オープンデータと生成AI

Open Data and Generative AI

国立情報学研究所所長／京都大学特定教授 黒橋 禎夫

Sadao Kurohashi, Director General, National Institute of Informatics / Specially Appointed Professor, Kyoto University

ORCID ID : https://orcid.org/0000-0001-5398-8399

1. はじめに

環境問題，格差問題，地域紛争など，複雑かつ相互に関連する現代の社会課題を打開するためには，多様なバックグラウンドや専門性を持つ人々の協力が必要である．そのためには，我々の根本的な考え方を競争から共創へ，クローズドからオープンへと変えていく必要がある．その具体的な第一歩はオープンデータの促進であろう．

オープンデータは，専門家の協業を促すだけでなく，昨今進展が著しい生成 AI をさらに発展させ，専門家と生成 AI の協業による社会課題解決という道を切り開く．さらに，生成 AI に支援された人々の活動が新たな知識やデータを生み出すという好循環を期待することもできる．そのような将来を見据えつつ，本稿ではオープンデータと生成 AI に関連する諸課題を議論する．

2. 学術分野のオープンデータ

COVID-19 パンデミック時に，研究者がオープンデータを活用して迅速にウイルスの特性を解明し，ワクチン開発を進めたことは記憶に新しい．ビッグデータ，データ駆動型社会など，データの重要性はさまざまな用語で語られてきたが，もはやその重要性を疑う人はいないだろう．問題は，いかにしてデータを継続的・網羅的に集積し，広く活用するかである．

2023 年 5 月に仙台で開催された G7 科学技術大臣会合では，新しい知識の創造，イノベーションの促進，地球規模の課題に対する解決策の開発に貢献するために，公的資金によって得られた研究データおよび学術出版物などの研究成果の公平な普及と，オープンサイエンスの推進が謳われた．この方針を受けて，我が国でも 2024 年 2 月に内閣府の統合イノベーション戦略推進会議が「学術論文等の即時オープンアクセスの実現に向けた基本方針」を示した [1]．2025 年度から，公的な競争的研究費の支援を受けた査読付き学術論文およびその根拠データについて，原則，学術雑誌への掲載後即時に機関リポジトリなどへ掲載することが義務づけられた．

我が国では，NII において 2017 年度から研究データ基盤システム（NII Research Data Cloud）の開発，提供が進んでいる．また，2022 年度に開始した「AI 等の活用を推進する研究データエコシステム構築事業」では，NII，理化学研究所，東京大学，名古屋大学，大阪大学が中核機関となり，各分野・機関の研究データをつなぐ全国的な研究データ基盤の高度化を進めている [2]．この中では，異なる分野間のデータ連携によるデータ駆動型研究のユースケースの創出を目的として 30 件超の公募研究も推進されている．また，2023 年度に東海地区と北陸地区，2024 年度に中国四国地区と九州地区に，研究データに関するコンソーシアムが立ち上がり，研究データ基盤の利活用に向けた普及・広報活動が推進されている．2024 年度には「オープンアクセス加速化事業」に 83 の大学，研究機関が採択され，即時オープンアクセスに向けた体制整備・システム改革が加速されている．

世界に目を転じると，研究データ基盤整備の先進地域は欧州であり，欧州オープンサイエンスクラウド（EOSC）には我が国とは桁違いの予算が投入され，研究基盤の統合環境の構築が進められている [3]．我が国においても，より大きな予算投資が行われるに越したことはないが，高齢化先進国として喫緊の課題が山積する中でそれは簡単ではない．遍在する縦割りの壁を乗り越えて，現状のプロジェクト間のシナジーを最大化し，学術オープンデータの成果を国民と政府に一歩ずつアピールするしかないだろう．

3. 生成AI技術の現状と課題

従来の AI は「分類」，すなわち，決められた課題に対して事前に用意された選択肢の中から最善のものを選ぶことが基本であった．これに対して，生成 AI はテキストや画像を「生成」することができる．その最大の効

果は，人間の思考における中心的メディアである言語を用いて人間とコミュニケーションできることであろう．

現在の生成 AI においては，言語，知識，対話などをつかさどる大規模言語モデル（Large Language Model; LLM）が，音声，画像，ロボット制御などの情報統合の中核を担っている．

