

2024年3月18日 全18頁

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

フロンティア研究開発センター長 田中宏太郎

[要約]

- 2022年11月に公開された米オープンAIのChatGPTを契機に、生成AI（人工知能）である大規模言語モデル（LLM）が世界で爆発的に普及し、多くの企業・組織がその効果を実感する一方、世界的に規制のあり方が議論されている。今後もAIの技術革新と社会の安全との両立を模索することがしばらくの間続くと思われる。
- 特定業務への適合性向上や開発・運用費の低減等のため、分野・業種等に特化した小型・省電力型のLLMや次世代半導体の開発競争が激化している。また「連合学習」や「AI コンステレーション」と呼ばれる、いわば「AI学習の分散処理」の進展が予想される。ITはこれまでも、大型機器での集中処理と小型機器での分散処理の交代サイクルを経て発展してきた。AIの分散処理も同様の経緯をたどる可能性がある。
- 言語処理が主だったLLMの、音声や画像も処理するマルチモーダル化の進化が止まらない。領収書や請求書の文字を読み取り、経理ソフトに反映させる事例も出てきており、デジタルトランスフォーメーション（DX）の前提となるアナログ情報のデジタル化に極めて有効と思われる。
- 検索拡張生成（RAG）と呼ばれる手法を用いた自社データの活用事例が増えてきており、自組織内に蓄積された貴重な情報をいかに業務・ビジネスに活かすかの巧拙が、企業の生き残りの条件の一つになりうると思われる。
- 生成AIが「何に使えるか」といった模索は2023年で終わり、2024年は「どう活かすか」という段階に確実に移るであろう。言い換えるならば、業務効率化や顧客サービスの差別化をもたらすAI・生成AIの開発や利用に取り組みないということが経営リスクとなる段階に突入するといっても過言ではない。

1. はじめに

本レポートは、『生成AI（LLM）の進展と今後の動向 ～ 企業として活用する上での留意点』（2023年11月21日）の続編である。前レポートでは、言語系生成AI（人工知能）であるChatGPTを中心に大規模言語モデル（LLM）の特徴や活用分野等を整理し、業務効率化のメリットとともに

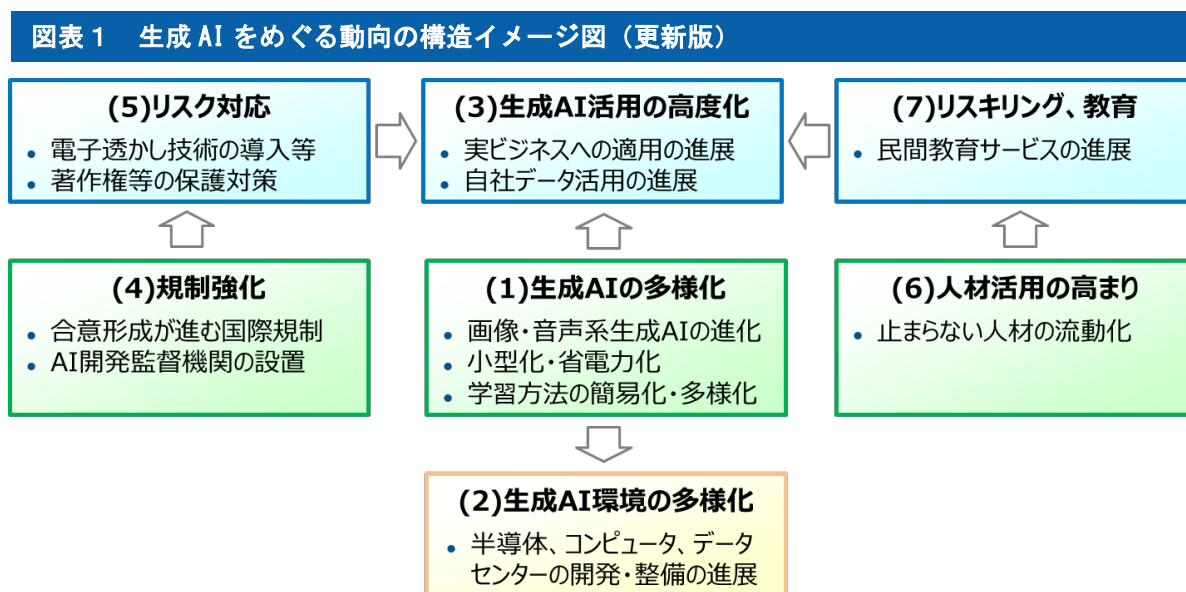
に、著作権を侵害する恐れ等のリスクにふれた。そしてリスクを正しく理解した上で、AI 利用のガイドラインやルールを策定する等の対策をしっかりとうちながら最大限活用するべきであろうと述べた。また今後の動向として7分野での考察を行った。

本レポートでは、2章で7分野の動向を追跡し、昨今の状況について述べる。そしてその中でも筆者が特筆すべきと考える、前レポートでふれられなかった生成AIの画像系処理について3章で述べる。また前レポートで述べた、文書データのベクトル（数値配列）化をさらに活かす検索拡張生成（RAG: Retrieval Augmented Generation）と呼ばれる手法について4章で述べる。

2. 生成AI（LLM）をめぐるその後の動向

前レポートでの7分野の構造について振り返る。2023年は生成AIが爆発的に普及する一方、特定業務への適合性向上や開発・運用費の低減等を目的に（1）の生成AIの多様化が起き、それらを開発するコンピュータ等の（2）の環境の多様化の動きが活発化した。そして（1）の多様化により（3）の活用の高度化がより容易になってきたが、効果を出すための補完活動として（4）の規制強化に対する（5）のリスク対応と、生成AIを使いこなす（6）の人材活用の必要性から（7）のリスクリング（学び直し）が切迫してきている。

この7分野について、前レポート以降にあった主な動きを記載する。



（出所）大和総研作成

（1）生成AIの多様化

画像・音声系生成AIの進化

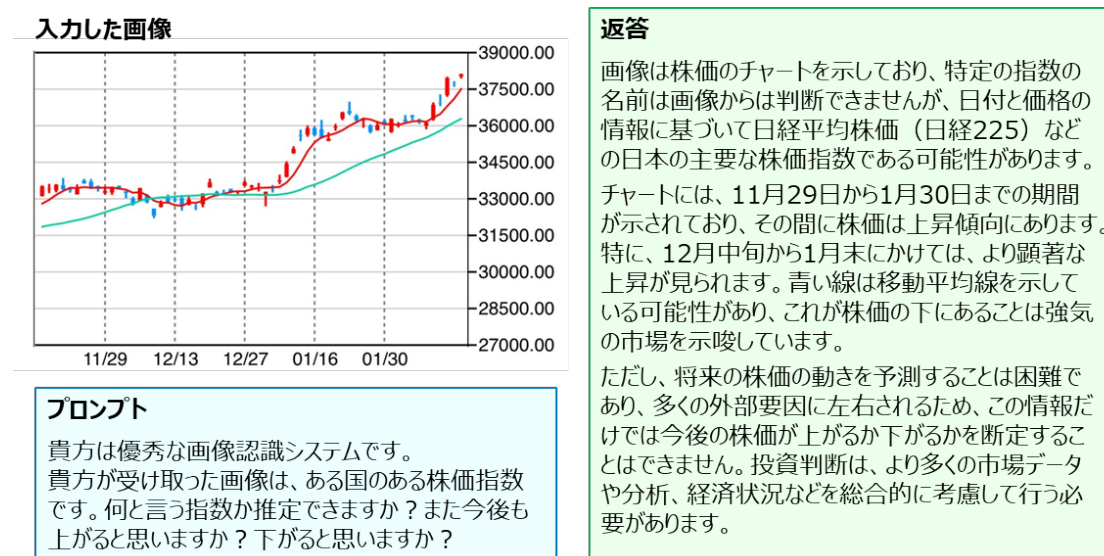
前レポートでは、米オープンAIが提供するLLMであるGPT（Generative Pretrained

