2023 年 10 月 10 日

大規模言語モデルのビジネス活用の新展開

― ChatGPT のプロンプトエンジニアリングを超えた活用法と今後の展望 ―

先端技術ラボ　近藤 浩史、門脇 一真、工藤 剛

《要 点》

- 大規模言語モデル（LLM：Large Language Model）をベースとした対話型 AI サービス ChatGPT の活用が進んでいる。ChatGPT に投げかける指示文の工夫（プロンプトエンジニアリング）により処理能力を引き出すことで、ChatGPT はアイディエーションや文書作成の効率化に使用されている。一方で、ChatGPT をプロンプトエンジニアリングによって有効活用することを超え、より進んだ LLM のビジネス利用方法を模索する動きもある。
- 第一に、Web 上の一般的な情報に留まらずに、LLM に組織の固有知識（業務知識、業務マニュアルなど）を反映させて利用する方法である。RAG（Retrieval-Augmented Generation）と呼ばれる手法が代表的で、LLM が組織の内部文書を元に応答する手法である。また、LLM の性能が不足する場合には、LLM 自体をチューニングして用いる方法もある。
- 第二に ChatGPT 以外の LLM の活用である。様々な組織が ChatGPT と同等の LLM を開発し、公開している。LLM の実行環境に高いセキュリティを求める領域や、LLM に高い専門性を求める領域で、ChatGPT 以外のオープンな LLM をチューニングして用いる需要がある。また、様々な LLM を使いこなすためのツール群も開発されることで、組織において ChatGPT 以外の LLM も活用の選択肢になり始めている。
- 今後の展望として、①LLM への過度な期待が落ち着き、慎重に活用先を見極めるフェーズに移行すること、②LLM 活用の深化として、まずは RAG の活用が進展すること、③データセットの整備やモデル開発の共同化の動きが活発化すること、などが挙げられる。

本件に関するご照会は、先端技術ラボ 近藤、門脇、工藤 宛にお願いいたします。

Mail：101360-advanced_tech@ml.jri.co.jp

日本総研・先端技術ラボについては、以下をご覧ください。

https://www.jri.co.jp/company/business/system/advtechlab/

本資料は、情報提供を目的に作成されたものであり、何らかの取引を誘引することを目的としたものではありません。本資料は、作成日時点で弊社が一般に信頼出来ると思われる資料に基づいて作成されたものですが、情報の正確性・完全性を保証するものではありません。また、情報の内容は、経済情勢等の変化により変更されることがあります。本資料の情報に基づき起因してご閲覧者様及び第三者に損害が発生したとしても執筆者、執筆にあたっての取材先及び弊社は一切責任を負わないものとします。

１．はじめに

米 OpenAI は 2022 年 11 月に大規模言語モデル（Large Language Model; LLM）をベースとした対話型 AI サービス ChatGPT をリリースした。従来よりも自然な言語理解や生成を行い、対話形式での指示や質問応答ができることで世界中に衝撃を与えた。その後も、ビッグテックを中心に LLM に関連したサービスの発表[1]が続き、LLM への注目が続いている。

2023 年に入って、日本国内では ChatGPT の活用を掲げるビジネス書や Web 記事が多く出版・公開されている。その多くは、アイディエーションや社内文書作成・加工の効率化といった特定の目的のための活用法を取り上げたものである。ChatGPT の能力を引き出すためには ChatGPT に投げかける指示文（プロンプト）の工夫が重要とされ、その工夫する活動はプロンプトエンジニアリングと呼ばれる。

一方、執筆時点（2023 年 10 月初時点）では、ChatGPT をプロンプトエンジニアリングによって有効活用することを超え、より進んだ利用方法を模索する動きも出てきている。1 点目は、Web 等にある一般的な情報に留まらずに、LLM で組織の固有知識（業務知識、業務マニュアルなど）を扱う方法である。組織の内部文書を呼び出す方法や、LLM 自体をチューニングして用いる方法などがある。2 点目は、ChatGPT 以外の LLM の活用である。米 OpenAI 社以外の組織も、ChatGPT と同等の LLM を開発し、公開している。また、それらの LLM を使いこなすためのツール群も開発されることで、各組織において ChatGPT 以外の LLM も活用の選択肢になり始めた。

本レポートは、上述のような、プロンプトエンジニアリングを超えたビジネスのための LLM の活用動向について述べる。また、それらの動向を踏まえた今後の展望についても述べる。本レポートが、今後の LLM のビジネス活用動向の理解ならびに、各企業における活用検討の一助となれば幸いである。

