、覆いかぶさる技術・サービスの要素と厳密に対応しているものではない。

（出所）大和総研作成

ここでは、生成 AI の活用度合いによって企業のレベルを Lv. 1～Lv. 5 で定義した。なお、図表では示していないが生成 AI にまだ取り組んでいない企業は Lv. 0 である。まずは、自社がどのレベルにあり、次に目指すレベルはどこかを確認することを推奨する。

Lv. 1 や Lv. 2 は、生成 AI を試行的に、あるいは汎用的な業務で活用している企業であり、これらを目指す企業は、一般に提供されている生成 AI サービスを活用することで十分な成果を得られる可能性がある。したがって、まずは上図におけるアプリケーションの領域に着目し、自社の業務に適したサービスを見つけるべきだろう。あるいは、すでに自社が導入しているサービスにおいて、生成 AI を活用した機能が導入されるようであれば、その活用を検討することもよい。ただ、いずれの場合でも、情報漏洩対策などをしっかり考えることは必要である。

Lv. 3 は、生成 AI を専門的な業務で活用している企業であり、これを目指す企業は、カスタマイズ・チューニングによって自社の業務に適した生成 AI の環境を整備することが必要である。それは自社で構築することもできるが、最近ではカスタマイズ・チューニングを支援するサービスが数多く登場していることから、そのようなサービスを利用することで比較的容易に環境を構築できる可能性がある。例えば、5 月に AWS が一般提供の開始を発表した「Amazon Q」のうち、ビジネス向けの「Amazon Q Business」は、企業内のデータや情報に基づいて、質問への回答や

要約の提供、コンテンツの生成などを実施できるとされている。

Lv.4 は、生成 AI をより高度に活用している企業であり、これを目指す企業は、基盤モデルについて深く理解し、直接アクセスして最適な活用方法を検討することが求められる。前述のように、最近は生成 AI サービスが充実してきたことで、基盤モデルに関する深い理解が無くても生成 AI を活用できる環境が整備されている。それ自体は望ましいことだが、より高度な活用を進めるうえでは、モデルごとの得意不得意などを理解し、課題解決に最も適した形で生成 AI を活用することが必要だろう。例えば、大和証券およびグループ各社は 5 月、社内向けの生成 AI 対話環境に、新しい基盤モデル「GPT-4 Turbo with Vision」と「Claude 3」を導入した。米国の Anthropic が提供する Claude 3 はより大規模なデータを入力可能であるうえ、長文の読解能力、日本語の流ちょうさといった分野において優れていると見込んでおり、有望な選択肢になると考えている。

Lv.5 は、生成 AI をより高度かつ大規模に活用している企業であり、これを目指す企業は、オンプレ環境での構築を含め、包括的に技術・サービスの選定を検討することが求められる。このレベルまで到達している企業は少なく、これらの取り組みを実施することによるコストとベネフィットをよく検討する必要があるだろう。

5. おわりに

本レポートでは、まず生成 AI をめぐる動向のアップデートを行った。さらに、最新の基盤モデルを用いたアプリケーションのデモを紹介し、ビジネスでの活用について考察した。そのうえで、乱立する基盤モデルやサービスを俯瞰するため、生成 AI の技術・サービスを構成するプレイヤーを可視化し、企業はどう生成 AI に取り組むべきか述べた。これらの情報が、読者の企業における生成 AI への取り組みを進める一助となれば幸いである。なお、生成 AI をめぐる動向は変化が激しいことから、継続的に情報をキャッチアップすることを推奨したい。

以上