IT Leaders × シンクイット Think IT™ | 特別編集号 | 2024 Summer
本冊子は株式会社インプレスの Web メディア「IT Leaders」と「Think IT」から注目の記事を抜粋・再編集した特別編集号です。

生成AI ChatGPT［最新像と基礎知識］

検討期から普及期へ！
ここまで進んだ「先行企業の生成AI“独自”活用」

開発のしかたが変わる！
全エンジニア必須の「ChatGPT/LLM“超”基本」

impress

グループ全社のデータ分析力向上へ「生成AI×データ活用」の新たな実践

グループ全社を挙げてデータドリブン経営に邁進する日清食品ホールディングス。2023年からは、データマネジメントの取り組みに生成AIの活用を加えて、大規模データベースからAIが自動でレポーティングする仕組みの構築などに取り組んでいる。2024年3月8日に開催された「データマネジメント2024」（主催：日本データマネジメント・コンソーシアム〈JDMC〉インプレス）のセッションに、日清食品ホールディングス 執行役員 CIO グループ情報責任者の成田敏博氏とデータサイエンス室の小郷和希氏が登壇し、取り組みを紹介した。　神 幸葉（IT Leaders編集部）

データドリブン経営の起点となる全社データ基盤

「カップヌードル」「チキンラーメン」などの国民的ブランドをはじめとした数々の商品で食の世界を追求する日清食品ホールディングス。堅調に見える同社だが、社内では、“カップヌードルシンドローム”と呼ぶ、新しいことへの挑戦や現状の見直しを怠る大企業病のような停滞への危機意識がある。経営トップが折に触れて、グループ全社に向けてこの言葉を発して挑戦や変革を促しているという。

近年の同社はグループを挙げた経営変革の一環として、「DIGITIZE YOUR ARMS（デジタルを武装せよ）」をスローガンに掲げ、さまざまな領域で現場主導型の業務デジタル化を推進している。これまで、脱・紙文化、常時テレワークが可能な環境整備、業務の効率化/自動化、AI活用によるルーチンワーク削減などに次々取り組み、2025年には完全無人の工場ラインの実現などを視野に入れる。

そして、同社のあらゆる取り組みの根幹にあるのがデータドリブン経営だという。全社データ基盤の構築、共通マスターの正規化、データ活用専任組織の設置などが進んでいる。

日清食品ホールディングス 執行役員 CIO グループ情報責任者の成田敏博氏（写真1）は、日清食品のデータ分析力の成熟度について、「限定的」から「組織的な強化」へと取り組みのレベルを上げている段階とし、「2025年には社内のさまざまな部門でデータ分析・活用が行われているレベルを目指す」とした（図1）。

データドリブン経営に着手した頃、グループを横断したデータ連携/分析基盤が存在せず、社内のデータが各部門に分散していたという。現在では、データウェアハウスにあらゆる業務データを集約し、各種のデータを同じ軸で分析可能なデータ基盤を整えている。その下で、現場のスタッフがBIツールを使って、みずから手を動かせる仕組みづくりが進んだ（図2）。

例えば、資材本部では資材情報を統合化し、データに基づく調達戦略の策定がなされ、経営企画本部ではグローバルな経営ダッシュボードを構築し、意思決定に役立てているという。

「情報システム本部がデータ連携/分析基盤を整備して、その上でグループ本部クラスの業務に活用している段階だ。今後は各部門にも展開して、全社的・組織的なデータ活用を促していく」（成田氏）

サプライチェーン/セールス部門のデータ活用事例

事業部門主導のデータ活用の事例として、日清食品ホールディングス データサイエンス室の小郷和希氏（写真2）が、同社サプライチェーン部門、セールス部門での取り組みを紹介した。

サプライチェーン部門では、地政学、

写真1　日清食品ホールディングス 執行役員 CIO グループ 情報責任者の成田敏博氏

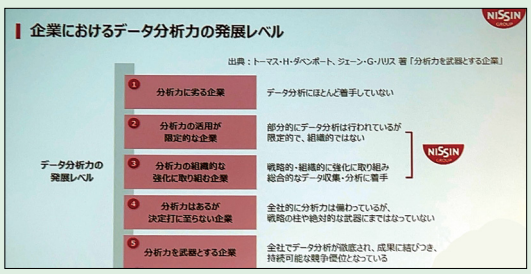


図1　企業におけるデータ分析力の発展レベル（出典：日清食品ホールディングス）

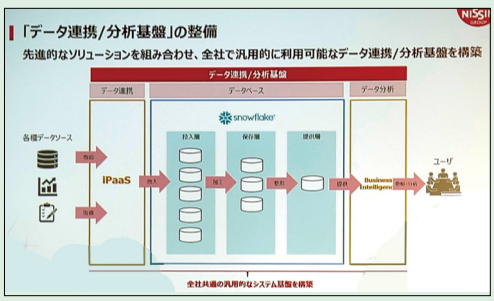


図2　データ連携 / 分析基盤の概要（出典:日清食品ホールディングス）

写真２　日清食品ホールディングス データサイエンス室の小郷和希氏

図３　共通データベースとAIの連携（出典：日清食品ホールディングス）

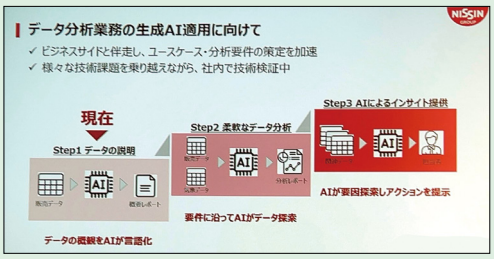


図４　データ分析業務の生成AI適用ステップ（出典：日清食品ホールディングス）

経済、パンデミックなど、経営を取り巻く環境変化の中、担当者には資材に関わる情報のスピーディかつ正確な可視化が求められている。そこで同部門は、データに基づいた迅速なアクションを可能にすべく資材統合データベースの構築に取り組んだ。

資材統合データベースは上述の全社データ基盤を基に構築されている。例えば、「ウクライナ情勢の影響でコリアンダーの調達が滞りそうだ」と判明した場合、資材担当者がサプライヤーに代替国からの仕入れを打診し、生産担当者が生産計画の見直しを指示するといったアクションを可能にする。

セールス部門の活動もデータ連携/分析基盤を起点にする。例として、ID-POSデータから商品単位で購入の傾向をモデル化して潜在顧客分析を実施。「高頻度でカップヌードルを購入する顧客と類似傾向がある顧客」といった情報の特定を行い、潜在顧客をターゲティングする。

「顧客Aはカップヌードルのほか、サバ缶、七味唐辛子、ビールを購入している」「顧客Bはサバ缶、七味唐辛子、ビールを購入しているが、カップヌードルは購入していない」といった傾向を把握した際、顧客Bを潜在顧客と捉え、営業担当者が施策を検討するといった具合だ。

生成AIがデータ活用を深化させる

日清食品では、Azure OpenAI Serviceを用いて独自に構築した生成AIチャットツール「NISSIN AI-chat」をグループ全社4800人で活用している。2023年からは、生成AIをデータ活用の取り組みに加えて、大規模データベースからAIが自動でレポーティングする仕組みの構築を進めている（図3）。

