AIをめぐる技術と業界の動向

青山 祐輔 ●株式会社 企

AI開発競争が過熱するも、学習データの権利問題やリソースの確保など課題は多い。一方で効率化やエッジAIの開発、マルチモーダル化が進み、顧客対応や画像生成など多くの場面でAIが実用化されつつある。

■進むAIの性能向上と制約

生成AIが登場したことで、AIによるさまざまなタスクの自動化や効率化が期待され、ここ数年は多くの実証に向けた取り組みがなされてきた。当初は技術的な検証にとどまっていたが、2024年には実際の業務で用いられるケースも増えており、AIは実用期へと突入しつつある。

改めてこれまでのAIの流れを振り返ると、2010年代にはディープラーニングによる画像認識技術の驚異的な向上や、プロを超えるほどの腕前を持つ囲碁・将棋プログラムが登場した。これによって、20世紀から研究開発が行われてきた人工知能＝AIという技術が、ようやく実用レベルになったことを人々に印象付けた。そして、2022年以降は「Midjourney」や「ChatGPT」が発端となった生成AIブームにより、人々はようやく「AIが人間を超えるかもしれない」と実感するに至った。特に前出のChatGPTをはじめとする大規模言語モデル（Large Language Model：LLM）は、登場直後から不完全ながらも、その高い言語処理能力がホワイトカラーのビジネスパーソンが業務に利用できるレベルだったことから、「AIによって人間の仕事が奪われる」ということが現実的な社会課題の一つとして議論されるまでになった。

2024年に入っても、生成AIをめぐる騒動は続いた。その一つがAI開発企業による性能競争だ。LLMは学習データの量とそれにともなうパラメータ数が増加するほど、性能も向上することから、この点で他社への優位性を示すことができる。OpenAIが開発するChatGPTのパラメータ数は、初めて一般向けに公開されたGPT-3.5（2022年11月リリース）において約3500億だったが、GPT-4（2023年3月）では1〜2兆、さらにGPT-4o（2024年5月）では100兆にまで達したと推定されている（OpenAIは公式に発表していない）。追いかけるグーグルも、最初のGemini（2023年12月）では約6000億、Gemini 1.5（2024年2月）では約1.5兆に到達した。

一方で、こうした物量での競争においては、学習に必要なデータやコンピューティングリソースも莫大なものとなるため、その確保においてさまざまな課題が生じている。学習データにおいては、2023年12月に米ニューヨーク・タイムズ紙がOpenAIを著作権侵害によって告訴したように、権利者側との十分なすりあわせが行われないまま、AIモデルの開発にコンテンツが使われている例もある。これまで生成AIの学習データについては、開発者自身が出所を明確にするケースは多くなかった。しかし、学習データの一部にウェブサイトから収集したコンテンツが利用されてい

ることは明白であり、そのデータ利用についての適法性は疑問視されてきた。ニューヨーク・タイムズ紙のような法的措置を取る権利者も出てきており、今後の学習データについては適切なガバナンスの下での収集と利用が求められるだろう。

AIの開発と利用に欠かせないコンピューティングリソースについても、データセンターとそこで使われるAI向けのプロセッサーの獲得競争という課題が生じている。AIに限らず世界的にクラウド利用は拡大しており、DXが進展することでさらにコンピューティングリソースの需要は高まり、データセンターも増え続ける。しかし、電力確保や経済安全保障の問題もあり、その設置には制限がある。また、AI開発においてはエヌビディア製の半導体がデファクトスタンダードとなっているが、需要に供給が追いつかない状況が常態化している。グーグルやマイクロソフトなどは独自にAI向け半導体を開発しているものの、長らくこの分野での半導体開発を行ってきたエヌビディアの開発力、供給能力との差は大きく、当面はエヌビディアへの依存が続くと思われる。

■実用に向けたAIの技術開発

こうした課題を背景に、単純な性能競争とは別に、2024年のAI開発は実用性にフィーチャーする流れが加速した。その大きなポイントがAIモデルの効率化だ。

LLMはパラメータ数が多いほど性能が向上するが、そうした巨大なAIモデルは利用時に必要とするコンピューティングリソースも大きいため、一般的にはクラウド上で動作するサービスとして利用される。だが、さまざまなユースケースを想定した場合、巨大なパラメータ数のLLMを必要としないことも多い。そこでLLMにおいても、数億から数百億程度にパラメータ数を抑え、より多くの環境で利用できるモデルが開発されている。

