URL https://journal.ntt.co.jp/article/25744
DOI https://doi.org/10.60249/24042001

期待高まる国産生成AI（前編）
――AIの歴史的変遷と大規模言語モデルの動向

2022年11月に登場した「ChatGPT」に代表される生成AI（人工知能）は世界を熱狂させ，ビジネスや生活に変革をもたらしています．その要素技術の1つが大規模言語モデル（LLM：Large Language Model）で，米国のビッグテックをはじめ，各国で研究，開発が進んでいます．特に昨年来，規模の大小に加え，特定言語に対応したタイプや，金融，医療といった特定領域に適したタイプなど多様化と細分化が加速しています．本稿では前後編2回にわたり，LLMを中心にAIの歴史や現行のさまざまなモデル，国内外の法制度，開発・規制の動向を紹介していきます．

AIの誕生，そして現在

■繰り返すブームと冬の時代

AI（人工知能）という言葉は半世紀以上前，1956年に米ダートマス大学で開かれたワークショップ，通称「ダートマス会議」を機に誕生したといわれています[1]．AIの名付け親で後に米スタンフォード大学のAI研究所を立ち上げたジョン・マッカーシー氏が開催を呼び掛け，「AIの父」と呼ばれるマービン・ミンスキー氏らが参加したこの研究会で現代AI研究の基礎が築かれました．人間の思考や論理，学習の仕組みをコンピュータによる機械的操作，記号処理での解明，再現を試みます．

会議は1カ月に及びました．この研究会において，数学の定理をコンピュータで自動的に証明することに成功し，史上初のAIプログラムとして実を結びます．研究会では，「コンピュータへの言語のプログラム方法」や「神経細胞（ニューロン）網」などが課題に挙がりました（**表1**）．自然言語処理，ニューラルネットワーク，機械学習，抽象概念と推論，創造性といった今なお研究が続く今日的テーマの基礎が，このとき整理して定められたのです[2]．

出席者らはAIの実現を楽観視していました．1960年代初頭，マッカーシー氏は「完璧な知能を持つ機械」の10年以内の実現をめざし，スタンフォード人工知能研究所を設立しました．MIT人工知能研究所を設立したミンスキー氏も「一世代のうちにAIの実現に向けた問題点はほぼ解決されている」と予測していました．

しかし期待とは裏腹に，AIはさまざまな課題に直面，希望は失望へと変わっていきます．AIをめぐる期待と失望は，これまで幾度となく繰り返されてきました．

概説すれば，1950～1960年代，ダートマス会議の後にAI研究への楽観論が広まり，政府や企業はAI関連の投資を増やしました．この時期，基本的なAIプログラムが開発され，機械が簡単な問題を解決できることが示されました．ただ，過大な期待に反し，AIが直面する問題の複雑さと，計算資源の乏しさを背景に，研究は停滞，政府なども資金を引き揚げてしまいます．

停滞期を経て，1980年代を中心に，AIは再び第2次ブームを迎えます．医療分野などの専門家の知識を教え込んだ「エキスパートシステム」が，その牽引役でした．ただ，人間が情報を与え続ける必要があるといった制約が足枷となり，AIは高まる期待にこたえられませんでした．

ただ，その冬の時代を経て，2000年代以降にAIはみたび脚光を浴びます．総務省の2019年版情報通信白書の言を借りれば，「AIは期待と失望を繰り返しつつも関連の研究が進んでいた中で，近時目覚ましい研究成果を出すようになってきた」といえます[3]．

その第3次AIブームは萎むことなく，生成AIによるリブートを受け，現行の3.5次とも第4次ともいわれるかつてない盛り上がりをみせています（**表2**）．

■第4次ブームは生成AI

第2次AIブームの機運が萎え，冬の時代を迎えていた間にも，第3次ブームの萌芽となる技術は静かに胚胎していました．すなわち，1990年代から2000年代前半にかけて世界的に広まったインターネットと，その上で蓄積された多種多様かつ膨大な情報です．ネットとビッグデータが，AIを大きく進化させていく起爆剤となりました．

第2次ブームのエキスパートシステムの

表1　ダートマス会議で扱われたAIの課題

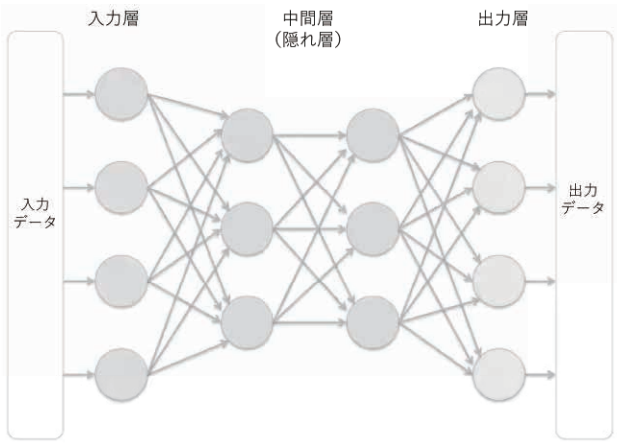
1. 自動計算
2. コンピュータへの言語のプログラム方法
3. 神経細胞（ニューロン）網
4. 計算量理論
5. 自己改善
6. 抽象化
7. 無作為性と創造性