データ活用にAIを適用していく中で課題も見えつつある。小郷氏は1つの例を挙げた。生成AIに次のような指示を与える。「あなたは高度な統計分析官です。オフィスAの2022年11月から2023年3月まで出荷実績を基にさまざまな角度から分析し、トレンド、ピーク、季節性、周期性、前年同月比較などの観点でインサイトを提供してください」。返ってきたのは「オフィスAカップヌードル出荷実績分析レポート」というタイトルの付いた、指示した5つの観点ごとに言及したレポートだ。小郷氏によると、「一見しっかりとしたレポートだが、読むと不正確な部分が混じっている」という。データサイエンス室で検証したところ、意味の読み取りや要約などに関しては問題ないが、データの比較、類推に関してはハルシネーションが現れており、改良の余地があるとした。

ハルシネーションを生む部分に対しては、生成AIが参照するデータを適切な範囲に限定し、分析の補強となる情報が必要であれば追加するといった工夫で回答精度を高めているという。

小郷氏は、「技術検証を踏まえたデータ基盤と生成AIの連携により、現場のスタッフに効果のインパクトをもたらしていきたい」と述べ、3ステップの推進計画を紹介した。現在は販売情報を基にデータの概観を言語化するステップにあり、その後、プロンプト投入に関連データを加えて分析レポートを生成するステップを経て、最終的にはAIが要因を探索し、担当者にアクションを提案するステップを目指している（図4）。

「特異なデータが発生した際、生成AIがそれに関してレポートを自動生成したり、担当者に対して情報をリマインドしたりといったことも将来的には目指していきたい」（小郷氏）

“デジタル武装”に生成AIが加わって、日清食品はデータドリブン経営に向けた取り組みに磨きをかけようとしている。全社データ基盤を中核に、データ分析の高度化や共通マスターの整備、データリテラシー教育などを並行して進めていくという。

先行ユーザーの取り組みに学ぶMicrosoft Copilot/Azure OpenAIの効果的活用法

日本マイクロソフトは2024年3月18日、同社の生成AIサービス群に関する説明会を開き、ユーザー3社がみずからの取り組みを紹介した。日本ビジネスシステムズ（JBS）が「Copilot for Microsoft 365」を、ベネッセホールディングスが「Copilot Studio」を、ソフトバンクが「Azure OpenAI Service」をそれぞれ用いて、生成AIによる業務効率化・自動化を図っている。

日川佳三（IT Leaders編集部）

日本マイクロソフトが開催した生成AIサービス群に関する説明会に、同社のユーザー企業である日本ビジネスシステムズ（JBS）、ベネッセホールディングス、ソフトバンクの3社が登壇。プロジェクトを主導したキーパーソンがそれぞれの取り組みを紹介した。

Copilotで議事録作成や契約書チェックを省力化 ―JBS

JBSは2024年3月、全従業員2500人を対象に、Microsoft 365のAIアシスタント「Copilot for Microsoft 365」を導入した。Microsoft 365の利用時、対話型で問い合わせて回答を得られる機能である。

日本マイクロソフトによる同機能の一般提供開始は2023年11月。JBSは一般提供に先立って、同年8月に初期導入プログラムに参加し、希望者300人にライセンスを付与してトライアル導入を行っている。

「Teamsにコミュニティを作ったり、勉強会を開催したりして、Copilotの活用を促した。12月には利用状況を分析し、さらなる活用に向けた体制づくりに取り組んだ」（JBSの後藤行正氏、写真1）という。

後藤氏らがトライアル導入の前後で業務がどう変わったかを調べたところ、社員1人あたり1カ月あたりの価値創造時間が、従来の40時間から54時間へと36%増えたという（図1）。

後藤氏は、Copilot for Microsoft 365が効果を上げた例として、マーケティング部門における議事録作成を挙げた。「Teamsは標準で会議要約機能が備わるが、プロンプトを工夫することで、他の資料で活用しやすい議事録を作成できている」（同氏、画面1）。また、法務部門では契約書チェックに活用し、これまで1件に平均15分かかっていたのが平均5分に短縮されたという。

写真1　日本ビジネスシステムズ　取締役 専務執行役員 ビジネスグループ統括 デジタルセールス本部 担当の後藤行正氏

カスタムCopilotでイントラネットのナレッジを活用 ―ベネッセ

ベネッセは、Copilot for Microsoft 365の利用に加えて、「Copilot Studio」を使って、イントラネットに蓄積したナレッジを基に従業員からの質問に回答するカスタム型のAIアシスタント「社内相談AI」をノーコードで開発した。情報の検索や各部門/担当者に相談する時間を減らすことを目指している（図2）。

2023年10月にトライアル版をリリースしてPoC（概念検証）を開始。イントラネットから750ページ相当分の情報、各部門の業務マニュアルをナレッジにしている。2024年2月に正式版をリリー

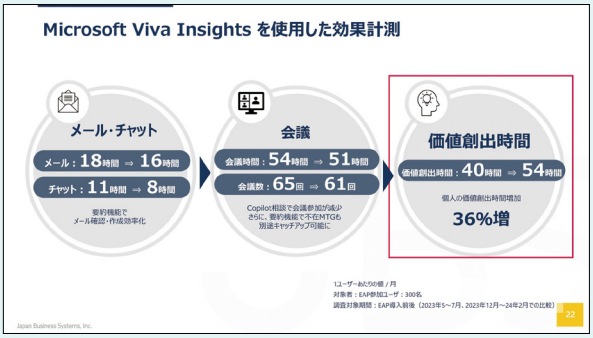


図1　Copilot for Microsoft 365 先行導入の効果（出典：日本ビジネスシステムズ）

画面1　Copilotを使った議事録生成画面（出典：日本ビジネスシステムズ）

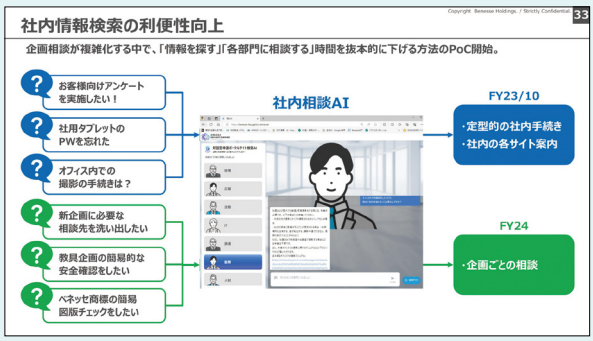


図２　Copilot Studio を使って開発した、イントラネットの知識をもとに回答するカスタム型 AI アシスタント「社内相談 AI」の概要（出典：ベネッセホールディングス）

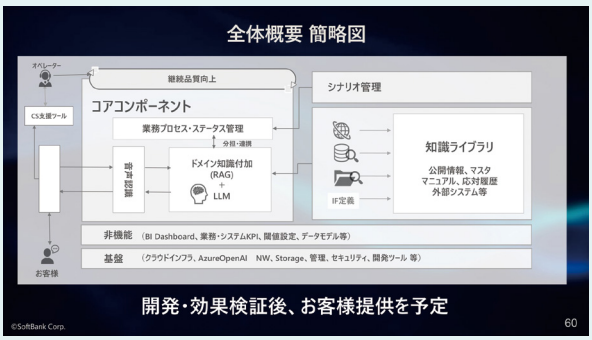


図３　改善したデータセットで向上した正答率（出典：ベネッセホールディングス）

スし、同年3月中旬には、ナレッジに社内データを追加する。

「PoCにおいては、データセットを作ること自体が大事ということがわかった」（ベネッセホールディングスの橋本英知氏、写真2）。初期状態の正答率は54%と低かったが、データセットを改善することで正答率が向上したという（図3）。