Transformer) の最新版である GPT-4V が登場し、画像解析機能が追加され、マルチモーダル AI¹ になったことにふれた。

GPT-4V では、例えば入力した画像から画像の説明や要約・翻訳等の画像内容のテキスト化、抽出した数字のデータ化 (CSV 形式での出力等)、画像にふさわしいタイトルのネーミング、写真画像による場所の特定や推察等が可能となっている。

画像は写真・イラストだけでなく、図表やチャート等の画像でも良い。図表 2 は、スマホで表示できる株価チャートをスクリーンショットで撮影し、数字とグラフのみにして、当センターが社内に公開している GPT-4V を利用したデモンストレーション・アプリに投入したものである。上昇傾向であることや移動平均線の動きは、それらしく説明されている。

図表 2 GPT-4V による画像解析の例①



(出所) 大和総研「画像解析 LLM アプリ」(2024 年 1 月)の実行結果 (2024 年 2 月 15 日)

また米メタが開発した新技術のうち「オーディオボックス」では、文字で指示した環境音を作成できたり、文章の読み上げにおいて声色や速度を簡単に指示できたりする (図表 3)。

さらに 2024 年 2 月 15 日に発表された米オープン AI の動画生成 AI 「Sora (ソラ)」は、最長 1 分の、しかもこれまでとは比較にならないくらいに高精度な動画を生成でき、世界に衝撃を与えている²。映像制作工程の自動化等、制作プロセスを根底から様変わりさせる可能性を秘めている。様々な分野での活用が期待されるが、企業においては顧客サービスの向上や広報活動の高度化、社内マニュアルの映像化等が考えられる。但し現在、米オープン AI 社内で社会に及ぼす影響・リスクを評価中であり、一般公開の日程は未定である (図表 3)。

¹ マルチ (複数) とモーダル (形式・手段) を合わせた用語。「テキスト、音声、画像、動画など複数種類のデータを組み合わせて学習、利用できる」という意味

² テキスト指示で動画生成 OpenAI が新技術「Sora」を公開 まるで実写のような動画を作成 | TBS NEWS DIG ([youtube.com](https://www.youtube.com)) 等多数

2024年はAI開発各社ともマルチモーダルAIの高機能化・実用化に一層力を入れるであろう。

図表3 生成AIの多様化の動き

時期	企業等	内容の概要
11月	オルツ	デロイトトーマツコンサルティングと協力し、金融業界を対象に顧客企業の社内データを学習させたLLM等を提供するサービスを共同で展開
	米アマゾン・ドットコム	年次イベント「AWS re:Invent」にて、ChatGPTに対抗する企業向けの対話型AI「アマゾンQ」の提供を始めると発表
	米メタ	画像や音声を使った生成AIの新技术を複数公開。例えば「オーディオボックス」では、「流れる川と鳥のさえずり」と文字で指示すると、環境音を作成できる
12月	米マイクロソフト	11月の「Ignite 2023」で発表した、自社開発の小規模言語モデル(SLM)「Phi-2」をリリースしたと発表。開発プラットフォーム「Azure AI Studio」にて利用が可能
	米グーグル	生成AIの新たな大規模言語モデル「Gemini」の対話サービスを6日に提供開始。また幅広い処理に適した普及版「Gemini Pro」をクラウド経由で企業に提供を開始
1月	NTTグループ サカナAI	NTTは米グーグル出身者が昨年東京に設立した、小規模な複数のAIを連携することで巨大なAIに匹敵する、省電力型AIの開発を目指す「サカナAI」に出資
	米オラクル	企業がもつ独自データを追加学習させる等してカスタマイズできる、米メタのLlama2やカナダのCohereのLLMを企業向けに提供するOCI Generative AIサービスを開始
	京都大学 製薬17社	2017年に米グーグルが提唱した「連合学習」手法を用いて、機密性の高いデータを共有することなくAIモデルを開発。新薬候補物質を探索する創薬AIの試作に成功
2月	ディープレナー	東京大学発のスタートアップ「ディープレナー」は、商用利用可能な70億パラメータの日本語LLM「blue-lizard」を開発、公開。各社独自のオンプレ型LLMの構築も支援
	米グーグル	「Gemini」を改良し、一度に処理できる量を従来の約30倍に引き上げ、1.5Pro版では最大1時間の動画、11時間の音声、70万単語を処理できるようにした
	米オープンAI	最長1分の高精度な動画を生成できる「Sora」というAIモデルを開発(これまでは数秒程度)。まだ一般公開されておらず、社内で社会的な影響やリスクを評価中

(出所) 2024年2月17日までの各種報道より大和総研作成

小型化・省電力化

米マイクロソフトは、現在、自社のクラウドサービス Azure 上で、米オープン AI の GPT、米メタの Llama2 等の各種 LLM を提供している。その一方で、自社開発の小規模言語モデル (SLM: Small Language Model) である「Phi-2」をリリースしたと発表³した (図表3)。Phi-2 のパラメータ数は 27 億であり、GPT-3 (1,750 億) の約 1.5% の規模である。

前レポートに記載したとおり ChatGPT は 1 時間当たり 100 万 kW、原発 1 基分相当以上の電力を消費するといひ、AI 開発・利用に必要な電力を含む膨大な計算資源は、世界的な課題となっている。2024 年 1 月、国際エネルギー機関 (IEA) は、世界のデータセンターの電力消費量が 2026 年には、2022 年比 2.2 倍、日本の年間総電力消費量に相当する規模に拡大するとの試算を出している⁴。SLM はコスト効率性に優れ、LLM に取って代わる可能性を秘めている。

また NTT グループは、米グーグル出身者らが 2023 年 8 月、東京に設立した「サカナ AI」に出資した (図表3)。サカナ AI の社名は、小さな魚が群れをなして優れた行動をする様子に由来しており、両社は複数の AI 開発会社がつくった多数の小型の AI を連携させることで、単一の巨

³ [Phi-2: The surprising power of small language models - Microsoft Research](#)

⁴ [データセンター電力消費急増 2026年は日本1年分に匹敵 - 日本経済新聞 \(nikkei.com\)](#)

大 AI に匹敵する仮想 AI モデルの構築を目指す。その技術は、元は「星座」の意味から転じて「点と線で連なっている」状態を表すコンステレーション (constellation) という単語を用い、AI コンステレーションと名付けられている。なおサカナ AI は、経済産業省が米グーグルから借り上げたクラウドを無償提供する七つの企業・研究機関の一つに選ばれている (図表 5)。