Anthropicが開発するLLM「Claude」は、バージョン3において、最も高性能で複雑なタスクも可能なOpus、性能と速度のバランスがとれたSonnet、軽量で処理速度に特化したHaikuという3つのモデルを提供し、用途に応じて使い分けることを提案している。

また、NTTが開発した「tsuzumi」は、日本語に特化しつつ、学習でOpenAIのGPT-3の数十から数百分の一、推論に数十分の一のリソースで開発されており、国内においてさまざまな用途や環境で利用できる。ほかにも楽天やNECなども独自に日本語に特化した軽量なLLMを開発しており、ビッグテックの大規模物量とは異なるAI戦略を狙っている。

さらに、Mixture of Experts（Moe）という手法をとるAIモデルも登場している。これは巨大な1つのAIモデルではなく、データを分野ごとにそれぞれ別のAIモデルに学習させ、それらを連携することで高度な推論を実現するという手法だ。それぞれの専門分野のAIは、比較的小規模なパラメータ数に収まるため、超大規模なパラメータ数のLLMよりも、該当する専門分野での性能や費用対効果に優れる。

また、AIの利用においてさまざまなユースケースが登場した結果、通信を必要とするクラウドではなく、ローカル（エッジ）で動作する方が効率的なケースもある。そこで、膨大なコンピューティングリソースが使えるクラウドではなく、小規模なデバイスでも動作するAI、いわゆる「エッジAI」にフォーカスした開発も広がっている。

例えば、アップルは独自のLLMである「MM1」をリリースしたが、これは用途に応じて30億、70億、300億という3つのパラメータ数を切り替えて利用できるようになっている。同社はiPhoneやMacといったユーザーが直接手に取るエッジデバイスに「Apple Intelligence」としてAIの搭

載を進めており、クラウドで動作する大規模なAIモデルではなく、ローカルで動作するエッジAIによるユーザー体験の向上を狙っている。

また、マイクロソフトはWindows 11にCopilotをはじめとするいくつものAI機能を追加した。それと同時に、Windows 11 バージョン24H2をAIの動作に特化したニューラルプロセッサーを搭載したPCと合わせて利用する「Copilot+PC」というブランディングを展開している。これによりウェブサービスに登録したり課金したりすることなく、手元のPCだけで文書や画像の作成にAIを用いたり、外国語を翻訳したりすることができる。こうした日常的なAI利用が手元のPCだけで完結することで、より多くの人々がAIに親しみ、その便益を得られるようになるだろう。

AIの利用シーンの拡大という観点から、実装された多くのAIサービスのマルチモーダル化も進んでいる。マルチモーダルAIとは、これまでの生成AIのようにテキストだけ、画像だけではなく、1つのAIで複数の種類のデータを同時に扱うことができるAIモデルのことだ。マルチモーダルAIならば、テキストで指示を与えた際にテキストだけでなく画像も合わせて生成したり、逆に画像や動画の入力に対して、その内容を文章で記述したり要約したりすることができる。このように、従来はテキストや画像などそれぞれに適したAIを複数組み合わせて使わなければできなかった作業が、1つのAIで行えるのである。AIの利用シーンの1つのゴールとして、人間に代わってさまざまなタスクを処理する「AIエージェント」が想定されているが、マルチモーダル化はその方向へと着実に歩みを進めるものだろう。

■さまざまな場面でAIの実装が進む

このほかにもさまざまな場面でAIの実利用が進んでいる。その一つが、顧客対応におけるチャットや電話対応のAI化だ。ユーザーサポートなどの電話やチャットの対応が、人間のオペレーターではなくAIでできるようになった。これは検索拡張生成（Retrieval-Augmented Generation：RAG）と呼ばれる技術を応用している。ユーザーへの対応マニュアル、製品・サービスの説明書などの文章をあらかじめデータベース化しておき、ユーザーからの入力に基づいてデータベースを検索し、その結果を基に回答を生成するというものだ。AIには、学習データにない架空の情報を生み出してしまうハルシネーション（幻覚）という現象があるが、RAGによってその発生を抑え回答や振る舞いの精度を向上できる。RAGは準備するデータベース次第でさまざまな場面で利用可能なため、教育、医療、学術など多くの分野で導入が進んでいる。