“自律思考型LLM”でコールセンター業務を改善 ―ソフトバンク

ソフトバンクは、生成AIサービス群「Azure OpenAI Service」を用いて、コールセンター支援アプリケーションを構築している。顧客からの問い合わせ内容を大規模言語モデル（LLM）が判断し、自動で案内したり、データソースから情報を収集して回答したりする。2024年7月以降、自社コールセンターに順次導入し、業務の自動化を拡大していく（図4）。

写真２　ベネッセホールディングス 専務執行役員 CDXO 兼 Digital Innovation Partners 本部長の橋本英知氏

これまでは、顧客からの個々の問い合わせに対しフロー追従型で対応していた。「LLMが問い合わせのインテント（検索意図）を分類し、決められた順序と固定化されたスクリプトで対応していた。IVR（自動音声応答システム）に近いことしかできなかった」（ソフトバンクの牧園啓市氏、写真3）という。

そこで、フロー追従型に代えて、LLM自律思考型のシステムの開発に取り組んでいる。同システムでは、顧客との会話内容に応じて必要な機能やデータソースを利用。さらに、対応精度を高度化するためにプロンプトを工夫するほか、「Azure AI Search」を活用してRAG（Retrieval-Augmented Generation：検索拡張生成）構成を採り、社内のデータベースをナレッジに回答を生成する（図5）。

写真３　ソフトバンク IT 統括 専務執行役員兼 CIO の牧園啓市氏

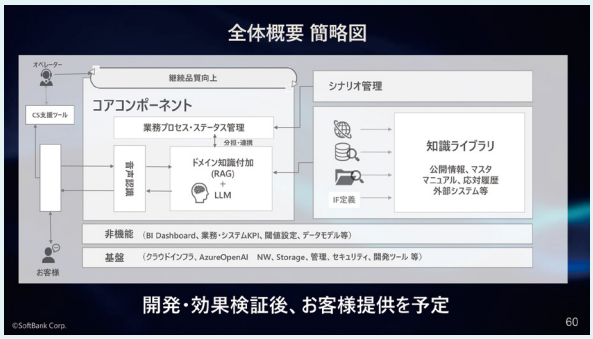


図４　Azure OpenAI Service を使って構築するコールセンター支援アプリケーションの概要（出典：ソフトバンク）

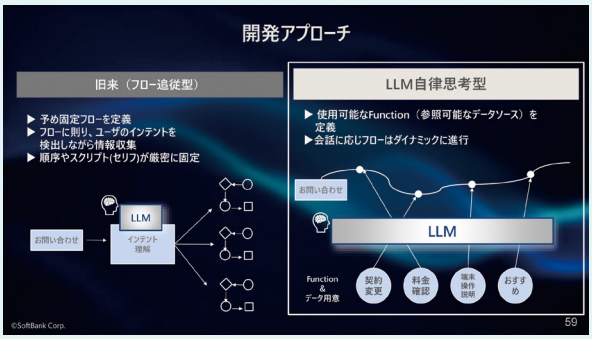


図５　フロー追従型からLLM自律思考型へと開発アプローチを改め、問い合わせに対して柔軟に対応できるようにする（出典:ソフトバンク）

グループ90社1000人で生成AIを活用 店舗運営、商品企画、システム開発など

イオン（本社：千葉県千葉市）は、グループ90社の約1000人で生成AIの利用を開始した。Exa Enterprise AIの生成AIサービス「exaBase 生成AI」を導入して、店舗運営や商品企画、IT開発のコード生成などの用途で活用している。生成AIの情報交換の場として掲示板を設け、プロンプトの共有などを行うほか、レベル別（初級・中級・上級）の勉強会を定期的に開催している。Exa Enterprise AIが2024年2月13日に発表した。

日川佳三（IT Leaders編集部）

イオンは、グループ90社の約1000人で生成AIの利用を開始した。店舗で使う文書の作成、商品企画・アイデア立案、IT開発のコード生成などの用途で活用している。総合スーパー、ディスカウントストア、専門店、ヘルスケア＆ウェルネス、金融、機能などグループの全業態で取り組んでいる（図1）。

利用を促進する取り組みとして、社内ポータルサイト内に情報交換掲示板を設置して、生成AIの技術動向、便利なプロンプト、失敗事例などを情報交換している。実際に、「掲示板で知ったプロンプトが業務に役立った」という報告が複数寄せられているという。

また、利用者のAIリテラシー向上を目指して、レベル別（初級・中級・上級）の勉強会を定期的に開催している。技術動向やグループ内外の事例を共有するほか、実際に生成AIを使って学ぶ機会を設けている。今後、ハッカソンなどのプログラムを通じて新規ビジネスの創出などにつなげていくとしている。

エクサウィザーズ子会社のExa Enterprise AIが提供する「exaBase 生成AI」を利用している。Azure OpenAI ServiceのChatGPTを用いた法人向けの生成AIサービスである（画面1、図2）。

イオンでは、社内に蓄積したナレッジを利用して対話・生成を行う手法も視野に入れている。現在、exaBase 生成AIの「データ連機機能」を一部の利用者に絞って提供している。今後、用途や回答の質などの課題を洗い出し、改善を図りつつ進めていく。

画面1　イオングループで導入した生成AIポータル画面（出典：Exa Enterprise AI）

図1　イオングループにおける生成AIの利用状況（出典：Exa Enterprise AI）

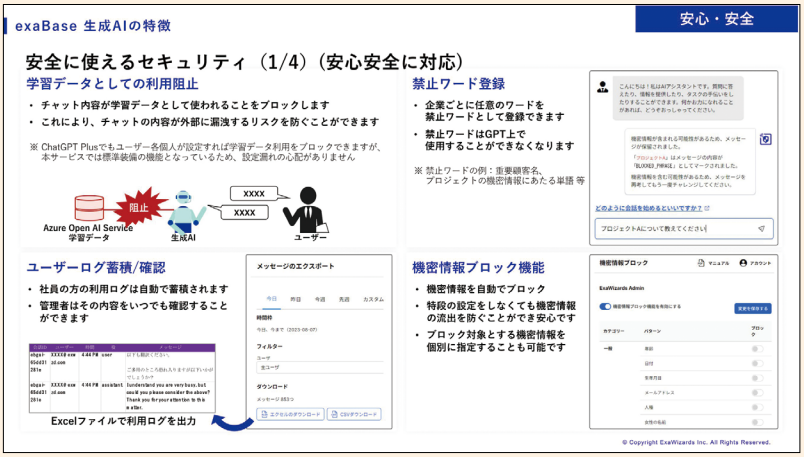
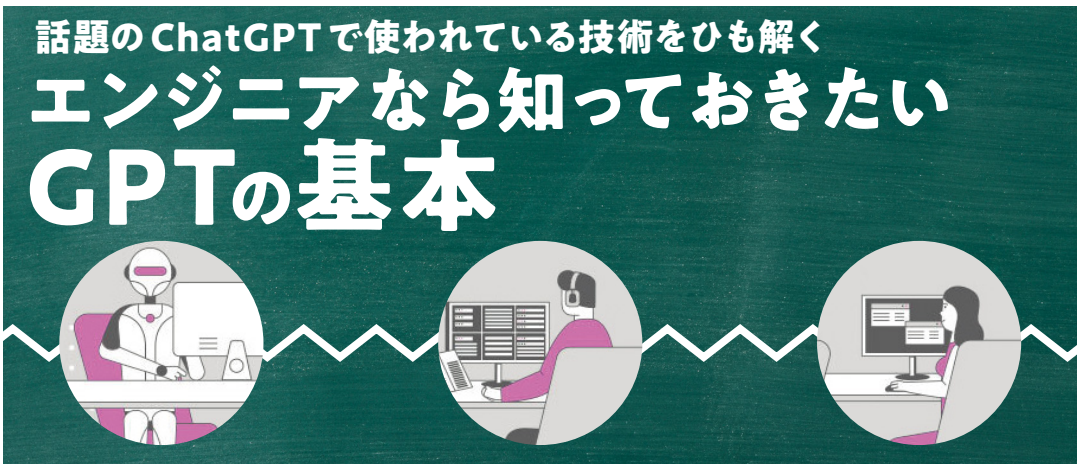