学習方法の簡易化・多様化

前レポートで、GPT は膨大な情報を用いた事前学習により汎用性が高いものの特定業務に対応できない側面を持っており、その有効な対策の一つである追加学習もデータの準備や学習に相当な手間がかかり、ハードルが高いと述べた。例えば米グーグルは汎用型の PaLM2 に追加学習して医療用の Med-PaLM2 を開発しているが、相応の資金力・体力が必要となる。

しかし昨今では生成 AI 提供事業会社の導入時の支援サービスを用いて比較的容易に行えたり (図表 3 のオルツや米オラクル)、米オープン AI や米マイクロソフトが提供する AI 開発プラットフォームにて、マウスを用いた GUI (Graphical User Interface) 操作で実行できたり、技術的なハードルが低くなってきている。

なお特定業務の用途等に対応する手法として、昨今、検索拡張生成 (RAG) と呼ばれる手法を用いて自社データを活用する事例が増えてきているが、これについては後述する。

図表 4 特定業務に対応できるように生成 AI を調整する方法

プロンプトエンジニアリング (プロンプトデザイン)	プロンプト(質問・指示文)に、自社データを貼り付けて渡すことで学習させる方法。但し渡せる情報量には限界がある。またセキュリティ要件としてオプトアウト申請を行っている場合には、再学習効果はない
追加学習 (ファインチューニング)	事前学習済の LLM に、専門分野等のデータを追加し、再学習する方法。米オープン AI は、GPT-3.5 のオプションとして追加学習機能をリリース済。また OSS [*] (米メタの Llama2 等) に対して追加学習する手段もある
LLM そのものの構築	独自の専門分野特化のモデルをゼロから構築する方法。しかしこれには大量の良質な学習データ、莫大な計算能力、データサイエンティストが必須であり、一般的な企業が単体でできるものではない
検索拡張生成 (RAG: Retrieval Augmented Generation)	AI のハルシネーション(幻覚)対策として、また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、データ等を検索して情報を抽出し、それを LLM にプロンプトとして渡す方法。 自社データの検索では、文章データをベクトル(数値配列)化したデータベースに対する類似検索(ベクトル検索)を用いることが多い

^{*}Open Source Software の略。ソースコードが公開されており、無償で利用できる。

(出所) 大和総研作成

また「連合学習」(Federated Learning) という AI の学習手法の事例も増えている (図表 3 の京都大学と製薬会社 17 社による事例)。そもそも AI は学習用のデータ量 (および質) に比例してモデルの精度が向上する特性があるが、不正な金融取引や医療における症例等、一企業の規模では十分なデータを用意するのが困難な場合がある。米グーグルが 2017 年に提唱した「連合学習」では、各社で共有する中央のサーバーから、まずベースラインの AI モデルを各社に配布

し、次に各社のサーバーにて各社のデータを用いて学習し、そして中央サーバーにモデルの変更点（パラメータ）を戻して統合、このサイクルを繰り返し行う。複数の企業が、社外秘情報や個人情報が含まれたデータを中央のサーバーに送って共有することなく、AI を共同開発することができる。

これら「連合学習」や「AI コンステレーション」は、いわば「AI 学習の分散処理」ともいえる。ITはこれまでも、メインフレームという大型汎用コンピュータでの集中処理、UNIXやWindowsサーバーといった中・小型コンピュータでの分散・連携処理、その変形である Web 型処理を経て、クラウドでの集約処理、エッジコンピューティングというデータ発生源に近いネットワーク機器内等での分散処理というように、集中処理と分散処理の交代サイクルを経て大きく発展してきた。AI での分散処理も同様の発展をするであろうと推測できる。

(2) 生成 AI 環境の多様化

AI 半導体で市場シェア 8 割（2022 年時点）を握るといわれる米エヌビディアを追うべく、米アドバンスト・マイクロ・デバイセズや米マイクロン・テクノロジー等、半導体メーカーが生成 AI に特化した半導体製品を次々と発表している。またソフトウェア大手の米マイクロソフトも半導体開発にこぎ着け、ソフトバンクグループが半導体を生産する新会社の設立を検討している。

図表 5 生成 AI 環境の多様化の動き

時期	企業等	内容の概要
11月	日立ハイテク	山口県に建設する、半導体回路の基板を加工する装置の設計や開発、製造を担う新工場の起工式を開催、2025年2月に完成予定。投資額は約245億円
	米アマゾン・ドット・コム	「AWS re:Invent」にて、処理性能を向上させたAIモデルを動かすための「Graviton 4」およびAIモデル訓練用の「Trainium2」という2種類の半導体を発表
12月	米アドバンスト・マイクロ・デバイセズ	米エヌビディアの主力AI半導体「H100」に匹敵する、新製品「MI300X」を発表。米マイクロソフトや米メタが採用を決めた
	三井物産	東京都内のデータセンターに設置した米エヌビディア製のスパコンを貸出し、アステラス製薬・小野薬品工業が新薬候補の化合物を絞り込む等、新薬の開発に活用
	米マイクロソフト	11月の「Ignite 2023」で発表した、ChatGPTの動作に最適化したクラウドサービス向けCPUの「コバルト」等の2種類の半導体の開発にこぎ着けた
	米マイクロン・テクノロジー	2025年から広島県の工場ですべて最大5,000億円をかけて最先端メモリーを生産。生成AI向けにDRAMを積層させた高性能品「HBM」を生産する計画などが明らかに
1月	米アマゾン・ドット・コム	急激に拡大している日本の企業・行政のクラウド需要に対応するべく、東京と大阪でのデータセンター設備の投資等、2兆2,600億円の投資計画を発表
2月	経済産業省	海外に先行されている生成AIの国産化を支援するため、データ学習等に使うクラウド・サービスを米グーグルより84億円で借り上げ、7つの企業・研究機関に無償提供する
	日米両政府	科学研究に特化したAIの開発において、学習データの共有や、開発に使うスーパーコンピュータの共同利用等の連携に乗り出す(日本は理化学研究所が中心)
	ソフトバンクグループ	孫会長兼社長は、AI向け半導体を生産する新会社の設立を検討。最大で1,000億ドル(約15兆円)の資金を調達、傘下の英アームと連携、調達網を強化する狙い

(出所) 2024年2月17日までの各種報道より大和総研作成

しかし必要な計算資源が不足している現状が続いており、三井物産がスパコンを製薬会社に貸し出したり、経済安全保障の観点も踏まえながら生成AIの国産化支援として経済産業省が米グーグルより借り上げたクラウドを企業・研究機関に無償提供したりしている。日本では、国産の資源供給体制が整うまでの間、このような先行的な動きが続くであろう。

なお米アマゾン・ドット・コムは、2027年までに2兆円超の日本での設備投資計画を発表した。日本で急増するクラウド需要に応えるとともに、クラウド自体は国産ではないものの、重要なデータを国外に持ち出さずに国内で処理したいという企業の要望を満たすことになる。

(3) 生成AI活用の高度化

実ビジネスへの適用の進展

オーナーズの譲渡提案書、リクルートの職務経歴書、デジタル庁の法律系文書など、LLMの得意とする文書生成の領域で業務の効率化や顧客サービスの拡充が進展している。

また画像系生成AIを活用した、オリックスの領収書の文字の読み取り、NECのドライブレコーダー動画の解析、サイバーエージェントの広告への適用等、より具体的な活用が目立つ。