情報通信総合研究所作成

表 2　AIの歩み

AI 草創期	1940年代	1945年	世界初のコンピュータ（ENIAC）登場
		1947年	チューリング、AIの概念を提唱
	1950年代	1950年	アイザック・アシモフが著書『われはロボット』でロボット３原則を提唱
第１次AIブーム		1956年	ダートマス会議でAIの言葉が登場
		1958年	フランク・ローゼンブラットが「パーセプトロン」を論文で発表
	1960年代	1964～1966年	ジョセフ・ワイゼンバウムが「ELIZA」を開発
		1965年	エドワード・ファイゲンバウムが「Dendral」を開発
	1970年代	1972年	スタンフォード大学で「Mycin」を開発
		1979年	米国人工知能学会（AAAI）が設立
第２次AIブーム	1980年代	1982年	日本で「第５世代コンピュータプロジェクト」が開始
		1986年	日本人工知能学会が設立
冬の時代	1990年代	1997年	「ディープブルー」がチェスの世界チャンピオン、ガルリ・カスパロフに勝利
		1999年	ソニーがペットロボット「AIBO」を発売
第３次AIブーム以降	2000年代	2005年	レイ・カーツワイルが著書『ポスト・ヒューマン誕生』にて、コンピュータが人類の知性を超えるときを予測
	2010年代	2011年	「Watson」が米国テレビクイズ番組「Jeopardy!」で人間のチャンピオンに勝利
		2012年	アップルがiPhone 4SにバーチャルアシスタントSiriを搭載 深層学習アプリケーションが画像認識コンテストで人間に圧勝 グーグルがAIによる猫認識の精度を公表
		2016年	「アルファ碁」が、囲碁の世界チャンピオンである李世乭九段に勝利
		2017年	グーグルの研究者らが深層学習モデル「Transformer」を発表
		2018年	OpenAIが大規模言語モデル「GPT」を開発
	2020年代	2022年	各社が相次いで画像生成AIをリリース
			OpenAIが対話型AIサービス「ChatGPT」を発表

各種資料より情報通信総合研究所作成



情報通信総合研究所作成

図１　ディープラーニングの構造

難点として，総務省の2016年版情報通信白書は「当時はコンピュータが必要な情報を自ら収集して蓄積することはできなかった」と指摘しています[4]．第３次ブームでは，その克服がカギになりました．

その期待を受けて登場したのが，1990年代以降に加速した機械学習，そして2000～2010年代に全盛期を迎えたディープラーニング（深層学習）です（**図１**）．ニューラルネットワークを発展させて入出力の間に隠れ層を組み込んで複層構造にした，このディープラーニングが３度目のブームのブレークスルーとなりました．

機械学習で賢くなったAIは，次々と人間を凌駕する能力を発揮していきます．IBMの特製コンピュータがチェスの世界王者を1997年に下し，世界を驚かせました．その後も，韓国のプロ棋士，李世乭（イ・セドル）九段を破ったGoogle傘下のディープマインド「アルファ碁」や，米国の人気クイズ番組「Jeopardy！（ジョパディ）」で優勝したIBMの「Watson（ワトソン）」など，AIによる面目躍如の快進撃が続きます．

その後もAIの進歩は目覚ましく，課題とされていた「必要な情報を自ら収集して蓄積する」ことができるようになりました．さらに，収集データを基に「生成」まで可能となった――．それがAIの現在地といえるでしょう．そして，生成コンテンツの種類はテキストに限らず，画像や音声，音楽など拡大を続け，１つのモデルで複数の種類に対応する「マルチモーダル化」が進んでいます．

■マルチモーダル化の追求

我が世の春を謳歌する生成AIですが，一口に「生成」といっても，LLMに基づくテキスト生成AIから，画像生成AI，音声・音楽生成AI，はたまたスライドや3DCGを生成できるモデルまで多岐にわたります．

今はPCやスマートフォンを通じてテキストで入力し，やはりテキストや画像で出力するといったタイプが主流です．一方，入力に音声や画像を組み合わせるモデルも出始めています．テキスト生成AIや画像生成AIのハイブリッド型など異種混合のケースが増えつつあります．

いずれは，そうした境目が消失し，あらゆる情報がデータ化されて入力要素となり，出力形態もテキスト，イラスト，図表，音声，動画など多様に選べる「マルチモーダル」型のモデルへと収れんしていくと見込まれます（**図２**）．

収れんするにしても，現行のテキスト生成AI，画像生成AI，音声生成AIが実現するまでに辿った道のり，それを支えてきた技術は異なり，それぞれ別個の構成要素のうえに発展してきた歴史があります．ただ，テキストも画像も音声も，それらを識別したり，分析したりする技術はここ10年ほど，特にディープラーニングの発達によって加速してきました．

（1）　画像認識技術

画像認識の技術としてもっとも歴史が古く，かつもっとも身近な例として，バーコードがあります．1940年代から実用化され，広く普及している技術です．

その認識技術に対するディープラーニングの革新性を印象付けたのが，カナダ・トロント大学教授のジェフリー・ヒントン氏らによる発明でした．教授らが開発した「AlexNet」は2012年，AIによる画像認識の精度を競うコンテスト「ISLVRC」で，２位以下のチームに大差で優勝しました．以来，ディープラーニングへの注目度が格

段に高まってきました．

（2） 音声認識技術

ディープラーニングは音声認識技術の進化も急加速させました．

最も初期の音声認識の研究は1970年代，米国で始まります．米軍など政府の野心的研究に続き，企業として世界初となる音声認識技術をIBMが開発しました．その後，綿々と発展を遂げながら，特に2010年代に米Appleのスマートフォン iPhoneに搭載された「Siri（シリ）」や米Amazonの「Echo（エコー）」など，音声デバイスが普及期に入りました．