図2　exaBase 生成AIの特徴（出典：Exa Enterprise AI）

https://it.impress.co.jp/articles/-/25957

本記事は抜粋版です。ぜひ、Webで全文をお読みください！



● 株式会社システムインテグレータ　梅田 弘之(うめだ ひろゆき)

[第1回] GPTで始まる大規模言語モデル時代

第1回は、GPTの概要を中心に解説します。

はじめに

AIは既に顔認証や音声認識、翻訳などさまざまな分野で実用化されていますが、ChatGPTの出現により自然言語処理能力についても十分実用レベルに到達していることが広く認識されました。それどころか、あまりにも急速に賢くなっていて、いったいどこまで行くのだろうと不安視する声も聞こえてきます。

何ごとも相手をよく知らないと不安になるものです。そこで、本連載ではGPTシリーズを中心に大規模言語モデル(LLM)がどのような技術や原理で人間の期待した回答を生み出しているのかをやさしく解説します。仕組みを知ると客観的に判断でき、ビジネスへの活用イメージが湧きやすくなります。

ChatGPTとは

「ChatGPTとは、OpenAIが開発した大規模な言語モデルで、テキスト生成や言語翻訳などの自然言語処理タスクに利用することができます。現在利用可能な最大かつ最先端の言語モデルの一つであるGPT-3(Generative Pretrained Transformer 3)モデルをベースにしています。ChatGPTは、人工知能チャットボットです。2022年11月に公開されました。」

はい、まずは“お約束”で、これはGPT-4を搭載しているBingに「ChatGPTとは」と質問した際の回答です。

Bingの回答をまとめると、次の３点になります。

- OpenAIが開発した大規模言語モデルで、2022年11月に公開された
- テキスト生成や言語翻訳などの自然言語処理タスクに利用できるAIチャットボットである
- GPT-3(Generative Pretrained Transformer)モデルをベースにしている

もう少し補足しましょう。

- GPTは、Generative Pretrained Transformerの略。Generative(生成できる)、Pre-trained(事前学習する)、Transformerという技術を使った言語モデルで、生成AIと呼ばれている
- ポリグロッド(多言語対応している)言語モデルである
- Attentionという技術で学習し、RLHFという強化学習でお作法を学んでいる
- 正確にはGPT-3.5をベースとしており、その進化版GPT-4もリリースされている
- ChatGPTやGPT-4は2021年９月までのデータで学習しているため、それ以降の情報に弱い
- Microsoftの新BingやCopilotでGPT-4が使われている

大まかな特徴はこんなところでしょうか。本連載は「エンジニアなら」という冠がついていますので、GPTとのやり取りの紹介は少なめにして、仕組みや技術について解説していきます。

GPT誕生まで

実は、私は2017年のThink ITの連載「ビジネスに活用するためのAIを学ぶ」で「自然言語処理は、音声認識や画像認識に比べると“人間レベル”に到達するまでまだ時間がかかりそうですが…」と書きました。実際、そのときのレベルはそんなものだったのですが、直後にTransformerという新技術が現れて急に進化が加速しました。

４年後の続編「エンジニアなら知っておきたいAIのキホン2021年版」の第６回(2021年10月26日掲載)では、GPT-3を紹介しています。そこでは「これがもっと進化して完成度が高くなれば、骨子やあらすじを示すだけでブログや記事、小説などをAIが書いてくれる時代が来る、そんなふうに期待されているのです」と書いています。そして、そのわずか１年後の2022年11月30日にChatGPTが公開され、これが現実的なものとして認識されたのです。

実際、ChatGPTのような大規模言語モデルは短い期間で急成長しており、その進化の速さが「このままだととんでもないことが起こるのでは」という不安を掻き立てている面もあります。そこでエンジニアらしく技術を理解して冷静に判断するために、まずはChatGPTがどのように作られてきたのか、その誕生までの流れを図１を使って説明しましょう。

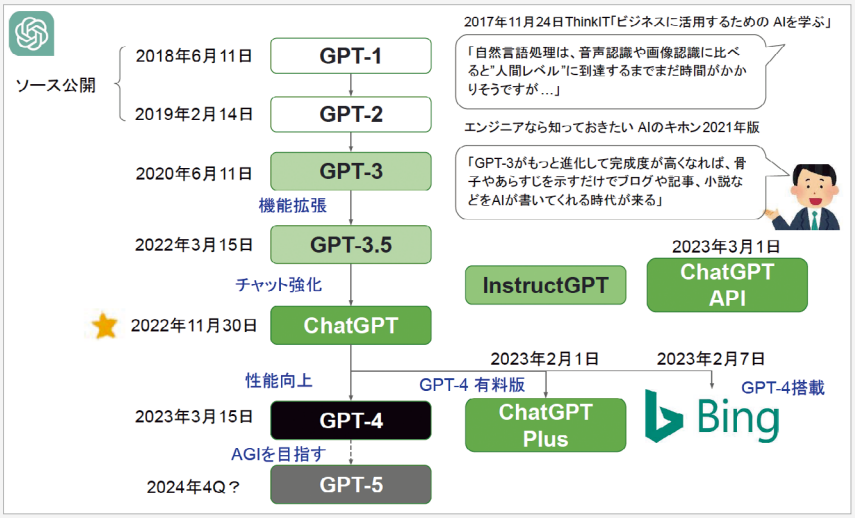


図1：GPTシリーズ

GPT-3

アメリカのAI関連企業OpenAI社は、2020年6月にGPT-3というAIを公開しました。これは、従来の自然言語処理AIに比べて格段にレベルが高く、AIが人間のように書けることを最初に示したと言われています。GPT-3はその前身のGPT-2と同じ言語モデル構造ですが、学習データ量が40GBから570GB、パラメータ数が15億個から1750億個と大幅に増えています。このGPR-3の登場により学習データとパラメータ数を大きくしてゆく大規模言語モデル競争が始まったのです。

GPT-3.5

OpenAIは2022年5月にGPT-3の機能を拡張し、2021年6月までのデータを用いて訓練したGPT-3.5というモデルをリリースしています。パラメータ数は未公開ですが推定3550億個くらいとも言われており、これがChatGPTのベースとなっています。

ChatGPT

GPT-3.5のチャット機能を強化したものが、話題沸騰のChatGPTです。チャット強化とは「人間の好む回答をする(話術の向上)」と「不適切な発言をしない(マナー向上)」という2点です。これをRLHFという強化学習により学び、一般公開しても大丈夫なレベルにしたのです。