図表6 生成AI活用の高度化の動き

時期	企業等	内容の概要
11月	LINEヤフー	質問投稿サイト「Yahoo!知恵袋」に、当初は脅威と考えていた生成AIを導入。AIの回答が契機となって人間同士の議論が活発化、利用者が増加
	オーナーズ	米オープンAIのGPT等を活用し、従来1日から数日かかっていた譲渡の提案書を1～2時間で作成できるようにする等、M&A業務に特化した生成AIを開発する
12月	リクルート	転職支援サイトにて、求職者が経験やスキル、希望する勤務条件などの質問に選択肢で答えていくと、自動的に職務経歴書を作る機能を導入
	オリックス	中小企業向けに提供する経理ソフトに2024年春から生成AIを導入。領収書や請求書の文字を読み取り、経理ソフトに反映させる技術を確認
	NEC	交通事故の際のドライブレコーダー等の動画を分析・要約する技術を開発。損害保険金請求用の報告書に活用。1時間以上の動画の説明文を数秒で作れる
	旭化成	生成AIを用いた社内データの検索で顧客との監査書類のやりとりが約25時間から約12時間に短縮する等の効果があり、1月から特許申請資料の作成などにも展開
	サイバーエージェント	広告に使う背景や商品の画像を作る生成AIを開発。商品に差し込む光を再現するなど複雑な表現にも対応し、ロケが不要となり、広告制作を効率化
	三菱UFJ銀行	自社ChatGPTによる労働時間の削減効果が月22万時間以上と試算。来春にもAIと社内システムが接続できる仕組みを整え、ウェルスマネジメント業務でも活用予定
	デジタル庁	法律を作る際に必要な文書を自動作成するシステムを2024年度にも導入。膨大な量の紙の印刷や読み合わせといった官僚の事務作業、文書のミスの削減につなげる
1月	三井物産セキュアディレクション	サイバー攻撃被害の分析に生成AIを使い、専門人材を補完。特にマルウェアの解析では、これまで専門家が見つけられなかった内部の脆弱性を発見できたという
	ディップ	生成AIを活用して高い精度で闇バイトを検知できるツールを開発、求人サイト「バイトル」に導入。報酬額、職務内容の不明瞭さ、安全性の主張等から判別
2月	NTTテクノクロス	少ない追加学習量で業界特化が可能なNTTのAI基盤「tsuzumi」を用いて、コールセンターの通話要約機能を高度化するシステムを保険業界等から順次提供開始
	キャピタル・アセット・プランニング	米AwakAppと、ChatGPTを活用したNISA向け投資提案アプリ「W2C」の試用版を共同開発、金融機関に提供。アバターによるバーチャルアドバイザー機能も搭載

(出所) 2024年2月17日までの各種報道より大和総研作成

自社データ活用の進展

図表6の旭化成や三菱UFJ銀行に見られるように、自社データの参照・検索による業務効率化事例も増えてきている。今後、検索拡張生成（RAG）の事例は急増すると考えられる。

(4) 規制強化

合意形成が進む国際規制

G7は2023年末、世界初の包括ルールとしてAIの国際指針に最終合意した。EUでは、AI規制案を早ければ2026年にも施行する見込みである。

このような動きの一方、非営利財団である世界経済フォーラムが毎年1月に開催する年次総会「ダボス会議」では、各国・地域はAI規制の重要性・必要性について一致したものの、規制強化に伴う技術革新の停滞を懸念する見解も多くあがった。日本や米国は法的拘束力を持たないガイドラインの策定、自主規制を重視するソフトロー主義、EUは法律による厳格な規制を目指すハードロー主義であるが、具体策でその温度差が露呈した格好となった。しかしその米国でも昨今、制裁金を科すことができる法案が提出される等ハードローの動きが出てきている。

規制の具体策はまだこれからという状況ではあるが、欧州の市場規模を考慮すると、EUのAI規制案が世界各国のモデルとなる可能性が高いとの見解もある⁵。

図表7 規制強化の動き

時期	企業等	内容の概要
11月	米議会	上院議員6人が超党派で新たな法案を提出。雇用や医療などに関連する「高い影響力を持つAIシステム」と、生体認証データを使う等の「重大な影響力を持つAIシステム」に区分し、前者については制裁金、後者についてはサービス差止め等も可能とした
12月	G7	G7の首脳がAIの国際指針に最終合意。世界初の包括ルールで、G7の12項目ある国際指針と、AI開発者を対象とする行動規範の2つの柱で構成
	EU	生成AIを含む包括的なAI規制案に大筋合意したと発表、早ければ2026年に施行の見込み。またAI事務局を設置し、規制の順守や生成AIをめぐる開発動向を監督
	日本政府	AIの安全性を評価する専門機関の設置について、月内にも開くAI戦略会議で方針を示す。AI関連製品について、軍事転用されるリスク、セキュリティ面の欠陥、制御不能要素の有無、偽情報や偏見の拡散の懸念がないか、市場に出る前に検証する
	日本政府 AI戦略会議	AI基盤モデルの「開発者」、サービス「提供者」、業務での「利用者」を対象に「人間中心」「安全性」「透明性」などの10原則を提示した「AI事業者ガイドライン案」を公表。本案は、適切なデータの学習、開発記録の保存、安全対策等を求め、点検リストの素案も公表
1月	ダボス会議	各国・地域の政府高官もAI規制の重要性にふれ、規制の必要性で一致したものの、「技術革新が止まってしまう」との見解も出、具体論では温度差が露呈した
2月	日本政府	経済産業省傘下の情報処理推進機構(IPA)に、AIの安全性評価基準の策定や実施方法の検討を行う「AIセーフティ・インスティテュート」を設置

(出所) 2024年2月17日までの各種報道より大和総研作成

⁵ AIガバナンスを考える(7)「ブリュッセル効果」の影響 - 日本経済新聞 (nikkei.com)

AI 開発監督機関の設置

EU は規制案の実効性を高めるために AI 事務局を設置し、規制の順守や生成 AI をめぐる開発動向を監督していく方針を明確にしている。日本も 2023 年末、AI の安全性評価基準の策定や実施方法の検討を行う専門機関の設置について AI 戦略会議で方針を示すとしたが、2024 年 2 月、経済産業省傘下の IPA に「AI セーフティ・インスティテュート」を設置した（図表 7）。

(5) リスク対応

電子透かし技術の導入等

偽情報の拡散防止を意図して生成 AI を安全に利用するための技術開発における企業横断の取り組みが米国で起こり、また具体的な対策として音声や画像・映像に対する電子透かし技術の製品・サービスへの組み込みが始まっている。

また 2024 年は、米国、EU、インドなど多くの国や地域で指導者を決める大型選挙が続くため、米オープン AI は世論の誘導や偽情報の拡散への対策として、電子透かしによる技術的な対策をとりつつ、同社技術の選挙活動への利用禁止を発表した。