現行の一般的な音声認識の仕組みは，4つの要素から成ります．すなわち，アナログの音声情報をデジタル化する「音響分析」と，そうしてつくられたデータから音素を抽出する「音響モデル」，その音素ごとにモデル化された膨大なデータベースを指す「発音辞書」，そして「言語モデル」です．

その言語モデルこそ，ChatGPTをはじめとするテキスト生成AIの中核技術となっています．なお，GPTは「Generative Pre-trained Transformer（事前学習済み生成トランスフォーマー）」を表し，末尾の「T」のTransformerが言語モデルの大規模化を促し，今の生成AIブームを支える革新的発明でした．

LLMの仕組みと特徴

■言語モデルの仕組み

LLMの構造を理解するには，その大本となる言語モデルを理解する必要があります．

言語モデルとは，テキストを生成したり理解したりするために使用される確率モデルです．テキストの過去の使用例に基づいて，テキストの次の単語を予測し，文章にしていく仕組みです．言語モデルはさまざまな方法で構成されますが，現在は脳の構造を模したコンピューティング技術「ニューラルネットワーク」を使うのが主流です．

AIにおける言語モデル発展の歴史，その始まりは1950年ごろまでさかのぼります．当初は文法規則に基づくモデルなど，主に「ルールベース」と呼ばれるアプローチが主流でした．1980年代になると，統計的手法が導入され，単語の並びや文脈から言語のパターンを学習するようになりました．

時代が下り，1990～2000年代初頭に主流はその統計的な手法へと移りました．大量のテキストデータから言語のパターンを学習するモデルが開発されるようになります．

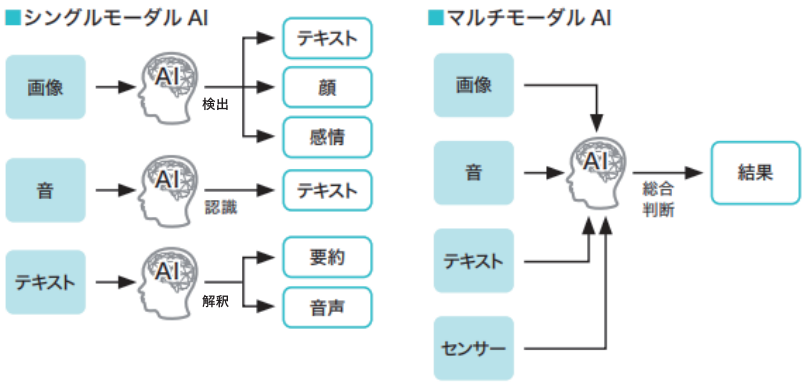
2000年代後半以降はインターネットの普及に伴いデータ量が急増するにつれて言語モデルが発展しました．ニューラルネットワークに基づくモデルの開発が加速し，より複雑な言語の特徴をとらえ，精度の高いモデルが実現されるようになりました．

さらに2010年代には，後述の「Transformer」が革新を起こし，それを活用したOpenAIの「GPT」シリーズをはじめとした事前学習モデルが登場しました．大量のテキストデータを用いた事前学習モデルは，特定のタスクに微調整（ファインチューニング）を加えることで，さまざまな言語処理タスクにおいて高い性能を発揮しています．

この言語モデルの構造を図式化すると**図3**のようになります．大まかに，言語モデル構築の基本プロセスとして，「生データ」の収集に始まり，そのデータに「クリーニング・正規化」を施し，トレーニング用のデータセットを作成，トークン化を経て，パラメータやアルゴリズムを加えたものが基本構造となります．

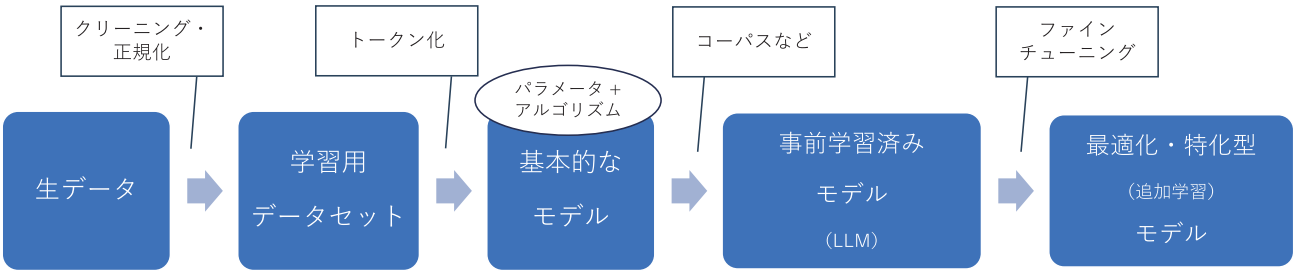
「パラメータ」は，モデルの構成要素であり，モデルの挙動を定義し，エラーを最小限に抑えるためにトレーニング中に調整されます．

まず生データは，誤字脱字など不純物が混じっているため，誤字の修正や日付形式の標準化といったかたちでノイズや不整合を取り除く作業が必要となります．これにより，トレーニングデータの品質が高まり



情報通信総合研究所作成

図2 シングルモーダル・マルチモーダル違い



情報通信総合研究所作成

図3 LLM構築のフロー

ます．さらにそのトレーニング用データを「トークン化」します．

トークンとは，テキストデータを処理する際に基本となる単位であり，生成AIを支えるLLMによるテキストの理解や生成に不可欠な要素です．一続きのテキストを個々の単語，文字，サブワードなどの小さな単位のトークンに分割して構成され，その手法は言語やモデルの要件に応じて異なります．