この戦略は非常にうまくハマりました。GPT-3やGPT-3.5は多くの専門家に注目される技術でしたが、あくまでも"通を唸らせる"存在でした。しかし、これをチャットで公開して誰でも利用できるようにサービス提供したことで一気にバズりました。世界中でこれをどのように活用するかという試みが爆発的に広がり、学習データも世界中の人々から集まっています。これまでの研究室レベルから世界に広がったことにより、進化が加速しているのです。

なお、現状、1つだけ注意が必要なのがChatGPTは2021年9月までの学習データで学んでいることです。そのため、それ以降の出来事に関係する質問をした場合に、回答のネタが古いというケースが発生します。例えば、ChatGPTに「日本の総理大臣は誰ですか」と聞いてみると「2023年5月現在、私が取得している情報によれば、日本の総理大臣は菅 義偉(すが・よしひで)氏です。」と回答します。わざわざ"2023年5月現在"という言葉を付けて平気で嘘をつくのがChatGPTらしいですね。

ChatGPTシリーズの成長

GPT-4

2023年3月にGPT-4がリリースされました。これはChatGPTを進化させたもので、次のように性能が大幅に向上しています。

a. 言語能力が上がり(言い回しがうまい)や信頼性も向上(嘘が減る)している
b. 文字だけでなく画像も取り扱える(マルチモーダルである)
c. ChatGPTに比べて、長い文章が取り扱える(8倍の長さ)
d. 脚本や音楽などの創造性が向上している
e. ポリグロット(多言語を操れる人)の能力がアップ
f. 差別や暴力など不適切な発言を回避する能力が向上(プロンプト・インジェクションに強い)

OpenAIのテクニカルレポートによると、アメリカの司法試験(模擬試験)を受験したところ、ChatGPTが受験者の下位10%程度のスコアだったのに対し上位10%程度に入って合格したそうです。こうした専門的な分野(さらに英語圏)によっては人間レベルの性能を発揮できそうですね。

ChatGPT Plus

2023年2月1日に、GPT-4を搭載したChatGPT Plusがリリースされました。月20ドルという価格ですが、GPT-4を使えるほかに、有料な分、優先的なアクセスやサポートが提供されます。また、マイクロソフトのBingのチャットはGPT-4を搭載しているので無料で利用できます。

なお、GPT-4の学習済みデータはいまのところChatGPTと同じく2021年9月までのデータです。Bingの場合は、プロメテウスという仕組みでネット上の新しい情報を使って回答してくれるので、日本の総理大臣を尋ねると図2のようにきちんと岸田文雄と回答してくれます。

図2：ChatGPTとBingチャットに総理大臣が誰かを質問

GPT-5

GPTモデルの性能向上はこうしている間も絶え間なく続いており、2023年第4クォーターもしくは

2024年には次のモデルとなるGPT-5がリリースされると噂されています。また、GPT-5のリリース前に中間バージョンとなるGPT-4.5があり、これがリリースされるかどうかは分かりませんが、長文入力に対する正確な対応、より正確な回答などの性能向上が実現しているそうです。

AGI

ChatGPTやBingで公開した結果、これまでの研究室での開発に比べて、いっきに実用的な対話データを取得できるようになりました。これらを学習データとして利用することで著しい進化を果たすのではと期待しています。

OpenAIの目標は、第3次AIブーム(2018年頃)に話題となったAGI(Artificial General Intelligence)、すなわち人間と同じレベルの汎用人工知能です。そして、GPT-5はその可能性を示す最初のモデルになると期待されているのです。

OpenAIとマイクロソフト

OpenAI Inc.は、2015年12月にサム・アルトマン氏やイーロン・マスク氏らが10億ドル出資して作った非営利法人です。マスク氏は、テスラで研究しているAIとの利益相反を理由に2018年2月に役員を辞任して離れました。その1年後の2019年3月にいくつかのファンドから出資を受けて営利部門のOpenAI LPが設立され、7月にマイクロソフト社から10億円の出資を受けて関係性を深めています。

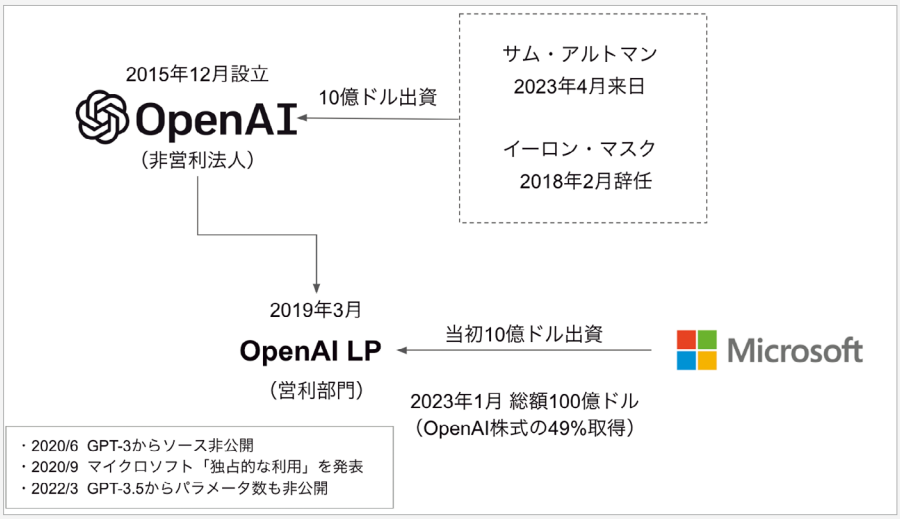


図3：OpenAIとマイクロソフト

マイクロソフトはさらに追加で出資を行い、2023年1月には総額100億ドルとなりOpenAIの株式の49%を取得しています。創業者でありCEOのサム・アルトマン氏は、2023年4月に来日して岸田総理と対談したことで、お茶の間でも有名になりましたね。

GPTシリーズはOpenAIの設立ポリシーのもとで2018年のGPT-1、2019年2月のGPT-2まではオープンソースとして公開されてきました。しかし、2019年3月に営利組織のOpenAI LPが設立されてマイクロソフトなどからの出資を受けたことでスタンスが変わり、2020年9月22日にマイクロソフトはGPT-3の「独占的な利用」を発表しました。

そして、2020年6月発表のGPT-3からはソース非公開になっており、ChatGPTではパラメータ数な

ども秘密になりました。AIという新たな武器で王座Googleに挑む立場としては当然のことだと思いますが、生い立ちがオープンだっただけに批判もあります。

大規模言語モデル

エピソード1と2

上記ではGPT-3から説明しましたが、スターウォーズと同じくエピソード1と2、すなわちGPT-1とGPT-2についても紹介しましょう。GPT-1のパラメータ数は1.17億個、GPT-2のパラメータ数は15億個でした。ただし、OpenAIは「悪意のある応用に対する懸念」を理由にGPT-2フルモデル版のリリースを見送り、1.24億パラメータの縮小版を2019年2月14日にリリースしています。

悪意ある応用とは、オンラインでのなりすましや不適切なコンテンツ、誤解を与える記事、ヘイト宣伝などです。これ以降も継続的に対策を続けていますが、いまだに問題視されることが多い宿命のような課題とも言えます。

スケーリングの法則(Scalling Low)

OpenAIは2020年に「スケーリングの法則(Scalling Low)」という論文で、ニューラルネットワークの言語モデルの性能が計算能力やモデルサイズ(パラメータ数)、データセットの量と関係が深いことを発表しました。つまり計算能力はもちろん、パラメータ数や学習データの量は非常に重要で、これが大きければ大きいほど精度が上がるということです(図3)。