２．ChatGPT 登場以降の振り返り

ChatGPT の登場以降、ChatGPT を含む LLM をビジネスに活用する方法の模索が活発となった。本節ではこの動きを 3 つに分けて振り返る。

（１） ChatGPT をそのまま活用する

まずは「ChatGPT をそのまま利用する」形態での活用が進展した。企業や自治体の組織内における業務効率化の観点から、主に一般的な情報の検索、文書のドラフト作成、要約、翻訳、アイディエーション等に利用されている。例えば、横須賀市では文書作成事務において 22,700 時間/年の業務時間短縮が想定されると 2023 年 6 月に公表[2]している。

ChatGPT をそのまま利用する代表的な方法として、①米 OpenAI の Web サービスを利用する方法、②米 OpenAI の API を利用する方法、③Azure OpenAI Service を利用する方法の 3 つが存在する[3]。ChatGPT の登場直後は、企業の中には情報漏洩を防止する観点などから ChatGPT の利用を制限する動き[4]があった。その中で、③Azure OpenAI Service は、セキュリティを確保した環境を構築可能となった[5]ことから、企業での採用が拡がった。例えば、金融[6]や保険[7]などの高いセキュリティを求められる企業で、Azure OpenAI Service の採用が進んでいる。

同時に、一部の組織では利用ルールの整備が進んだ[8]。ChatGPT には前述の情報漏洩のリスクに加え、事実と異なる不正確な文が生成される課題などがあり、ChatGPT を妄信した業務利用にはリ

スクも伴う。組織内のルール整備が完了してから ChatGPT の利用を開始するなど、リスクを低減する形で導入に至る事例も複数公表されている[9]。

セキュリティを確保した環境の構築が可能となったこと、利用ルールの整備が進んだことなどにより、ChatGPT をそのまま利用する形態での活用は拡大している。株式会社エクサウィザーズの調査(2023 年 8 月時点)[10]によると、生成 AI を業務で日常的に活用する人は 2023 年 4 月時点から大幅に増加した。

（2） 組織内の知識を反映させる

組織の固有知識（業務知識、業務マニュアルなど）を踏まえて LLM を活用するニーズが顕在化している。組織の固有知識の多くは、営業秘密として組織内に保有され、Web 上には公開されていない。一方で、LLM は Web 上のテキストデータで学習されたモデルが多く、組織の固有知識を反映した文は生成できない。このことが LLM の業務活用の幅を限定的にしている。

このニーズを満たす手法として RAG（Retrieval-Augmented Generation）と呼ばれる手法がある。これは、外部の情報源から情報を検索し、その検索結果を踏まえて LLM が文を生成する手法である。外部の情報源として組織内の文書を用いることで、組織の固有知識を反映させて LLM を活用することができる。

RAG の概要の理解のため、組織内に何らかのチャットシステムが存在する場合を想定して、仕組みの概要を説明する（図 1）。まず、利用者がチャットシステムにテキストで質問する（図 1 -①）。その質問の回答に関連しそうな文書を文書 DB から検索[11]し、結果を得る(図 1-②)。検索結果と質問文を LLM に渡す(図 1-③)。このとき、LLM は単に自身が持っている知識だけでなく、関連する組織内の文書（知識）を踏まえて回答を生成できるようになる。また、組織内の文書に基づいて回答するため、誤った回答を生成する可能性も低減できる。最後に LLM が生成した回答文をユーザに返答する(図 1-④)。

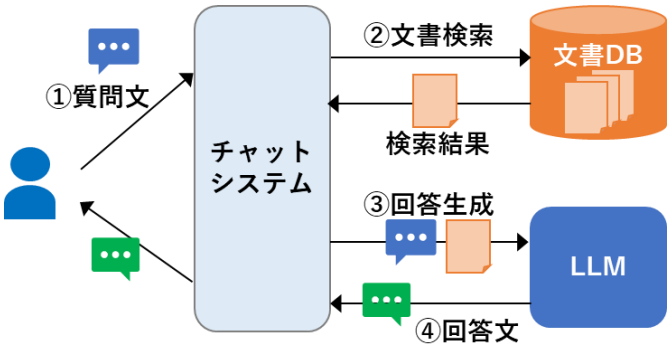

図 1：組織内の知識を取り込む手法の簡易図

現在、LLM の活用に積極的な組織は、RAG の仕組みについて実証実験(PoC)や社内環境の整備を進めている。RAG の仕組み自体をサービスとして提供する事例[12]もあり、RAG の仕組みを実現するハードルは今後低くなっていくと考えられる。