例えば，単語トークン化はテキストを単語ごとに分割する方法であり，"I like apples."という文を["I", "like", "apples", "."]というトークンに分割します．一方，テキストを個々の文字に細分する「文字トークン化」や「句読点トークン化」といった方法など万別です（**表3**）．

この時点で一定規模の語彙が集まり，そこに言語モデルのアルゴリズムやパラメータを付加することにより，体系化された基本的なモデルが出来上がります．

■言語モデルのスケーリング則

まだ粗削りともいえるこの序盤の言語モデルを，大規模モデルへ進化させるためには，トレーニングをする必要があります．

一般的に，このパラメータ数のほか，「計算量」「データ量」を巨大化させることで，言語モデルは，ChatGPTのように正確で自然なテキスト生成が可能なLLMとなります（**図4**）．ChatGPTを開発したOpenAIが2020年に発表した論文で，その効果が明らかになり，各社で大規模化をめざす競争が過熱していきました[5]．

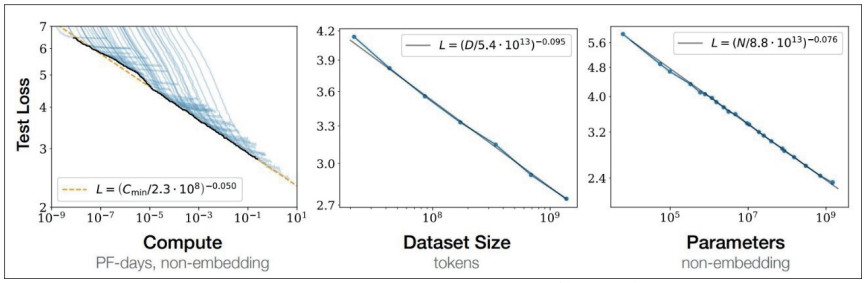
ただ，規模を追うこうした開発は，コンピュータの処理量の増大に伴う電力の大量消費などの問題をはらみ，持続可能性の観点から課題も少なくありません．そうした中，NTTの「tsuzumi」のようにパラメータ数を小さく抑えながら，高性能なモデルの重要性が高まっています．

粗削りのモデルを大規模化するためには，まずネット上の記事や論文，Wikipediaなどのオープンソースや SNSの投稿など，多種多様なデータセットに基づいた事前学習を行います．事前学習は，言語モデルが最初に大規模なコーパス上でトレーニングされるプロセスです．これにより，特定のタスク向けのデータで微調整される前に，幅広い言語の特徴を学習します．

表3　トークン化の種類と差異

種類		トークン化された結果
単語 トークン化	句読点を無視して、意味を持つ最小単位の単語に、文章を分割	['今日','は','いい','天気','です','ね']
文字 トークン化	文を、句読点も含めて個々の文字に分割	['今','日','は','、','い','い','天','気','です','ね','!']
句読点 トークン化	文を単語に分割すると同時に、句読点を独立したトークンとして認識	['今日','は','、','いい','天気','です','ね','!']

情報通信総合研究所作成



大規模化に伴い乗数が増える法則性を示したグラフ．左から「計算量」「データ量」「パラメータ数」で，それぞれの規模の巨大化により言語モデルの性能が格段に向上することを示している．

出典："Scaling Laws for Neural Language Models"

図4　言語モデルのスケーリング則

その「コーパス」とは，AIが人間の言葉を理解するうえで欠かせない辞書のようなデータ集です．自然言語の文章や使い方を大規模に収集し，コンピュータで検索できるよう整理されたデータベースです．コーパスは「言語全集」とも呼ばれ，自然言語を扱うAIにとって最重要ツールの1つです．

コーパスは，新聞や雑誌，本，インターネット上のテキストなど，さまざまなメディアから収集された自然言語によって構成されています．これらのデータは構造化され，品詞や文法情報も付与されています．AIが非構造化データとして存在する無数の自然言語を「読む」ための辞書のような役割を果たします．

図3の最終工程にあるファインチューニングは，パラメータの微調整などによりモデルを最適化します．あるタスクに特化して性能を高めるためのプロセスといえます．

まとめれば，LLMは，膨大なテキストデータから，単語やフレーズの出現パターンを学習し，一定の回数をこなした後，検証用データでテストし，その結果を踏まえて微調整する2段階のプロセスを経て完成します．

こうした一連のプロセスは，自然言語処理と総称されます．

■日進月歩の自然言語処理

自然言語処理（NLP：Natural Language Processing）は，人間が日常で使う言葉「自然言語」をコンピュータが識別，抽出する技術を指します．

コミュニケーション上の話し言葉や，行政文書や論文のフォーマルな書き言葉などが自然言語処理の対象であり，言葉の意味を多面的に解析します．「言葉」をコンピュータが理解する技術である自然言語処理は，近年飛躍的に進歩し，生成AIの台頭に結び付きました．

自然言語処理では，文章の構造や全体像を読み解く「形態素解析」，単語どうしを結び付ける「構文解析」，フレーズごとの相関性を表す「意味解析」，文章の流れの整合を確認する「文脈解析」といった各工程で自然言語を処理します．

自然言語処理技術の進化は，深層学習の進歩に支えられています．最先端のシステムは，数千億にも及ぶパラメータを持つ「LLM」を学習させることで，高度な言語処理能力が備わっています．自然言語処理技術が進化することで，質問やリクエストへの違和感のない応答，機械翻訳やWeb検索がより高い次元で可能となりました．