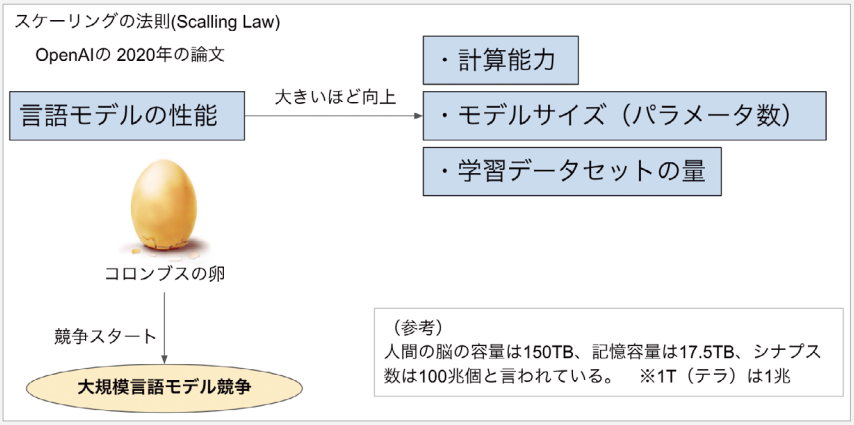


図4：スケーリングの法則

シンプルで当たり前に見えますが、まさにコロンブスの卵のような法則です。通常は性能がサチる(飽和する)と思われそうですが、サチらず強い相関関係が続くのです。この法則のモデルはGPT-3でした。モデルサイズ(パラメータ数)はGPT-2が15億個だったのに対し、GPT-3は1750億個と大幅に増えています。また、学習データ量もGPT-2が40GBだったのに対しGPT-3は570GBに拡大しています。このように言語モデルを大規模化したことにより、初めて「AIが人間に近い対話をできそうだ」と思わせる実力が示されたのです。

大規模言語モデル競争

論文とGPT-3の出現により、一気に大規模言語モデル(LLM：Large Language Model)競争が始

まりました。現在、世界中で表のような言語モデルの大規模化が著しく進んでいるのです。

言語モデル	リリース日	開発元	最大パラメータ数
GPT-3	2020年6月	OpenAI	1750億
GShard	2020年6月	Google	6000億
Swich Transformer	2021年1月	Google Brain	1.57兆
悟道(WuDao)2.0	2021年6月	北京智源人工知能研究院	1.75兆
HyperCLOVA	2021年11月	LINEとNAVER	390億
Gopher	2022年1月	DeepMind	2800億
日本語GPT	2022年1月	rinna	13億
GPT-3.5	2022年3月	OpenAI	(推定)3550億
PaLM	2022年4月	Google Reserch	5400億
GPT-4	2023年3月	OpenAI	(推定)5000億～1兆

表：大規模言語モデル(GPT-3.5/4のパラメータ数は推定)

Google社は、GPT-3と同時期の2020年6月に6000億のパラメータからなる「GShard」を発表しました。その翌年の2021年1月にはGoogle Brain社が1.6兆ものパラメータを持つ「Swich Transformer」をオープンソース化し、半年後の2021年6月には北京智源人工知能研究院が悟道(WuDao)2.0を発表しています。また、2022年1月にはAlphaGoを開発したGoogleの子会社であるDeepMindが2800億のパラメータを持つGopherを発表しています。

GPT-4のパラメータ数は非公開ですが、5000億～1兆個くらいではと言われています。人間の脳の容量は150TB、記憶容量は17.5TB、シナプス数は100兆個と言われています。1T(テラ)は1兆ですから、だんだん人間の領域になっているような感じですね。

日本語に特化した言語モデルも登場して来ました。2021年11月にLINE社と親会社の韓国NAVER社が共同開発したHyperCLOVAのパラメータ数は390億個ですが、820億個のモデルを開発中とのことです。また、チャットボットりんなを開発していたチームがスピンアウトして2020年6月に設立したrinna社は、オープンソースのGPT-2をベースにした13億パラメータからなる日本語GPTをオープンソース化し、研究や開発に役立つように提供しています。

まとめ

今回は、以下の内容について学習しました。

- ChatGPTはGPT-3.5をベースとし、チャットを強化した大規模言語モデルである
- ChatGPTはAttentionという技術で学習し、RLHFという強化学習でお作法を学んでいる
- GPT-4は言語処理能力や信頼性、ポリグロットなどが大幅に向上し画像も扱える
- ChatGPTやGPT-4の学習データは2021年9月までのものである
- マイクロソフトはOpenAIに出資し、新BingやCopilotにGPT-4を搭載している
- スケーリングの法則により、大規模言語モデル競争が始まっている

今回は第1回ということで概要を中心に紹介しましたが、次回からはAttentionやRLHFなどの言語モデルの技術を解説していきます。難しい式を使わず、図や例でわかりやすく説明しますので、新しい技術に対する好奇心を一緒に楽しんでいきましょう！

[第2回] 大規模言語モデルの概要

第2回は、「ChatGPT」が誕生した背景を中心に、大規模言語モデルの本質について見て行きます。

はじめに

前回はスケーリングの法則により、大規模言語モデル(LLC)の競争時代が始まったことを解説しました。今回は、大規模言語モデルがどのように生まれたかを解説します。我々はChatGPTが人間のように会話したり、とても博学なのに驚かされましたが、いったいどうやってそんなことができたのでしょうか。

人の一生とAIの短期トレーニング

例えば人間であれば、いろいろな人と話したり、たくさんの文章を読み書きした結果で知識や会話能力、文章力が増えて、そのノウハウがその人の脳に刻まれています。実は、言語モデルAIの場合も似たようなアプローチです。膨大な文章を学習データとして読み書きのトレーニングをし、そこで得たナレッジをパラメータ化しています。

人間が一生かけて読み書きするテキスト量は馬鹿になりませんが、AIはそれをはるかに超える量のテキストを短期間で学習し、それをニューラルネットワークという脳に刻んでいるので、とんでもなく博識なのです。

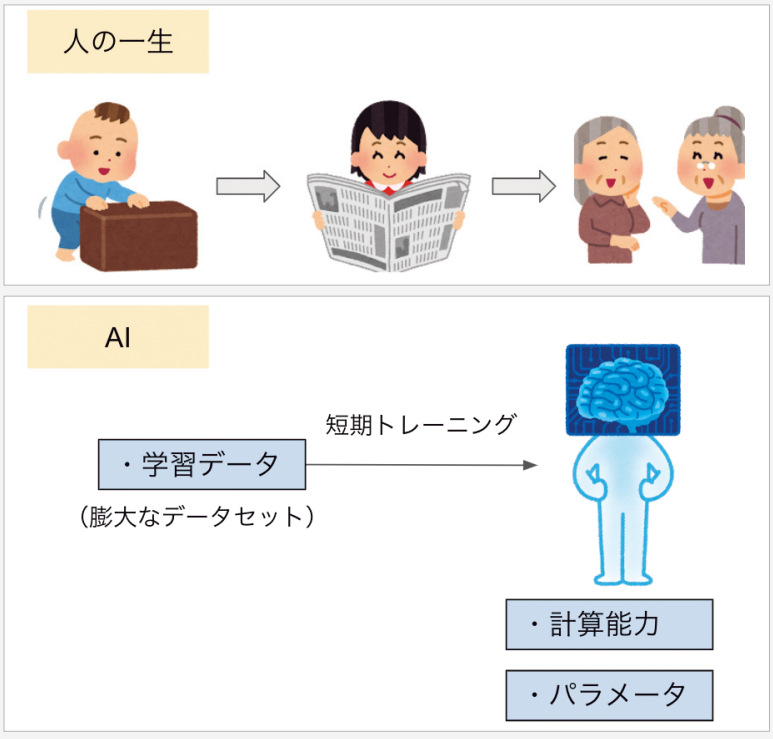


図1：人の一生とAIの短期トレーニング

そう考えると、スケーリングの法則が、言語モデルの性能は「計算能力」に加えて「モデルサイズ(パラメータ)」と「学習データセットの量」に比例すると言っている意味が理解できます。今や、パラメータは数千億から兆単位に拡大し、チャットを通じて世界中の人々から生きたデータを(個人情報を除いて)集めているので学習データも莫大に増えています。そして、大規模になればなるほど、それを高速処理する計算能力が必要となり使用されるGPUの数も多くなっているのです。