図表 8 リスク対応の動き

時期	企業等	内容の概要
11月	米アマゾン・ドット・コム	「AWS re:Invent」にて Amazon Titan Image Generator 等を発表。同サービスは、見えない電子透かしをすべての画像に追加、生成された誤情報の拡散を緩和
12月	日本IBM	AIモデルのリスクを管理できる「watsonx.governance」の提供を開始。将来的に予想されるAIに関する規制・政策に対して、順守度合いをダッシュボードで可視化
	米オープンAI	独メディア大手のアクセル・シュプリンガーと提携、ChatGPTが記事の要約を作成。米オープンAIは対価を支払う一方、回答の精度を高める学習に記事を使う
	エヴィクサー	動画の音声に人間では聞き取ることが難しい暗号化した「透かし音」を組み込む等して偽動画を判別するサービスを2024年1月より開始と発表
	ニコン、ソニー等	撮影画像に日時、位置情報、撮影者、編集履歴などの情報を埋め込み、デジタル署名技術で改ざん防止を行うことで、認証情報のないフェイク画像を見分ける
1月	文化審議会	生成AIによる検索や回答の生成が権利侵害になる恐れがあるケースについて考え方をまとめた文化庁の素案を了承。著作物の「無断学習」に一定の歯止めをかけた
	米オープンAI	2024年は大型選挙が世界で続くため、世論の誘導や偽情報の拡散対策として、同社の技術を選挙活動に利用することを禁止。技術対策として「電子透かし」も導入
	アマナイメーجز	同社と取引する国内カメラマンの写真を学習し、「日本風」の画像を生成するAIを開発。画像販売の貢献度に応じて権利者に還元、企業も安心して商業利用できる
2月	米政府	偽情報の拡散防止等、生成AIの安全な利用を促すための技術開発で、200超の企業等と取り組む。日本や英国、欧州の「AI安全研究所」と協力していく
	米グーグル	生成AIが作った画像を判別できるよう、編集履歴等を埋め込む規格の標準化を進めてきた国際標準化団体C2PAのステアリングコミッティに参加

（出所）2024年2月17日までの各種報道より大和総研作成

著作権等の保護対策

米オープン AI は、米ニューヨーク・タイムズから同社の記事を生成 AI の学習に許可なく使用し、著作権を侵害していると提訴された一方、独メディア大手のアクセル・シュプリンガーとは対価を支払って ChatGPT が要約の作成や学習に記事を使えるように提携した（図表 8）。

日本においては、国の文化審議会の小委員会が生成 AI による検索や回答の生成が権利侵害にあたる場合もあるとして文化庁がまとめた素案「AI と著作権に関する考え方について」を了承した。著作権法は、生成 AI に著作物を学習させる際は原則として許諾は不要としているが、文化庁は現行法の解釈を明確にし、著作物の無断学習に一定の歯止めをかけた。しかし著作権法の改正の議論には踏み込んでおらず、海外の事例や技術の進展等を踏まえ、引き続き見直されるものと思われる。

(6) 人材活用の高まり

米国中心に IT 業界は、コロナ禍での「巣ごもり需要」を背景に大量採用したが、対面での経済活動の回復に伴ってその反動が表面化し、利上げにより景気後退の懸念もあり、収益が悪化した。そのため人員削減を続けるとともに、成長分野である AI 向けに投資をシフトさせている。

図表 9 人材活用の高まりの動き

時期	企業等	内容の概要
12月	大和総研 (レポート)	(略)本レポートは(略)職業ごとにタスクのうち生成AIが自動化可能な割合を示す「自動化対象率」を推計した。その結果、日本の就業者の約80%が何らかの形で生成AIの影響を受ける可能性があり、同様に約40%の就業者が仕事の半分以上を自動化できると推計された。 職業別では、事務従事者や管理的職業従事者、専門的・技術的職業従事者などホワイトカラーの職種が、自動化の影響を強く受けるとの結果が得られた。(略)
1月	米マイクロソフト 米グーグル等	米国中心にIT業界では、成長分野であるAI向けに投資を増やすため、コロナ禍等に伴って余剰となった人員の削減が続いている。1月の削減は1万人を超えた
	独SAP	AI事業を強化するための構造改革の一環として、従業員7%超の8,000人を対象に人員削減を含めた再編計画を発表。新規採用とリスキリングによりAI事業を強化
2月	米雇用調査会 社等	チャレンジャー・グレイ・アンド・クリスマス社がまとめた2024年1月の米国企業のレイオフ状況では、前月比約2.4倍の8万人強が削減
	大和総研 (レポート)	(略)本レポートでは(略)生成AIと協働関係にある職業グループ(以下、協働グループ)、および生成AIと代替関係にある職業グループ(以下、代替グループ)を推定した。(略)全就業者のうち、 生成AIの影響を受けやすい協働あるいは代替グループ に分類された割合は 共に20%前後 であり、残り約60%の就業者はその他のグループに分類された。(略)生成AIの普及は、協働グループの雇用と所得にプラスの効果をもたらすほか、新たな職業を誕生させるだろう。一方 代替グループ では、非正規社員を中心に、 雇用や所得が伸び悩む 可能性がある。(略)生成AI時代の本格的な到来に向けて、 生成AIに代替されないスキル の習得・活用支援(略)が求められよう。

(出所) 大和総研レポート『生成 AI が日本の労働市場に与える影響②③』のサマリー部分から抜粋（太字は筆者）。その他は 2024 年 2 月 17 日までの各種報道より大和総研作成

また日本でも就業者全体のうち、生成 AI と協働してサブ的な作業を自動化しつつ、メインの

仕事に一層注力する傾向が強い協働関係にある人々の割合、およびメインの仕事が生成 AI に取って代わられる傾向が強い代替関係の人々の割合は、それぞれ約 20%との推計がある(図表 9)。とりわけ後者の人々に対して、今後リスキリングを実施する必要性が高まるであろう⁶。

(7) リスキリング、教育

米マイクロソフトは AI の雇用への影響を軽減するため、2024 年から 2026 年にかけて労働者や学生が AI を学べる体験講座を開催する。また米国最大の労働団体「労働総同盟産別会議」と協議したり、労働団体と共同でリスキリングを推進する政策を提案したり、取組みを始めている。日本では厚生労働省が、デジタル関連を中心にリスキリング支援対象の講座数を拡充したり、助成率も上げたりする方針である。

但し前述の人員削減が先行し、リスキリングについては個人に対する e ラーニングによる知識習得の機会は拡充されてきているものの、法人向けの実践的なトレーニングについてはまだ少ない印象である(IT 部門をターゲットとした IT ベンダーによる技術支援や認定資格制度は急速に増えている)。そのような中、サイバーエージェントや LINE ヤフー等の民間企業からも、生成 AI の活用法等を学べる法人向け講座が拡充されてきた。

図表 10 リスキリング、教育の動き

時期	企業等	内容の概要
12月	米マイクロソフト	AIの雇用への影響を軽減するため、米国最大の労働団体「労働総同盟産別会議」と協議していくと発表。また2024年から26年にかけて、働く人や学生がAIを学ぶための体験講座も開催。リスキリング(学び直し)を推進する政策を労働団体と共同で提案し、AIの技術取得を後押しする資金も提供
	厚生労働省	支援対象の講座数をデジタル関連を中心に2025年度までに現在より60%多い300以上に拡充、助成率も最大8割に上げ、労働移動を促す
1月	サイバーエージェント	子会社CAリスキリングパートナーズを通じて、生成AIの活用等を学べる法人向け講座の提供を開始
	LINEヤフー	キラメックス社と連携し、企業内大学の「LINEヤフーアカデミア」を通じて、生成AIの活用等を学べる法人向けリスキリングプログラムの外販を開始
2月	米ユードミー	企業のリスキリング需要への対応策として、役職に応じて必要なスキルや学習教材を分析する機能をeラーニングシステムに導入予定。また企業ごとのカスタマイズも検討