（3） 大規模言語モデルのチューニング

先進的な組織では、組織内のデータを用いて LLM 自体のチューニングに取り組んでいる[13]。ChatGPT などの LLM を初期状態で使用する場合、プロンプトエンジニアリングだけでは LLM の性能を引き出すことが困難なことがある。特定の業務ユースケースでの性能を向上させたいなど、既存の LLM では期待する性能水準に達しない場合にチューニングが用いられる。特定ユースケースでの性能向上が、顧客に提供するサービスの競争優位性の向上や、自組織内の業務効率化に大きく寄与する場合には、チューニングのノウハウをいち早く獲得することが重要と言える。

LLM のチューニング手法を大別すると、①言語モデルの追加学習、②特定タスクに特化した学習、

③指示に対応するための学習(Instruction Tuning)、④人間のフィードバックによる強化学習(RLHF : Reinforcement Learning from Human Feedback)の 4 つがある(図 2)。米 OpenAI 社の GPT モデルは、OpenAI API[14] または Azure OpenAI Service[15]を通じて、②と③を実施できる。また、米 Google 社の LLM である PaLM2 の一部のモデルは、④も実施できる[16]。

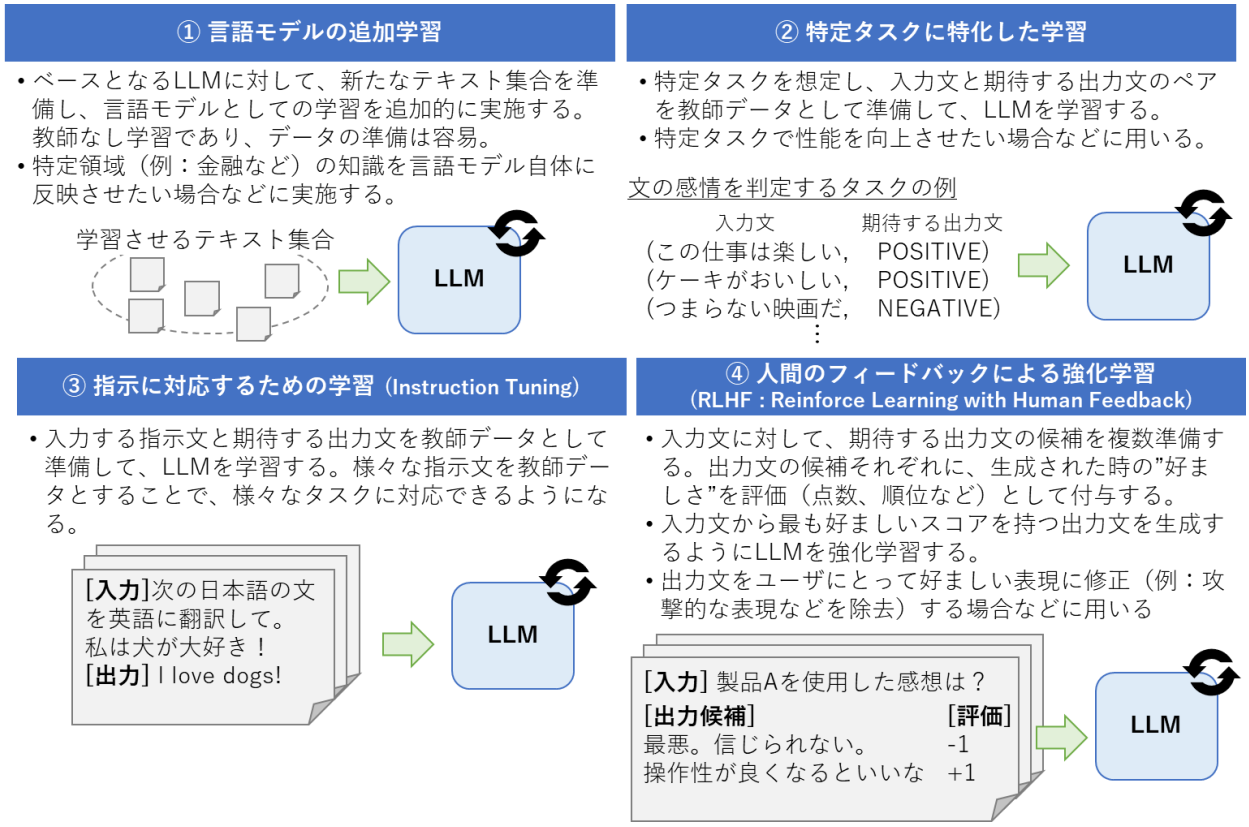

図 2 : LLM のチューニング手法

これらのチューニング手法のうち、特に③指示に対応するための学習（Instruction Tuning）を適用する取り組みが進展している。Instruction Tuning を行うには、チューニング用データの整備が重要となる。Instruction Tuning により LLM の性能が向上することは様々な研究[17]で指摘されているが、LLM に望ましい挙動をさせるために適した（例えば、データ量が十分にある、多様性がある、一貫性がある）データセットの整備方法は研究途上である。データセットの整備には、人手での作業を要し、作成ガイドを整備するなどの品質確保の取り組みが重要であり、コストや時間も要する。なお、Instruction Tuning のデータセットを LLM 自身に作成させることで性能向上を目指す研究[18]も行われており、今後はデータセットの作成を省力化するための研究が活発になるであろう。

なお、従来、LLM などの大規模なモデルのチューニングには大規模な計算資源（CPU、メモリ、GPU など）が必要であった。これはチューニング時に全てのモデルパラメータを更新していたからである。2023 年に発表された QLoRA[19]など、一部のパラメータのみを更新することで効率性を高める手法の研究も進んだことで、大規模な計算資源を確保できない組織でもチューニングを行うことも可能となりつつある。

3．注目すべき動向

（1） オープンな LLM の開発