GPT-3の学習データ

GPT-3.5は2021年3月までのデータ、ChatGPTとGPT-4は2021年9月までのデータで学習しています。これらのモデルがどの学習データを利用しているかは非公開ですが、GPT-3の学習データをベースにしているので、GPT-3がどんな学習データでトレーニングしたのか見てみましょう。

GPT-2の学習データは40GBだったのに対し、GPT-3は570GBにボリュームを拡大しています。図2はその内訳です。約6割の寄与度を締めているのがCommon Crawlコーパスです。これはインターネットのWebサイトをクロール(巡航)し、そこに掲載されているテキストデータをスクレイピング(データ取得)してコーパスにしたものです。

コーパスとは、文章や会話などを大量に集めて、コンピュータで処理しやすいように構造化した言語データベースです。Common Crawlのホームページを見ると「12年間のクロールで収集されたペタバイト単位のデータ」と記載されています。2008年からクロールとスクレイピング技術で集めたデータは、メタデータ(WAT)も付けられてAmazon S3に保存されており、誰でも無料でアクセスできます。このうち各言語の含有率は英語が50%程度であり、日本語は5%となっています。

WebText2は、Common Crawlコーパス以外のWebページのデータ、Book1とBook2は書籍データですが、どのような情報が使われたかについては明確にされていません。

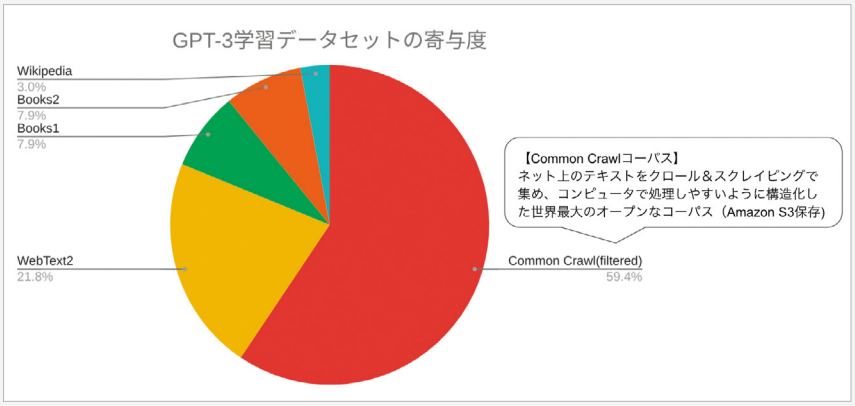


図2:GPT-3学習データセットの寄与度

日本語コーパス

Common Crawlのスナップショット(ある時点のデータを抜いたもの)で、フリーの日本語コーパスが公開されています。MC4はGoogleの多言語コーパスで、そこから日本語部分を抜き出したものです。

OSCARはフランス国立情報学自動制御研究所(INRIA)のOrtizチームが各国語に分けて提供しているもの、そしてCC-100はFacebookのコーパスで、そこから日本語を抜き出したものです。

また、国立国語研究所がさまざまな書籍や雑誌、新聞、白書、ブログなどからデータを集めたBCCWJという日本語コーパスや音声データのコーパスCSJなどもあります。大規模言語モデルの出現により学習データの重要性が再認識されたことで、日本語のコーパスのさらなる充実が期待されます。

表：主な日本語コーパス

データセット	サイズ	提供元
mC4	830GB	Google製データセットの日本語部分
OSCAR	260GB	INRIAが各国語のコーパスを提供
CC-100	82GB	Facebook製データセットの日本語部分
BCCWJ： 現代日本語書き言葉均衡コーパス	約1億語	国立国語研究所
CSJ： 日本語話し言葉コーパス	約700万語 (音声データ)	国立国語研究所

表：主な日本語コーパス)

言語モデルの本質

ChatGPTのTransformer技術を説明する前に、言語モデルの本質について理解しておきましょう。AIには、分類(Classify)や予測(Prediction)、異常検知(Anomaly detection)などさまざまな技術分野があります。その中で、言語モデルは自然言語処理(NLP:Natural Language Processing)を中心としたAIです。

言語モデルの機械学習

言語モデルの本質は「次の単語を予測するAI」で、そのトレーニングは学習フェーズと推論フェーズからなる機械学習で行います。図3のように事前学習(Pre-training)で徹底的に人間の文章を叩き込み、学習済みモデル(Pre-trained Transformer)を本番で使って「推論」させます。

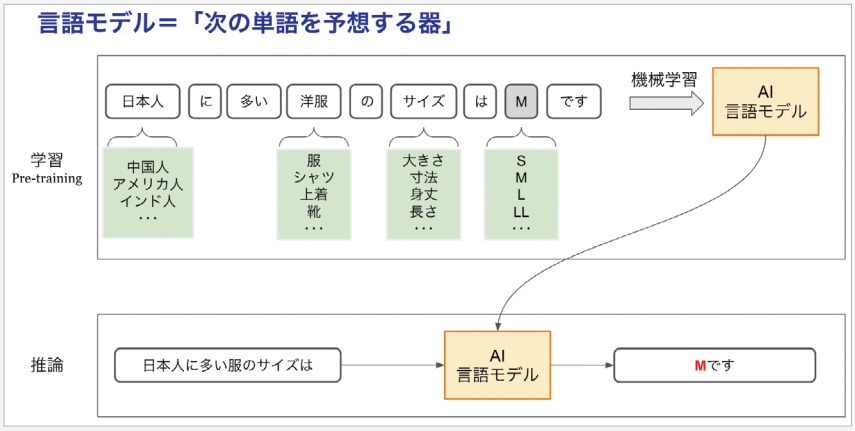


図3：言語モデルの機械学習

ここでは「日本人に多い洋服のサイズはMです」という文章を学習データとしてインプットしています。学習データには膨大な量のテキストが用意されており、「日本人」「洋服」「サイズ」「M」などのところに別の単語が入った文章データもあるわけです。これらを徹底的に学習することで「日本人に多い洋服のサイズは?」と聞かれた場合に、Mと回答する「次の単語予想」に強い言語モデルができるわけです。

この例はシンプルですが、実際は助詞の「の」にしても「は」や「も」や「に」などもさまざまなバリエーションがあり、それにより回答も変化します。ものすごい数の組み合わせがあるため、到底不可能に見えます。でも、人間だって長い年月をかけてこのような学習をし続けて、普通に会話できるレベルに達しているわけです。AIの場合は、24時間365日すごい処理速度のマルチGPUで大量データで勉強を行っているわけで、そう考えれば不可能ではないと理解できるのです。