(出所) 2024 年 2 月 17 日までの各種報道より大和総研作成

3. 画像系処理の可能性

画像解析の進化例① ～ 印刷文字

図表 11 はあるコンビニの領収書をデモンストレーション・アプリの GPT-4V に解析させた結果である。1 回目は税込価格を、2 回目は直接的には印字されていない税抜価格を回答するよう

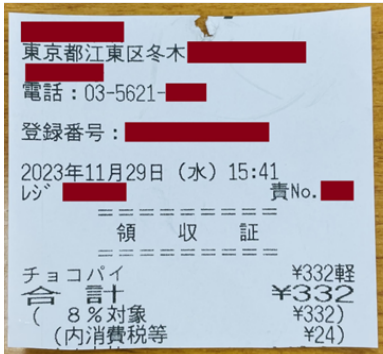
⁶ [生成 AI が日本の労働市場に与える影響③ 2024 年 02 月 06 日 | 大和総研 | 新田 堯之 \(dir.co.jp\)](#)

要求してみた（なお隠した「店名」についてどう回答するかに興味もあり、試してみた）。

店名については2回とも判読できないと回答してきた。2回目の税抜価格については、内訳である消費税等を引いた価格を推定値として回答している。これはおそらく膨大に学習してある知識の中で、税込価格・税抜価格を理解している、あるいは「内」という概念を理解していることから回答したものと推定される。

図表 11 GPT-4V による画像解析の例②

入力した画像 上段：プロンプト、下段：回答



1 回 目	貴方が受け取った画像を可能な限り正確に読み取って、店名、商品、 税込価格 を日本語で回答してください。
	店名は画像では隠されていますが、商品は「チョコパイ」で、税込価格は「¥332」です。
2 回 目	貴方が受け取った画像を可能な限り正確に読み取って、店名、商品、 税抜価格 を日本語で回答してください。
	店名は画像では部分的に覆われていて判読できません。商品名は「チョコパイ」です。税抜価格は「¥308」と推定されます。これは、税込価格「¥332」から内消費税等「¥24」を引いた金額です。

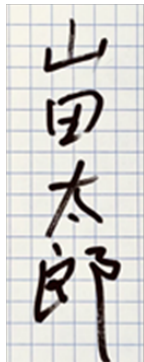
（出所）大和総研「画像解析 LLM アプリ」（2024 年 1 月）の実行結果（2024 年 2 月 22 日）

画像解析の進化例② ～ 手書き文字

図表 12 はメモ用紙に記載した手書きの文字を同様に GPT-4V に解析させた結果である。1 回目は、画像は日本語で書かれた「文字」と説明し、2 回目は「氏名」と説明し、抽出を指示した。

図表 12 GPT-4V による画像解析の例③

入力した画像 上段：プロンプト、下段：回答



1 回 目	貴方が受け取った画像は日本語で書かれた 文字 です。これについて、可能な限り正確に読み取り、日本語で回答してください。
	画像には、縦に書かれた四つの日本語の漢字が含まれています。上から順に、「山」「川」「大」「田」と読むことができます。
2 回 目	貴方が受け取った画像は日本語で書かれた 氏名 です。これについて、可能な限り正確に読み取り、日本語で回答してください。
	画像には、日本語の氏名が書かれています。それは「山田太郎」と読むことができます。

（出所）大和総研「画像解析 LLM アプリ」（2024 年 1 月）の実行結果（2024 年 2 月 22 日）

1 回目は「文字」と説明した関係からか、縦に書かれた漢字であるなどの特徴の回答は正しいが、「太郎」の字は読み取れておらず、字の順番も合っていない。2 回目は、「氏名」と説明した関係からか、おそらく膨大に学習してある人名からの補正が入り、正しく抽出されたもの

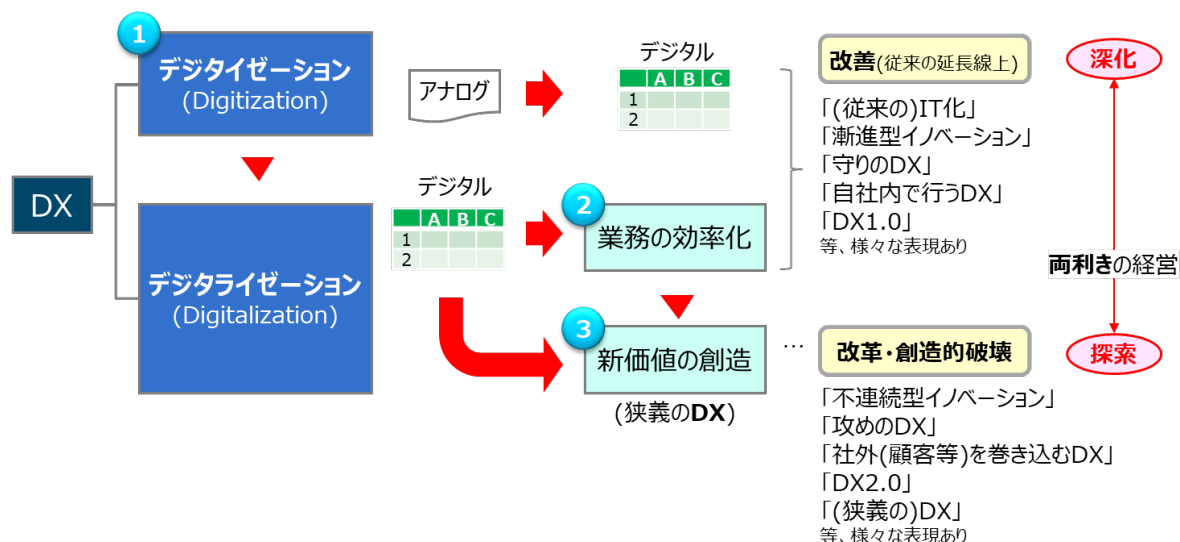
と推定される。漠然と文字の抽出を要求するのではなく、欲しい情報項目をプロンプト（質問・指示）に明示的に羅列して要求することで、より良い回答が得られている。

画像系処理への期待

2018年9月、経済産業省から「DXレポート」が公表され、2023年3月には情報処理推進機構（IPA）の「DX白書2023」が発刊された。ある設問の回答結果について、デジタルイゼーションとデジタルライゼーションの合計は米国と差がなくなっているが、デジタルトランスフォーメーション（DX）については米国とは大きな差がある旨の記載がある⁷。それゆえサブタイトルが「進み始めた『デジタル』、進まない『トランスフォーメーション』」になったと推測するが、米国との差の比較の詳細を見ると、「すでに十分な成果が出ている」と「すでにある程度の成果が出ている」の合計での比較であり、「すでに十分な成果が出ている」だけに限ると日本の回答（%）は米国の約3分の1である。

昨今日本ではトランスフォーメーション、つまりビジネスモデルの転換や、組織・制度・風土の変革の遅れがマスコミ等で指摘されることが目立つように思う。図表13は、デジタルイゼーション、デジタルライゼーション、DXの関係を記載したものだが、筆者は①のデジタルイゼーションがまだまだ不十分なのではないかと感じることが多い。例えば銀行では、来店せずにスマホアプリで各種変更手続きをできるようにする等デジタル化を進めており、それは素晴らしいことなのだが、特定の契約等においては来店時や郵送時に紙の伝票に手書きすることもまだ多い。