そして，自然言語処理において革新的な役割を果たしたのが，「Transformer」でした．

■Google発の革新的技術，Transformer

Transformerとは，「Attention Is All You Need」というGoogleの研究者らによる2017年の論文で紹介されたディープラーニングモデルです[6]．従来の言語モデルが単語の出現確率を学習していたのに対し，Transformerは単語の順序を考慮した「自己注意」のAttention層のみを用いた点が最大の特徴です．

このAttentionのメカニズムは，例えば人間が見聞きした情報の特定個所に「注意（Attention）」を払うように，その仕組みを模倣し，AIが入力データの一部に注意，着目するよう学習させる技術です．例えば，人間が猫を認識する際，「顔」や「体形」から「これは猫」と判別する，あるいは英文の穴埋め問題では解答個所の周辺の単語に特に注意を向けます．それらと同様の行為を機械的にプログラムし，入力データの一部に対する注目度を高め，相対的に他の部分では低める効果を持たせることにより，人間に近いかたちで画像や文章を認識可能としました．

以前のニューラルネットワークは，AIの学習に必要な正解ラベル付きデータを大量に用意しなければなりませんでしたが，Transformerではラベル付きデータがかなり少なくて済みます．その結果，ラベルが付いていないような，Web上の膨大なデータや企業のデータベース内の情報も，効果的に利用できるようになりました．さらに，Transformerの計算は並列処理に適しているため，高速なモデルの実行が可能です．このモデルは，GoogleやMicrosoftの検索エンジンをはじめ，多くのAIのアプリケーションやサービスに採用されています．また，2018年にGoogleが発表した自然言語処理モデルであるBERT（Bidirectional Encoder Representations from Transformers）やGPTといったTransformerを基盤とした発展形のモデルが生まれ（**図5**），LLMの開発をめぐり各企業が切磋琢磨しています．一方，Transformerのさらなる効率化やシンプル化をめざす研究も進められており，より少ないパラメータで最大の性能を引き出すことをめざしています．

なお，Transformerは，言語モデルとして使用されるだけでなく，画像認識や音声認識などのタスクにも適用されています．自然言語処理の多くのタスクでモデルのベースとして使用されるTransformerのように，高い汎用性を示すLLMは，「基盤モデル（Foundation Model）」とも呼ばれます．

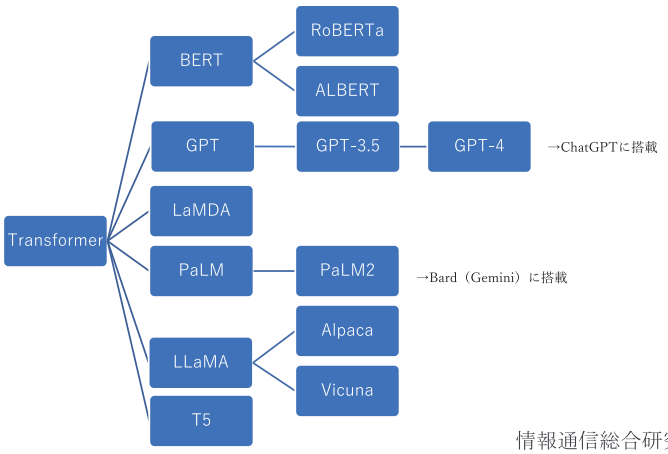
■増えるプレイヤ，広がる市場

こうしたLLM全般にわたり，新聞社や雑誌社をはじめとするメディアなどのデータを保有する企業や，それらを整理して構造化データにまとめ上げる企業，それを指南するコンサルティングファームやプラットフォームサービスを提供する企業など，プレイヤの裾野は広がっています．さらに，AIの高機能化，高速計算に欠かせないGPU（Graphical Processing Unit）の供給源である，NVIDIAに代表される半導体メーカが最重要プレイヤとして注目度が高まっています（**図6**）．

多国籍調査会社QYリサーチによると，LLMを取り巻く市場は，2022年の105億ドルから，2029年には408億ドルまで，年平均成長率21.4%で大きく伸長すると予測されます．

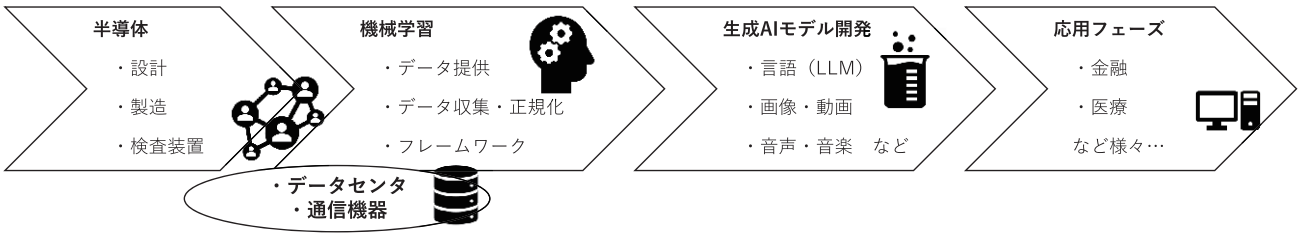
生成AIの普及・拡大による波及効果でもっとも潤う市場の1つが半導体産業です．米調査会社ガートナーの2023年12月4日の発表によると，同年の世界半導体売上高は前年比10.9%減の5340億ドルだったのに対し，2024年は同16.8%増の6240億米ドルまで伸長し，過去最高を更新する見込みです[7]．