米 OpenAI（GPT-4）や米 Google（PaLM2）が提供する LLM の技術詳細[20]、モデルパラメータ、ソースコードなどは非公開である。LLM をベースとしたサービスが、Web インターフェースや API を通じて提供されているのみである。また、LLM のカスタマイズやチューニングもサービスとして提供された範囲に限定される。そのため、一部の企業が LLM の技術を独占することを懸念する声があがっている。

一方、ベンチャー企業、大学、有志コミュニティなど、様々な組織が独自に開発した LLM を公開する事例が増加している。そのなかには商用利用が可能な LLM もあり、LLM 自体の開発が困難な組織でも、活用可能な LLM の選択肢が増えつつある。モデルサイズ、特定言語(日本語など)や特定ドメインへの特化などの点で、様々な特徴を持つ LLM が公開されており、適切な LLM を選択して活用することが重要である。例えば、日本語での利用を念頭に、日本語に特化した LLM の代表例を表 1 に示す。

オープンな LLM の中には、他のオープンな LLM に対して、特化させたい部分だけチューニングして公開する事例も増加している。特に、米 Meta が公開した Llama2[21]をベースとした事例が増加している。Llama2 は技術詳細が公開されていること、オープンな LLM の中でも大規模なモデルであり性能が良いことから、チューニングのベースモデルとして使用されている。

GPT-4 や PaLM2 などの LLM が API などで提供されているなかで、オープンな LLM を独自に活用する動機を整理すると表 2 のようになる。これらの動機を持つ組織はオープンな LLM の活用も検討の余地がある。

なお、様々な LLM が公開されているが、その仕様情報のみから LLM の性能の優劣を判断することは困難である。LLM の性能はユースケースや使用するデータによって変わるため、オープンな LLM を組織内で評価する環境を整備し、どの LLM が使えそうか試用することも必要であろう。

表 1：日本語に特化した LLM の代表例

公開年月	名称	開発組織	最大パラメータ数(※)	商用利用可否	備考
2023/5	japanese-gpt-neox-3.6b	rinna	3.6B	○	GPT-NeoXをベースに学習
2023/5	open-calm-7b	サイバーエージェント	7B	○	
2023/7	japanese-mpt-7b	Lightblue	7B	○	MPT-7Bをベースに学習
2023/8	japanese-large-lm	LINE	3.6B	○	
2023/8	gpt-neox-japanese-1.4b	ストックマーク	1.4B	○	GPT-NeoXをベースに学習
2023/8	Japanese StableLM	Stability AI	7B	○	
2023/8	Weblab-10B	東京大学 松尾研究室	10B	×	
2023/8	ELYZA-japanese-Llama-2-7b	ELYZA	2.7B	○	Llama2をベースに学習
2023/9	PLaMo-13B	Preferred Networks	13B	○	