図表13 デジタルトランスフォーメーションの段階



（出所）大和総研 社外向け役員研修資料「ビジネスのためのデジタル化から、デジタル前提のビジネスへ」（2021年6月）より抜粋

紙伝票については Web 画面やスマートフォンからの入力への移行というデジタル化が本来策

⁷ DX白書2023 表紙～第1部 総論 (ipa.go.jp)

であることはいうまでもない。しかしながら総務省の令和 5 年版情報通信白書によれば、年齢階層別のインターネット利用率（2022 年）は 13 歳から 59 歳までの各階層では 90%を超える一方、60 歳から 69 歳層では約 87%、70 歳から 79 歳層では約 66%と利用率が低い⁸。高齢化社会、人生 100 年時代における顧客サービスを考慮すると、まだしばらくの間は紙伝票に対応せざるを得ないのが多くの業界での実情であろう。紙伝票に関わる業務効率化を追求する上で、DX の前提となるアナログ情報のデジタル化に寄与する生成 AI の活用は、現時点では有望な選択肢ではないだろうか。

なお「アナログ情報のデジタル化」については、既存技術として OCR⁹や、AI 技術を組み合わせることで文字の認識精度を向上させた AI-OCR がある。但し OCR 製品の多くは、あらかじめ利用者が読み取る項目の位置や領域を定義する、あるいは OCR 製品が自動解析した項目定義を補正する必要がある。しかし GPT-4V では図表 11・12 のように事前定義は不要である。

この特長を活かしてさくら情報システムは 2024 年 1 月、AI-OCR と生成 AI を組み合わせた「AI Textract（仮称）」を 6 月から提供すると発表した。OCR 製品には読み取った文字が違う場合に代替候補を提示して修正しやすくする等のような機能があり、いわば文字認識エンジンである GPT-4V が OCR 製品にとってかわるものではない。しかしながら、読み取る画像の品質が一定程度確保でき、GPT-4V を組み込むシステム構築力があれば、「アナログ情報のデジタル化」は内製化も可能になるであろう。

4. 検索拡張生成（RAG）の可能性

検索拡張生成（RAG）とは

RAG は、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、また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。

図表 14 に RAG の基本的な構造を示したが、①コンピュータが自然言語を扱えるようにドキュメント（テキストデータ）のベクトル（数値配列）化を行い、ベクトルデータベースに格納する。②検索において、検索の内容（テキスト）もベクトル化し、ベクトル同士の比較を行う類似検索（ベクトル検索）を行う。③その検索結果を LLM にプロンプト¹⁰として渡し、要約なり箇条書きなり、指示に応じた回答を生成させる。

ベクトル化(図内①)には「埋め込み (Embedding)」という技術を用いる。また LLM に指定した

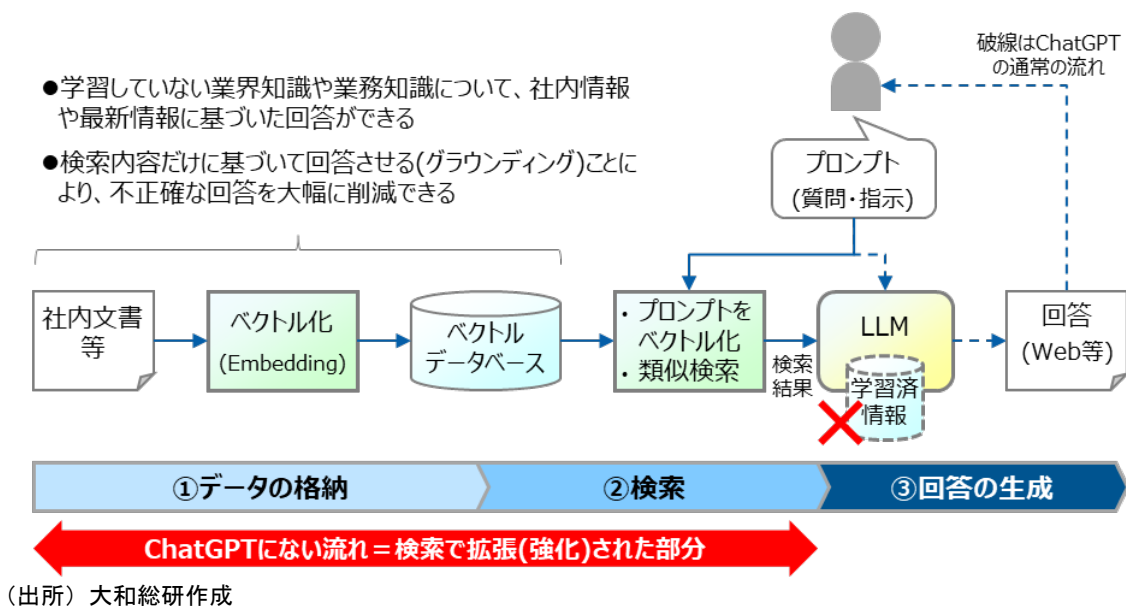
⁸ 総務省 | 令和 5 年版 情報通信白書 | インターネット (soumu.go.jp)

⁹ Optical Character Recognition/Reader. 光学的文字認識。手書きや印刷された文字をスキャナやデジカメ等で読み取り、コンピュータが利用できる文字コードに変換する技術

¹⁰ プロンプトについてはこれまで「質問・指示」と説明してきたが、プロンプトは主に、(1)質問・指示、(2)より良い回答のための外部情報・追加情報（コンテキストと言う）、(3)出力形式から構成される。RAG はこの(2)にベクトル検索の結果を用いる

情報（図内②の「検索結果」）だけに基づいて回答させる手法を「グラウンディング (Grounding)¹¹」という。なお類似検索は意味での検索が可能ということからセマンティック検索と呼ばれる場合もある。

図表 14 検索拡張生成 (RAG) の構造図



なおRAGについては昨今、当社へのレクチャーや支援のご要望をいただくことが増えている。図表 15 は当センターが社内公開している、RAG を用いた検索システムのデモンストレーション・アプリの実行結果である。見栄えは通常の検索画面とあまり変わらないが、類似度順に表示されること、また真偽判定のために容易に情報源の本文にジャンプできることが特徴である（検索対象が社内規定のためイメージ図となるが、ご容赦いただきたい）。

図表 15 RAG を用いた検索システムの例

社内情報検索LLMアプリ

社内の事について質問する

ABC休暇はいつからいつまでに取得することが可能ですか？

回答

ABC休暇は原則として10月から翌3月までの期間に取る事ができます。(「X規定14頁」を参照)

※結果はGPTが作成したものです。必ず真偽を確認してから利用してください

文書名	頁	該当部分	更新日	担当部署
X規定	14		202x/x/xx	人事部
X規定	12		202x/x/xx	人事部
Y規定	6		202x/x/xx	人事部
...				