関連して半導体の製造装置や検査装置の需要も大きな伸びが予想されます．製造過



情報通信総合研究所作成

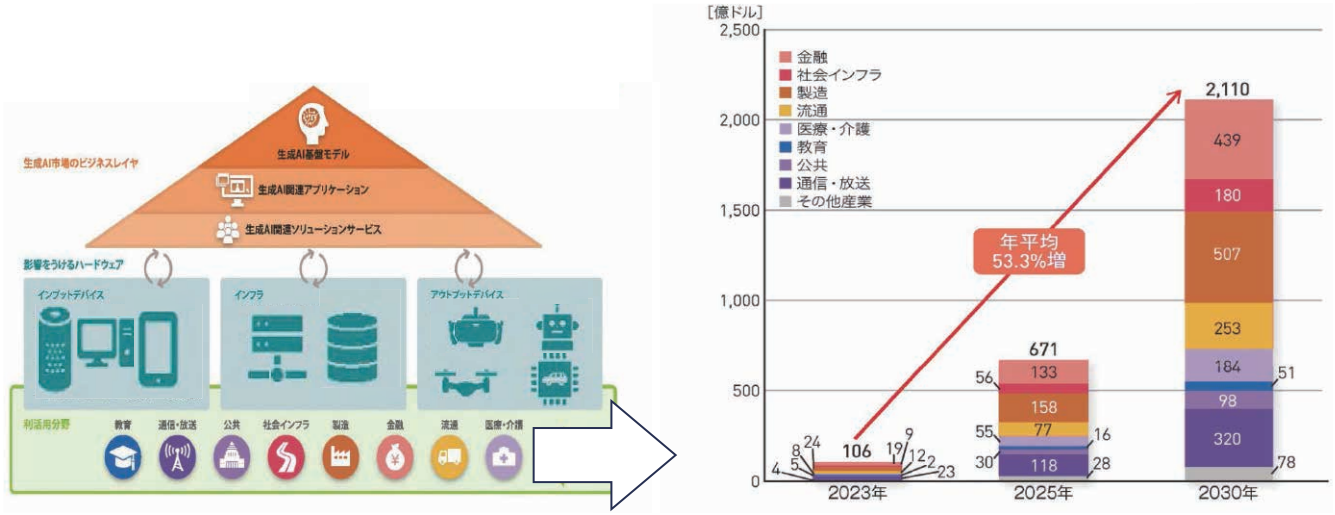
図5　Transformerから派生した言語モデルの例



※半導体産業の主要関連企業は「For The Future:“世界中が熱い！半導体政策・動向を紐解く－前編－，”NTT技術ジャーナル，Vol.35, No.7, pp.10-15, 2023」を参照

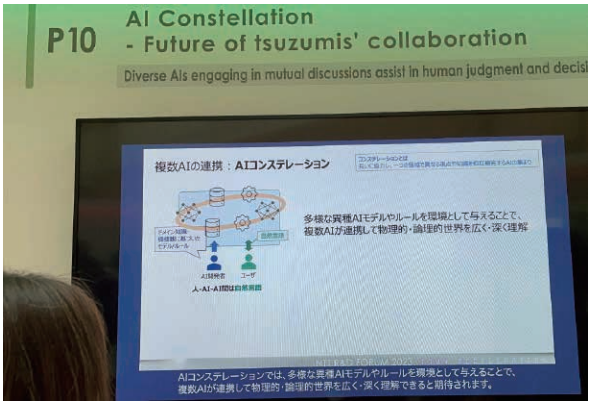
情報通信総合研究所作成

図6　生成AIブームが波及する主な関連市場



出典：JEITA:“生成AI市場の世界需要額見通しを発表”を一部加工

図7　生成AIの世界需要額見通し



複数のAIが連携して議論を深める「AIコンステレーション」（2023年10月，筆者撮影）

図8　AIコンステレーションの展示

程において欠かせない薬剤フォトレジストなどの原材料も需要増が見込まれます．例えば半導体計測および検査装置の市場規模は，2024年に約105億ドルと推定され，2029年までに約135億ドルに達すると，調査会社モルドールインテリジェンスは予測しています[8]．

当然ながら世界中で扱われる情報量が急増し，データセンタの需要も増します．インドの市場調査会社ストレーツリサーチによると，世界のデータセンタ市場規模は，1926億ドル強だった2021年から，2030年には5544億ドルに達すると予想されます[9]．

このほか，生成AIの応用フェーズではありとあらゆる分野が影響を受けます．米コンサルティングファーム，マッキンゼー・アンド・カンパニーによると，教育やアート，法務など多岐にわたるジャンルで，生成AIが業務の効率化に資するとの見通しを示しています[10]．

生成AI全体としては，2023年の106億ドルから，2030年には2110億ドルまで急速に成長すると電子情報技術産業協会（JEITA）は予測しています[11]（**図7**）．

■新たな職業の創発

LLMを実装した生成AIの普及に伴い，新たな職業も生まれつつあります．代表例として，テキスト生成AIへの指示文「プロンプト」を，最適化して出力の精度を高める「プロンプトエンジニア」のスキルが重視されています．その職種において「年収5000万円」といった好待遇からも，いかに重要なスキルか分かるでしょう[12]．

今後は記事を書いたり，スライドを作成したりといった作業も，AIが担う割合が増してくるはずです．ただ，生成AIは事実に基づかない回答をするケースが少なくありません．そうした真偽のチェック，誤字脱字の校正や文章の校閲といった新聞社のデスクのような仕事，「ファクトチェッカー」の仕事が増えると見込まれます．