同じように実社会での導入が進んでいるのが、画像の生成だ。MidjourneyやStable Diffusionといった先行するサービスやアプリに続き、グーグルも写真と見まがうほどの画像を生成できる「ImageFX」を2024年8月にリリースした。また、これまでは画像生成で人物を描画する際、どのような人物が描かれるかはランダムで、気に入った人物が描画されても、その人物をほかの画像に再登場させるのは困難だった。しかし、Midjourneyの「Cref」という機能ならば、生成された人物を再利用してほかのシチュエーションやポーズで画像を生成できることから、クリエーターにとって画像生成AIはより使いやすいものとなった。

また、画像に続いて動画の生成AIも進化が著しい。これまでも動画生成AIは存在したが、登場する人物や背景が崩れるなど、出来の悪い悪夢のような映像しか作ることができない実験室レベルのものだった。しかし、2024年2月にOpenAIが「Sora」において人物の一貫性を維持した動画生成を実現し、さらにLuma AIの動画生成AIサー

ビス「Dream Machine」（2024年6月リリース）や「Runway Gem-3 Alpha Turbo」（2024年9月リリース）などの新たな動画生成AIが登場するたびに、着実に性能向上を見て取ることができた。

AIをめぐる技術開発のスピードは衰えを知らず、以前は不十分だと思われていた技術が、1年後には実用的なものとなり、多くの場面で使われるようになっている。その処理能力は、大学入試レベルの問題なら合格できる水準に達しており、この急速な発展を受けて、OpenAIのあるエンジニアは「すでにAGI（Artificial General Intelligence、汎用人工知能）のレベルに到達した」とX上で発言した。さらには本項を執筆している2025年1月末には、中国のDeepSeekが開発したLLM「DeepSeek R1」が登場した。このLLMは最新のGPUを利用せずに開発されており、かつオープンソースとして公開されたことで世界中のAI開発者を驚愕させている。ロボティクス技術とAIの融合も進んでおり、サイバー空間だけでなく物理空間においてもAIが活躍する世界が近づいている。

その一方で、AIの利用に伴うさまざまな課題への対応は十分とは言えず、世界各地で問題を引き起こしている。世界各国もAI開発を推進する一方で規制の網を掛けようとしており、その綱引きのバランスの落とし所は見えていない。

インターネット白書 ARCHIVES

1996,1997,1998,1999,2000....

［インターネット白書ARCHIVES］ご利用上の注意

このファイルは、株式会社インプレスR&Dおよび株式会社インプレスが1996年～2025年までに発行したインターネットの年鑑『インターネット白書』の誌面をPDF化し、「インターネット白書ARCHIVES」として以下のウェブサイトで公開しているものです。

https://IWParchives.jp/

このファイルをご利用いただくにあたり、下記の注意事項を必ずお読みください。

- 記載されている内容（技術解説、データ、URL、名称など）は発行当時のものです。
- 収録されている内容は著作権法上の保護を受けています。著作権はそれぞれの記事の著作者（執筆者、写真・図の作成者、編集部など）が保持しています。
- 著作者から許諾が得られなかった著作物は掲載されていない場合があります。
- このファイルの内容を改変したり、商用目的として再利用したりすることはできません。あくまで個人や企業の非商用利用での閲覧、複製、送信に限られます。
- 収録されている内容を何らかの媒体に引用としてご利用される際は、出典として媒体名および年号、該当ページ番号、発行元などの情報をご明記ください。
- オリジナルの発行時点では、株式会社インプレスR&Dおよび株式会社インプレスと著作権者は内容が正確なものであるように最大限に努めましたが、すべての情報が完全に正確であることは保証できません。このファイルの内容に起因する直接的および間接的な損害に対して、一切の責任を負いません。お客様個人の責任においてご利用ください。

お問い合わせ先 | インプレス・サステナブルラボ | ✉ iwp-info@impress.co.jp