(※)BはBillion(10億)の略

表 2：オープンな LLM を活用する動機

動機	解説
セキュリティの向上	• 自組織データセンター内の基盤上に、クラウド環境に送信できない機密性の高いデータを入力可能なLLMを構築する。
高度な専門領域または特定言語の処理能力向上	• 高度な専門知識（例：医療、金融など）や、日本語などの特定言語の知識をLLMに追加学習させ、処理野力を向上させる。
提供するAIに責任を持つ	• クラウドで提供されるLLMの技術詳細が不明である場合、当該LLMを使用したアプリの挙動について、信頼性・安全性・公平性などの責任を確保することが難しい。 • 独自のLLMを開発・運用することで、挙動に一定の責任をもってユーザにアプリを提供できる。
従量課金からの脱却	• クラウドで提供されるLLMは、使用量に応じた従量課金が一般的である。 • LLMで複雑な処理を実行する場合、使用料が膨らみがちである。独自のLLMを開発・運用することで、使用料金を固定できる（ただし、独自LLMの開発・運用コストは生じる）。
バージョンコントロール	• クラウドで提供されるLLMが廃止またはアップデートされることで挙動が変わり、そのLLMを使用していたアプリに不具合が生じることがある。 • 独自のLLMを開発・運用することで、LLMの挙動を固定できる。
処理性能の確保	• 2023年10月初時点では、クラウドで提供されるLLM利用時の処理速度が遅い、処理上限にすぐ達してしまう、といった意見もある（クラウド側のアップデートにより、将来的に解消する可能性は高い） • 高い処理性能を確保するために、オープンなLLMを使って独自に計算基盤を確保する
エッジでの稼働	• クラウドとの通信が難しいケースでLLMを使用したい

（2） LLM を活用したサービスの開発ツールの整備

2023 年 9 月までに、LLM を活用したサービス（但し、プロトタイプの水準）を簡単に開発するためのツールが整備されてきた。代表的なツールとして、米 Microsoft が OSS として公開した Semantic Kernel[22]（OpenAI や Azure OpenAI Service との親和性が高い）や、LangChain[23]などがある。これらのツールは、①LLM との入出力をインターフェースする機能、②LLM との対話履歴を記憶する機能、③LLM に外部データを投入し、外部データをもとに回答を生成させる機能、などがある。これらの機能を組み合わせることで、2.(2)で述べた RAG を比較的容易に実装できる。また、自組織が提供するアプリケーションに LLM を組み込むことも比較的容易に実現できる。

プロトタイプの水準を超えて、実際の業務サービスを開発する際に使用できるツールも開発されている。具体的には、いわゆる LLMOps と呼ばれる概念を実現するツールである。LLMOps は、従来の AIOps または MLOps といった概念に対して LLM 特有の要素を追加した概念であり、LLM の開発から運用管理までのサイクルを効率化・自動化するための実践のことである[24]。LLM 特有の要素の一例として、従来の AI では存在しなかった、プロンプトエンジニアリングとその管理のための実践がある。

LLMOps の実現を目指す具体的なツールには、LangSmith[25]、Weight & Bias[26]、Prompt Flow[27]などがある。例えば LangSmith では、LLM の実行ログを収集して確認する機能や、開発したモデルを評価する機能などが提供されている。

LLM を活用したサービスの開発には、アジリティの高い開発組織を整備することが重要である。LLM 自体の進展が早いこと、また、紹介した各ツールは登場して間もないことから、ツールが改良されていくと想像する。開発した資産が再利用しにくくなるなどのデメリットもありうるが、変化に迅速に対応しながら積極的に活用するべきと考える。

４．今後の展望

（１）LLM への期待が落ち着く

2023 年初は ChatGPT に対する過度な期待の声がよく聞かれた。現在は冷静に適用先や使用方法を見極めようとする組織が多い。この理由は２つある。第一に、ChatGPT が人間の文書作成を代替する「夢のツール」ではなかった点が挙げられる。前述の通り、LLM は誤った情報を生成することがあるため、LLM の出力に対して人間のチェックが必須となる。これでは、一定の業務効率化が図れたとしても、完全な自動化にはならない。

第二に、LLM は不適切なコンテンツ（差別的、攻撃的なコンテンツなど）を生成する可能性があり、企業の顧客向けサービスへの適用が難しい点がある。企業の顧客に不適切な回答を返し、企業の信頼が失墜するなどのリスクがある。一方、LLM のチューニング手法が確立することで、誤情報の生成や、不適切なコンテンツの生成などの課題は緩和される可能性もある。

今後、LLM の活用に着手する組織は、過度な期待に基づき安易な PoC を繰り返すのではなく、LLM の特性に合ったユースケースに対して PoC を実施することが重要である。PoC に投資したが成果が得られない、いわゆる PoC 貧乏にならないようプロジェクトを推進することを推奨したい。