また，出力された内容が事実だとしても，倫理観にもとるような出力をしてしまうことが想定されます．そうした問題を招かぬよう温和な表現にとどめる，手心を加えるスキルも今後一層求められるでしょう．「経営倫理士」ならぬ，「AI倫理士」といった職種ができるかもしれません．非倫理的，差別的な表現の発信は企業のレピュテーションリスクに直結し，ややもすれば「炎上」しかねないため，「防炎請負人」といった役目を担いそうです．

AIの生成物をめぐっては，入力から出力の間の因果関係が見えにくい「ブラックボックス問題」が長年課題となってきました．XAI（Explainable AI：説明可能なAI）の発展が期待される中，入出力の過程，そのブラックボックスの中身を，説得力のある根拠とともに示す「仮説構築力」，「論理的思考」を持ち合わせた人物の役割は一層高まりそうです．いうなれば「仮説検証士」「立証士」といったところでしょうか．

さらにAIの未来の姿として，「AIどうし

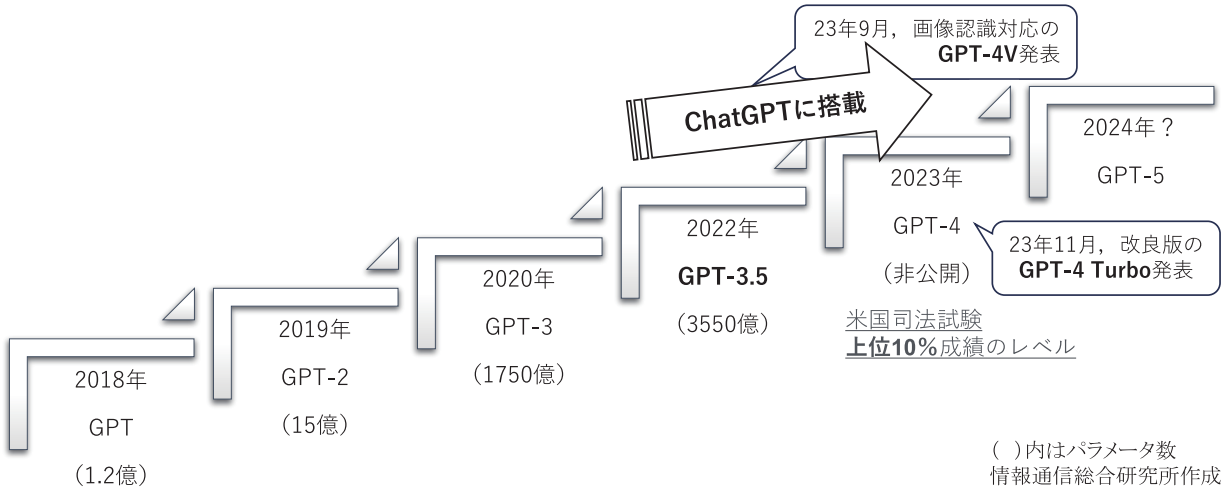
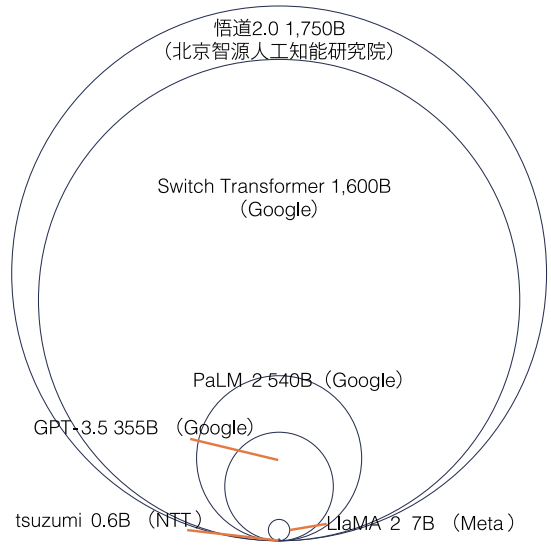


図9　ChatGPTの進化



単位のBは十億パラメータ．パラメータ数の大規模化が進む一方，小規模で高性能なモデルも増えつつある．

情報通信総合研究所作成

図10　主な大規模言語モデル

の対話」が繰り広げられる将来像も浮かんできます．2023年10月にNTT武蔵野研究開発センタで開催された「NTT R&D FORUM 2023」においても，生成AIの展示が注目の的となっていました．その1つに，将来的なイメージとして「AI コンステレーション」が紹介されていました（**図8**）．あるテーマについて，法律家や教員，政治家など専門家の知識を持った複数のAIどうしが議論し，望ましい方向性を導き出していくといったAIの活用法です．とはいえ，最終的に結論を下す，判断するのは人間です．そのため，AIどうしの議論を見守り，まとめ上げていく「ファシリテータ」「モデレータ」のような役割も一層重要視されるようになるでしょう．

代表的な言語モデル

■海外勢，パラメータ数1兆超えも

LLMについて，その区分はさまざまありますが，パラメータ数1000億を1つの基準として分ける分類法があります．

GPTシリーズでいえば，2019年リリースのGPT-2のパラメータ数が15億だったのに対し，翌2020年に出たGPT-3は一気に1750億となり，性能も格段に向上しました．この大規模化がChatGPTの成果へつながる突破口となりました（**図9**）．

海外の主だったLLMのうち，ChatGPTに搭載されたGPT-3.5は3550億のパラメータ数です．さらにはそれを大きく上回る1兆超えの言語モデルをGoogleなどが相次いでリリースしています（**図10**）．