類似性が高い順に表示

(出所) 大和総研「社内情報検索 LLM アプリ」(2023 年 7 月)の実行結果(2024 年 2 月 21 日。イメージ図)

¹¹ 裸足になって自分と地球を意図的に繋げて心身のバランスをとるといったスピリチュアルな考え方を意味するが、AI の分野では AI が言葉・概念と実世界(実データ、自社データ)を結び付けて理解することを言う

ベクトル化・類似検索における留意点

テキストデータのベクトル化においては、それまでの自然言語処理よりも高精度な手法として登場したのが、米グーグルが2013年に公開した「Word2Vec」である。これは単語をベクトル化するものであったが、単語の意味を計算する四則演算の例として「King - Man + Woman = Queen」が大変有名になった。そして2014年に考案された「Doc2Vec」では文章・文書全体をベクトル化できるようになった。その後、様々なベクトル化・ベクトル検索のツールが開発されている¹²。これらのツールは、類似度の計算や処理速度、保守・運用コスト等に差が見られるため、導入にあたっては実際に自社データ等を用いた技術検証・比較検証を行う必要がある。

また類似度の精度に大きく関わるのがベクトル化の単位（チャンクという）をどうするか、である。これは対象の文書が比較的短文の条文のようなものなのか、比較的長文の解説文なのか等の内容にもよる。一般的には200文字等の固定長とする、もしくは句点・段落等の可変長とすることが多い。固定長においては、チャンク間で少量のテキスト（10～15%）を重複させることが効果的とされている。この効果の度合いも検証事項となろう。

そして通常のAIのモデル開発でも事前作業として、学習データのクレンジング（標準化や異常値や重複データの除外等）を行うように、ベクトル化する対象文書の品質によっては、事前の整形等が必要になる場合もある。

既存のデータベース（RDB）のベクトル対応

業務系システムによく使われるリレーショナル・データベース（RDB）が、昨今、ベクトルデータを保持・活用できるよう進化してきている。この機能を用いれば、売上数値等の業務データと顧客との対話内容等のベクトルデータを一つのデータベースで管理できるようになる。図表15のベクトルデータベースはOSSのChromaDBを用いて構築しているが、ベクトルデータもRDBで扱えるメリットは大きい。

業務機能面でいえば、通常の属性や数値範囲等での検索とベクトル検索を一つのSQL¹³で実行できるため、アプリケーション開発者がベクトル検索のための特別なスキルを習得する必要がなく、また当然のことながら業務データとベクトル検索結果を突合する処理をコーディングする必要もない。保守・運用面でいえば、新たにサーバーを用意したり、サーバー内のRDBプロセスを増やしたりする必要がないため、バックアップ等の運用や管理の負担が増えない。

図表16の米オラクルのOracle Databaseでは、バージョン19cでベクトルの格納・検索は可能ではあったが、バージョン23cではベクトルデータ型を列属性として追加した。既存のパーティショニング（表や索引等の細分化）や並列検索（Oracle Real Application Clusters (RAC)）

¹² そのうちの 하나가前回のレポートでふれた米マイクロソフトの「Azure Cognitive Search」（現在は名称変更され、「Azure AI Search」）である

¹³ Structured Query Language の略。RDB のデータ格納場所であるテーブルに対するデータの追加、更新、削除等の操作や、テーブルの定義・権限設定等の管理を行う言語のこと

等が活用でき、ベクトル検索性能も大幅に向上しているという。

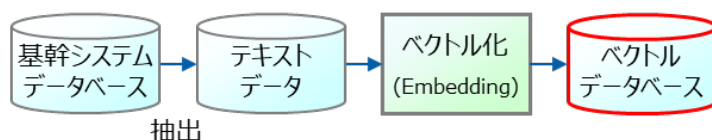
バージョン 23c ではデータ更新のタイミングでリアルタイムにベクトル化を行うため、既存業務処理への負荷が少し気になるころではあるが、もし影響がある場合はベクトル化の処理を夜間バッチ処理（一括処理）に切り出してベクトルデータの鮮度を落とすとしても、データベースが一元管理できる効果は極めて大きい。

なおバージョン 23c に対するベクトル検索は Oracle AI Vector Search を用いるが、これらは Oracle Cloud Infrastructure (OCI) というパブリッククラウドサービスでも利用できるし、OCI Dedicated Region というサービスを通してオンプレミス環境（自社環境）でも利用できる。

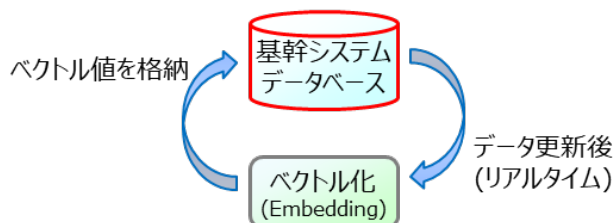
今後、社内データをいかにビジネスに活かしていくかを考える場合、このような RDB をベクトルデータベースとしても利用する手段は、選択肢の一つとして十分検討に値する。

図表 16 RDB のベクトル対応の効果

【通常のベクトル化】



【Database23c】



- 既存のデータベースをベクトルデータベースとして利用可能(一元管理)
- リアルタイムでベクトル化(同期更新)
- パーティショニング等による高速検索
- 1SQLで、業務データと文書データを検索(突合処理が不要)
- 基幹系の厳格なアクセス権限設定を流用

(出所) 日本オラクルの説明資料・報道内容等から大和総研作成

RAG への期待

RAG では自らが指定した信頼できる情報源からの回答になるため、情報の正確性が格段に向上する。また従来のキーワード検索では、キーワードの選択や聞き方によっては思い通りに情報を得られない場合が多いが、RAG では LLM のプロンプト（投入された質問・指示部分）に関する解釈能力と、意味による検索が可能なベクトル検索により、目的の情報が見つけやすい。

エンタープライズサーチ「Helpfeel」を提供する株式会社 Helpfeel が、従業員数 2,000 人以上の企業で働く 400 名を対象に行った「エンタープライズサーチ(企業内検索)に関する実態調査レポート」(2023 年 2 月)によると、社員が制度・ルールなども含む社内情報を調べている時間は 1 日当たり平均 1 時間 5 分であったという。

RAGにより10分間に短縮できると仮定すると、年間で55分間×240日(/年)÷8時間(/日)＝約27日(約1.4か月)の削減効果に相当する。この時間を事業拡大等への企画や遂行に充てるのはもちろんのこと、新規事業・新分野に必要な学習時間や、休暇取得等のウェルビーイングの向上に活かすことも可能となる。

人手不足が深刻化している昨今、情報検索・収集後の本務遂行に一層集中できるよう、RAGを始めとするAI、生成AIによる社内情報の活用は急務といえる。

5. 最後に

LLMが「何に使えるか」といった模索の段階は2023年で終わり、2024年は「どう活かすか」という段階、特にマルチモーダル化による音声や画像データの取込みや、自社内に蓄積された/される貴重な業務データを、いかにビジネスに活かすかという段階に確実に移ると考える。その結果、LLMを利用することで自らの業務・事業の効率化・高度化をイメージできる企業と、できない企業との差が大きく開いていくであろう。

AI・生成AIはコスト削減、CX(顧客体験)の向上、新たな収益機会の創出等、企業の競争力の強化に直結する。2024年は、AI・生成AI活用の巧拙が企業の生き残りの条件の一つとなる段階、言い換えるならば、AI・生成AIの開発や利用に取り組まないということが経営リスクとなる段階に突入するといっても過言ではない。

以上