（２）LLM 活用の深化

2.(2)で紹介した RAG を活用し、組織内の知識を LLM の出力に反映させるための実証検証や環境整備に力を入れる組織が増加すると考える。RAG の仕組み自体の構築には、パッケージ製品や社外の技術者を活用する場合であっても、自組織自身が取り組むべき推奨事項がある。それは、組織内の有用なドキュメントを、LLM と連携できる場所に収集しておくことである。組織内では、関連が高い情報であっても、用途や部署別に別々のシステムやストレージに分散されていることもあり、上手く活用できないといったことも想定されるためである。

また、自組織が提供するアプリケーションに LLM を組み込む動きも拡大すると見込まれる。LLM 自体と周辺の開発ツール（LangChain など）の進化スピードが速いため、内製開発が可能な組織であれば、自組織内で素早くプロトタイプを作成し、試行できるようにすることが望ましい。また、実現したいユースケースに応じて、GPT-4 以外の適切なモデルを選択できるように、複数の LLM を使いこなせるように環境整備や特徴の検証を進めておくと良いであろう。

2.(3)で紹介したチューニングを、独自に実施可能な組織は限定されるであろう。なぜなら、チューニングには、システムエンジニア、AI や LLM の知識を有する人材、計算資源の確保が少なからず求められるためである。そのため、一般の組織がチューニングに取り組む場合は、外部の IT ベンダーの力を借りることか、または、後述するように複数組織が協力して取り組むことが現実的であろう。一方、内製開発が可能な組織や、小規模な計算資源を持つ組織では、少ない計算資源で LLM をチューニングする手法(2.(3)に記載)などを用いて、チューニングを試行錯誤しても良いであろう。

（3）共同でのデータセット整備／モデル開発

LLM 自体の開発には大規模な計算資源や大量のデータが必要となる。そのため、モデルの開発や、データセットの整備を、政府が主導することや、複数組織が共同で行う事例が増加することが想定される。

既に検討・着手されているモデル開発の事例としては、国立情報学研究所(NII)を中心に大学・企業が集まり、GPT-3 の規模のモデルを構築して原理解明に取り組む体制を作ることを目指した LLM 勉強会[28]が開催されている。データセット整備の事例として、理化学研究所では、複数の民間企業が参加する形で日本語のインストラクションチューニング用のデータセット構築[29]が始まっている。また、日本政府においては、AI 戦略会議にて総務省・NICT が整備する学習用言語データの提供について議論されている[30]。

上記のデータセット整備の事例では、汎用的な用途でのデータ整備を目指しているが、今後は業界ごとに専門的なデータを整備することも重要と考える。例えば、金融領域に特化した学習データを金融業界が構築するといったことである。業界内では競争関係にあったとしても、基盤となる LLM を協力して整備することで、業界自体の活性化、世界における競争力の強化に資すると考える。また、データを学術界に還元することで、日本の LLM 研究にも貢献できるのではないだろうか。

（4）テキストと画像を組み合わせた活用の進展

近年、画像とテキストを同時に処理する AI（マルチモーダル AI とも呼ばれる）の技術が進展している。与えられた画像を説明するテキストを生成する AI や、与えられたテキストの内容を表す高品質な画像が生成する AI が登場している。

2023 年 9 月末に画像認識の機能を持った対話型 AI である ChatGPT(GPT-4V) [31]が利用できるようになった。ChatGPT にアップロードした画像の内容を踏まえ、AI と対話できる。また、2023 年 10 月には、DALL-E 3[32]とよばれる画像生成モデルを ChatGPT 上で使用できるようになった。生成する画像の細部の条件をテキストで指定して、画像が生成できる。

ChatGPT の登場によりテキストを処理する AI に注目が集まっているが、テキストと画像を組み合わせることで、ビジネス上の応用範囲がより広がることが想定される。実際に、執筆時点で、RAG の仕組みの整備を進める組織においては、次の課題として、組織内文書中の図表を上手く活用できないという課題が生じている。これは、図表はテキストデータではなく、LLM への入力情報としてそのまま活用できず、文の生成に反映できないためである。GPT-4V のようなテキストと画像を同時に処理できるモデルの開発が進展することにより、この課題は解消に向かう可能性がある。

（5）ＡＩ開発の指針やガイドラインの動向

G7 や日本政府は、AI 開発に関する指針や統合ガイドライン(案)の作成を進めており、2023 年末に公表される見通しである。2023 年 9 月 7 日には G7 広島 AI プロセスの閣僚級会合が開かれ[33]、AI システムに関する国際的な指針や行動規範を 2023 年中に策定する方向性が示されている。また、日本国内に存在していたいくつかの AI に関するガイドラインを統合し、新 AI 事業者ガイドライン（案）が 2023 年中に提示される見込みである。2023 年 9 月 8 日にガイドラインのスケルトンが公表[34]された。