ただ，1兆を超えるLLMが，相応の成果を上げ，評価を得られるかは未知数です．また，GPT-3の1回の学習に要する電力量は1300 MWhで，原発1基の1時間分に相当し，省電力化が課題とされています．それをしのぐ兆単位のモデルは推して知るべしです．

むしろ，パラメータ数を抑えつつ，高精度の出力を発揮する手法が，持続可能性の観点からは支持されつつあります．とりわけ，OpenAIやGoogleといった米企業が席巻する生成AI市場にあって，日本勢の勝機はまず英語ベースのLLMが適応しきれていない，日本語に特化した生成AIの充実にこそ見出せるといえます．中でも専門用語や専門知識が多い医療や金融といった特定領域に絞るほうが価値を生みやすいと見込まれます．そうした観点から，パラメータ数が6億の「超軽量版」と70億の「軽量版」をそろえ，金融や医療といった特定分野に強い，NTTのLLM，「tsuzumi」は1つの勝ち筋になり得るかもしれません．実際，米中はじめ生成AIで先行する各国でも業界特化型の生成AIは続々登場しています．さらに，Sakana AI社へ出資したNTTドコモ・ベンチャーズによる次世代

表 4　生成AIの各国市場トップ5（10億ドル）

	2020年		2023年		2025年		2030年	
1.	米国	2.33	米国	16.14	米国	30.25	米国	65.71
2.	中国	0.49	中国	5.45	中国	11.61	中国	29.55
3.	英国	0.28	ドイツ	1.90	ドイツ	3.65	**日本**	8.68
4.	ドイツ	0.26	英国	1.82	**日本**	3.62	ドイツ	8.31
5.	**日本**	0.23	**日本**	1.79	英国	3.39	英国	7.38

出典：Statista

表 5　日本の主なLLM

サービス名	提供元	パラメータ数	特徴
tsuzumi	NTT	6億/70億	日本語処理に特化した軽量モデル．金融や医療など特定領域に強み
japanese-large-lm	LINE（現LINEヤフー）	17億/36億	日本語に特化したオープンソースとして公開，商用利用も可
CyberAgentLM2-7B	サイバーエージェント	70億	チャット形式にチューニングされたバージョンも．2023年5月の進化版
Japanese StableLM Alpha	Stability AI Japan	70億	学習データは主に日本語と英語，加えてソースコード約2%．画像生成と連動も
Weblab-10B	東京大学松尾研究室	100億	日本語と英語のデータセットを用いた高精度多言語モデル．事前学習済みモデル・事後学習済みモデルの商用利用不可
PLaMo	Preferred Networks	130億	自社スーパーコンピューター「MN-2」を利用して学習
cotomi	NEC	130億	1回最大30万字の長文プロンプトに対応．ことばにより未来を示し，「「こと」が「みのる」ように」という想いを込めた名称
（開発中）	NICT（情報通信研究機構）	400億	350GBの日本語テキストを用いた高品質な日本語特化型モデル．さらに大規模な1790億パラメータのモデル開発中
LLM-jp	産業技術総合研究所，東京工業大学，国立情報学研究所	1750億	日本語特化の大規模モデル構築に着手．産総研の計算資源であるAI橋渡しクラウド（ABCI）を使用

発表年はいずれも2023年，会社名等は2024年2月1日時点
情報通信総合研究所作成

生成AI基盤モデル開発など，「tsuzumi」を有効活用しようとする動きは出始めています(13)．

日本語の市場のみといって，過小評価されるべきではありません．ドイツの調査会社スタティスタによると，生成AI市場は拡大し続けており，日本市場は2023年時点で米国，中国，ドイツ，英国に次いで5番目に位置付けられています（**表4**）．注目すべきは，2024～2025年までには英国を抜いて4番手につき，2030年には3位に浮上すると見込まれていることです．予測の不確実性は常に付きまといつつも，有望視されている以上，今後海外プレイヤの攻勢は必至でしょう．そうした中，日本企業が自らの市場を守る，あるいは新たに築く意味でも国産生成AIの意義は決して小さくないはずです．

なお，日本勢が海外で抗戦するにはさらなる工夫や戦うべき市場の精査が欠かせません（**表5**）．ChatGPTや2023年に発表されたGoogleの対話型AIサービス，Bard（現Gemini）のような大規模で総花的なマルチモーダルをめざすのか，あるいは後塵を拝している先駆者の台頭に伴って新たに生まれる市場，ビジネスチャンスに先鞭を付けるのか．選択を見誤ると，すでにレッドオーシャンで血みどろの争いをしているプレイヤから返り討ちに遭うだけかもしれません．

■参考文献

(1) https://www.ai-gakkai.or.jp/whatsai/AItopics5.html
(2) https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html
(3) https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r01/html/nd113200.html
(4) https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/html/nc142120.html
(5) https://arxiv.org/pdf/2001.08361v1.pdf
(6) https://arxiv.org/pdf/1706.03762.pdf
(7) https://www.gartner.com/en/newsroom/press-releases/2023-12-04-gartner-forecasts-worldwide-semiconductor-revenue-to-grow-17-percent-in-2024
(8) https://www.mordorintelligence.com/ja/industry-reports/semiconductor-inspection-equipment-market
(9) https://straitsresearch.com/jp/report/data-center-market
(10) McKinsey Global Institute："Generative AI and the future of work in America，" 2023.
(11) https://www.jeita.or.jp/cgi-bin/topics/detail.cgi?n=4724&ca=1
(12) https://www3.nhk.or.jp/news/html/20230518/k10014071011000.html
(13) https://kyodonewsprwire.jp/release/202401175415

株式会社 情報通信総合研究所
主任研究員　南龍太