LLMは連想ゲームの達人

我々はChatGPTの出現によりLLMの言語能力に驚いているわけですが、彼らは考えて答えているわけではありません。それよりも連想ゲームの特訓を死ぬほど行った次の単語予測器と考えた方が良いでしょう。

「日本人は」でピンと来たいくつかの続く単語を確率付きで出力し、「日本人は、夏は」で、また次に来る文章を予測する。そんなふうにインプットされた文章で反射的に浮かんだ単語を回答する連想ゲームの達人なのです。つまり、言語モデルが「次の単語を予測する器」だとしたら、大規模言語モデルは「連想ゲームの達人」なのです(図4)。

図4:大規模言語モデルは連想ゲームの達人

まあ、突き詰めて考えたときに、人間だって知識の中から考えて答えているのと、ぱっと浮かんだ単語を反射的に答えているのと、どこに境界があるのかは微妙です。しかし、大規模言語モデルが考えて答えているわけではないことを知れば、とりあえずは過度に脅威を感じる必要はなさそうです。と言いつつ、最近大規模言語モデルの「思考の連鎖」の研究が著しいので、これから考える力も身につけてきそうな不気味さも感じます。

大規模言語モデルは言語の天才

ところで、言語モデルをトレーニングして次の単語を推測できるようにしたものが、なぜ、プログラミングや作曲などもできるようになったのでしょう。図5を使って説明しましょう。

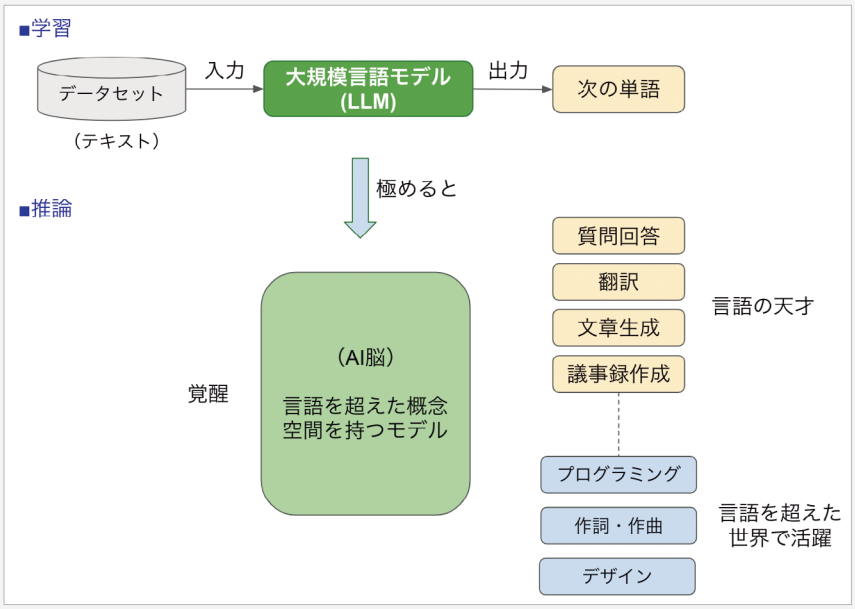


図5：大規模言語モデルの覚醒

大量のデータセットを使って、次の単語を連想するゲームを学習しまくった言語モデルは徐々に覚醒していきます。そして「質問回答」だけでなく「翻訳」「文章生成」「議事録作成」など、言語をいろいろ操れる天才の素質を見せてくれます。

このとき、言語モデルの頭の中には「言語を超えた概念空間を持つモデル」ができてきます。人間が自分の脳の仕組みをきちんと説明できないように、このAI頭脳がどのようになっているかは誰も説明できません。

翻訳の場合は、日本語の文章を入力して英語を出力させるトレーニングを行います。これも基本は「次の単語を予想する言語モデル」で、入力された日本語をAI能の空間でパターン化し、それを英語で出力するという形になります。例えば、和英翻訳を学習したLLMに「私はふるさとが好き」とインプットすると、AI脳がどういうロジックを使っているかはわかりませんが、「I love my hometown」とアウトプットするのです。

面白いのは、日本語や英語は単なる入出力のフォーマットにしか過ぎないということです。例えば、中国語と英語の翻訳を学習した言語モデルに「私はふるさとが好き」とインプットすると（たとえ日本語と中国語の翻訳を学習してなくても）「我愛我老家」などと翻訳してくれます。

つまり、日本語と英語というような2つの言語のペアで覚えているわけではなく、ある言語でインプットされた情報がいったん「AI脳」に変換され、それを別の言語にアウトプットしているだけなのです。GPT-4がポリグロット（多言語を話せる）なのは、この仕組みによります。

ChatGPTやGPT-4の学習データは圧倒的に英語が多く、日本語データはそれほど多くないのです。そのため英語で使用したほうが性能は良いわけですが、その割には日本語でもびっくりするほど優秀です。

これも、英語圏の知識は英語データの方がもちろん豊富ですが、知識以外の部分はAI脳が処理しているからだと推定されます。

大規模言語モデルは言語を超えた叡智

大規模言語モデルの「次の単語予想」トレーニングを極めてゆくと、いつのまにか「質問回答」や「翻訳」などの言語処理だけでなく「プログラミング」「作詞・作曲」「デザイン」など多彩な採用を見せるようになります。まるで、子どもが急に、やらせてみたらなんでもできるように覚醒したかのようです。

これも原理はAI脳です。プログラミングも言語であり、音楽も譜面を読むという言葉があるように言語のようなものです。AI脳にとっては、日本語も英語もプログラムコードも楽譜もすべて言語のようなものなので、世界中のデータを教師データに使って、Attentionという技術で連想ゲームを教えさえすれば、ジャンルを問わず人間のようなアウトプットを作り出す才能を発揮できるわけです。

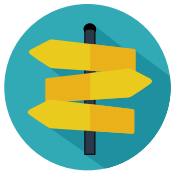
まとめ

今回は、以下の内容について学習しました。

- 大規模言語モデルが学習したテキストの量は、人間が一生で読む量より桁外れに多い
- GPT-3の学習データの6割は、ネットから取得したCommon Crawlコーパスである
- 言語モデルは「次の単語を予測する器」で、大規模言語モデルは「連想ゲームの達人」
- AI脳が「言語を超えた概念空間」を持つからこそ、連想だけでマルチになんでも行える

今回は大規模言語モデルの本質を中心に説明しました。次回からは、Attentionを中心としたTransformerモデルの説明や解説していきます。難しい式は使わず、図や例でわかりやすく説明するので、新しい技術に対する好奇心を一緒に楽しんでいきましょう！

Think ITサイトへ掲載されている記事は、さらに[第3回]～[第13回]へと続きます。
ぜひこの続きはサイトでお楽しみください。※2024年5月現在
https://thinkit.co.jp/article/22231 [第3回]
https://thinkit.co.jp/series/10884 (記事一覧)



IT Leaders × シンクイット Think IT™
特別編集号
2024 Summer
●AD／表紙デザイン：岡田章志 ●デザイン／制作：坂本房子／岡田デザイン事務所 ●編集：株式会社インプレス IT Leaders編集部／Think IT編集部

impress

株式会社インプレス

法人営業局
E-mail: customer@impressbm.co.jp
https:// www.impress.co.jp/
〒101-0051　東京都千代田区神田神保町1-105　神保町三井ビルディング