LLM の開発・運用を実施する組織は、これらが公表されたら速やかに自社への影響の有無を確認する必要があるだろう。

執筆者

近藤 浩史（Hirofumi Kondo）

2011 年入社。三井住友銀行の決済系／市場系システムの企画・開発業務等を経て、2018 年から現職。AI 分野全般の技術調査・検証に従事。

門脇 一真（Kazuma Kadowaki）

2011 年入社。三井住友カードのチャネル系システムの開発業務を経て、2017 年から現職。自然言語処理の技術調査・検証に従事。

工藤 剛（Tsuyoshi Kudo）

2014 年入社。三井住友銀行の市場系システムの開発業務や、三井住友銀行のデータ分析業務を担当。2020 年から現職。自然言語処理の技術調査・検証に従事。

参考文献・補足

[1] 例えば Microsoft 365 Copilot や Duet AI for Google Workspace などがある

[2] 横須賀市. 「ChatGPT 活用実証結果報告」. 令和 5 年 6 月 5 日. https://www.city.yokosuka.kanagawa.jp/0835/nagekomi/documents/yokosuka-chatgpt-2-houkoku.pdf (参照:2023-10-2)

[3] その他、明示的に ChatGPT を使わずとも、Bing Chat 検索、Microsoft Edge の Copilot 機能として利用しているケースもある

[4] ChatGPT、ソフトバンクなどが利用制限 ルール作り急ぐ. 日本経済新聞. 2023 年 3 月 11 日. https://www.nikkei.com/article/DGXZQOUC069HD0W3A300C2000000/ (参照:2023-10-2)

[5] 具体的には、Azure OpenAI Service では、データ保護の機能（データ暗号化、入力データをモデルの再学習に利用しないことなど）を備えていること、閉域化されたネットワーク環境で利用できることなどがあげられる。(参照：Microsoft Learn. 「Azure AI サービスの仮想ネットワークを構成する」. https://learn.microsoft.com/ja-jp/azure/ai-services/cognitive-services-virtual-networks (参照:2023-10-2))

[6] SMBC グループの専用環境における AI アシスタントツール「SMBC-GPT」の実証実験の開始について. 三井住友銀行ニュースリリース. 2023 年 4 月 11 日. https://www.smfg.co.jp/news/j110416_01.html (参照:2023-10-2)

[7] 保険領域に特化した対話型 AI の開発および活用の開始. 東京海上日動火災保険株式会社ニュースリリース. 2023 年 4 月 19 日. https://www.tokiomarine-nichido.co.jp/company/release/pdf/230419_01.pdf (参照:2023-10-2)

[8] 一般社団法人日本能率協会が実施したアンケート(n=1265)によると、「ChatGPT 利用ガイドラインや社内規則はありますか？」との質問に「作成済み、作成中、作成検討中」と答えた割合は 68.8%であった。(出所：一般社団法人日本能率協会,「「ものづくり人材が ChatGPT を使いこなす方法」講演会 参加者事前アンケート調査結果報告書」. 2023 年 8 月 22 日. https://www.jma.or.jp/img/pdf/pdf-2023-chatgptreport.pdf (参照:2023-10-2))

[9] 神戸市. 「ChatGPT の試行利用を開始します～独自の利用環境のもと本格利用に向けた検討を進めます～」. 2023 年 6 月 22 日. https://www.city.kobe.lg.jp/a08691/886672664922.html (参照:2023-10-5)

[10] ChatGPT など生成 AI を「業務で日常使用」は 2 割、4 カ月で 13 ポイント増、全社導入で利用が定着. 株式会社エクサウィザーズ ニュースリリース. 2023 年 9 月 7 日. https://exawizards.com/archives/25353/ (参照:2023-10-2)

[11] 組織内の文書を Embedding と呼ばれる数値ベクトルに変換しておき、ベクトル空間上で検索する手法が一般的。数値ベクトルを保存でき、かつ、効率よく検索可能なベクトル DB を用いる。

[12] Azure AI Services Blog. 「Introducing Azure OpenAI Service On Your Data in Public Preview」, 2023 年 6 月 19 日, https://techcommunity.microsoft.com/t5/azure-ai-services-blog/introducing-azure-openai-service-on-your-data-in-public-preview/ba-p/3847000 (参照:2023-10-2)

[13] NEC 社が提供するカスタマープログラムでは、NEC 社と参加企業が「LLM を活用した企業内データによる個別チューニングを実施する」としている。(出所：NEC、日本市場向け生成 AI を開発・提供開始 ～業種ナレッジの構築を目指したカスタマープログラムを開始～. NEC ニュースリリース. 2023 年 7 月 6 日. https://jpn.nec.com/press/202307/20230706_01.html (参照:2023-10-2))

[14] OpenAI Platform. 「Fine-Tuning」. https://platform.openai.com/docs/guides/fine-tuning (参照:2023-10-2)
[15] Microsoft Learn. 「Azure OpenAI Service を使用してモデルをカスタマイズする」. https://learn.microsoft.com/ja-jp/azure/ai-services/openai/how-to/fine-tuning (参照:2023-10-2)
[16] 執筆時点で正式リリース前のプレビュー段階である。(参考： Google Cloud ブログ. 「Google Cloud、I/O にてジェネレーティブ AI を加速 - Vertex AI の新しい基盤モデル、エンベディングとチューニング ツール」. https://cloud.google.com/blog/ja/products/ai-machine-learning/google-cloud-launches-new-ai-models-opens-generative-ai-studio. (参照:2023-10-2）)
[17] Jason Wei et al. Finetuned Language Models are Zero-Shot Learners, ICLR2022, https://arxiv.org/pdf/2109.01652.pdf
[18] Yizhong Wang et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions, 2023, https://arxiv.org/abs/2212.10560
[19] Tim Dettmers, et al. QLoRA: Efficient Finetuning of Quantized LLMs. 2023. https://arxiv.org/abs/2305.14314
[20] テクニカルレポートなどの形で一部の仕様は開示されている。
[21] Meta. 「Llama 2 – Meta AI」. https://ai.meta.com/llama/ (参照:2023-10-2)
[22] Microsoft Learn. 「Semantic Kernel documentation」. https://learn.microsoft.com/ja-jp/semantic-kernel/ (参照:2023-10-2)
[23] LangChain. 「LangChain」. https://docs.langchain.com/docs/ (参照:2023-10-2)
[24] LLMOps は登場したばかりの概念で、現時点で統一的な見解は存在しない認識。ここでは著者らの見解で記載。
[25] LangSmith Docs. 「LangSmith」. https://docs.smith.langchain.com/ (参照:2023-10-2)
[26] Weights & Biases: The AI Developer Platform. https://wandb.ai/site (参照:2023-10-2)
[27] Microsoft Learn. 「Get started with Prompt flow (preview)」. https://learn.microsoft.com/en-us/azure/machine-learning/prompt-flow/get-started-prompt-flow (参照:2023-10-2)
[28] LLM 勉強会. 「LLM 勉強会」. https://llm-jp.nii.ac.jp/ (参照:2023/10/2)
[29] RIKEN-AIP, LIAT. 「LLM のための日本語インストラクションデータ作成プロジェクト」. https://liat-aip.sakura.ne.jp/wp/llm%E3%81%AE%E3%81%9F%E3%82%81%E3%81%AE%E6%97%A5%E6%9C%AC%E8%AA%9E%E3%82%A4%E3%83%B3%E3%82%B9%E3%83%88%E3%83%A9%E3%82%AF%E3%82%B7%E3%83%A7%E3%83%B3%E3%83%87%E3%83%BC%E3%82%BF%E4%BD%9C%E6%88%90/ (参照:2023-10-2)
[30] 総務省・NICT が整備する学習用言語データのアクセス提供について. 内閣府 ＡＩ戦略会議 第５回 資料 3-4. 令和５年９月８日. https://www8.cao.go.jp/cstp/ai/ai_senryaku/5kai/datateikyou.pdf (参照:2023/10/2)
[31] OpenAI Updates. 「ChatGPT can now see, hear, and speak」. 2023 年 9 月 25 日. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak, (参照:2023-10-2)
[32] OpenAI Updates. 「DALL·E 3」, 2023 年 9 月 20 日. https://openai.com/dall-e-3 (参照:2023-10-2)
[33] Ｇ７広島 AI プロセス 閣僚級会合の概要. 内閣府 ＡＩ戦略会議 第５回 資料 1-1. 令和５年９月８日. https://www8.cao.go.jp/cstp/ai/ai_senryaku/5kai/kakuryoukyuu.pdf (参照:2023/10/2)
[34] 新ＡＩ事業者ガイドライン スケルトン（案）. 内閣府 ＡＩ戦略会議 第５回 資料 1-2. 令和５年９月８日. https://www8.cao.go.jp/cstp/ai/ai_senryaku/5kai/gaidorain.pdf (参照:2023/10/2)