3.1 LLM の技術的背景と歴史

OpenAI が 2022 年 11 月に公開した LLM に基づく ChatGPT は，人々に大きな衝撃を与えた [4]．その後も生成 AI は急速な進化を続け，徐々に社会での活用が広がっている．

そもそも言語モデルとは，自然言語処理の分野で古くから研究されてきたもので，あるテキスト（文脈）が与えられた時に，それに続く単語の確率を予測するものである．この確率は，コーパス（大規模なテキスト集合）における単語並びの出現を数えることで計算でき，機械翻訳や音声認識などの精度向上に寄与してきた．しかし，単語を単に記号として扱っていた時代には，文脈として考慮できるのはたかだか直前の数単語であった．この時代の言語モデルの実態は，「言語のモデル化」というには少々大げさな用語であった．

これが，ニューラルネットワークを用いて，単語を数百次元の実数値ベクトルとして表現するようになり，またネットワークの巨大化と，このあと述べる種々の技術によって，言語モデルにおいて考慮できる文脈が一気に広がった．現在では標準的に数千単語，Google の Gemini では 200 万単語の文脈が考慮されるようになっている．その学習方法は，与えられたコーパスにおいて次の単語を予測できるように，ニューラルネットワークのパラメータを逐次的に更新するという極めてシンプルなものである．

言語モデルの進展には，ベクトル表現の導入に加えて，機械翻訳研究の中から生まれた注意機構とよばれる技術も大きく貢献している．注意機構は，ニューラル機械翻訳において翻訳先の文を一単語ずつ生成していく際，原文のどのあたりに注目すべきかを計算する技術として生まれた．現在の LLM の基本となっている 2017 年に発表された Transformer も，もともとは注意機構を精緻化した機械翻訳のためのモデルであった．

そして，2022 年 3 月に発表された GPT-3.5 において取り入れられたファインチューニング（fine-tuning）が，LLM の能力を一気に現在のレベルに引き上げた．ファインチューニングでは，一般的なコーパスでの学習（これを事前学習とよぶ）を行ったあとに，質問と応答のペアの学習データを用意し，質問を与えた時にその応答が生成されるような調整を行う．その学習自体は事前学習と同じ枠組みであり，応答の表現が生成されるようにパラメータを逐次更新する．このファインチューニングを対話データに対して行ったものが 2022 年 11 月に発表された ChatGPT である．

時を移さず，2023 年 3 月に GPT-4 が発表された．GPT-4 はパラメータ数を含めてモデルの詳細は公開されていないが，そのサイズはおそらく 2 兆パラメータ程度と推測されている．それまでのモデルとの大きな違いとして，画像の入力・出力も可能となり，また，米国の大学入試の共通テスト，司法試験，医師試験などに合格するレベルのパフォーマンスとなった．

その後も生成 AI の進展はとどまるところがない．まず，RAG（Retrieval-Augmented Generation）と呼ばれる技術が普及した．これは，質問に対して，ある文書集合を対象とした情報検索を行い，その上位数件の文書，もしくはその中のより関連する箇所をプロンプトとして LLM に与え，その文書の情報を踏まえて LLM が回答する仕組みである．Google の Gemini や OpenAI の ChatGPT では，最新のウェブ検索と組み合わせることで，事前学習コーパスに含まれない新たな情報についても回答することができる．また，社内文書，マニュアルなどを検索対象とすることで特化型のサービスを実現することもできる．RAG のサービスを提供する IT ベンダーも多数存在する．また，LLM が必要に応じて自律的にプログラミングを行って問題を解くこともできるようになった．

さらに，OpenAI o1 に代表されるように，モデル内部でより長い推論，試行錯誤を行った上で回答を行うモデルも提案され，数学やコーディングなどのベンチマーク性能が大きく改善し，今後，ソフトウェア開発やさまざまな分野の研究開発での活用が期待されている．

モデルのアーキテクチャとしても，オープンなモデルでは Mixtral で初めて導入された MoE（Mixture of Experts）が進化を続けている（GPT-4 も非公開ではあるが MoE と推測されている）．Transformer では，注意機構のモジュールと，中間層 1 つの FFN（feedforward neural network）をつなげたものを 1 ユニットとし，これが何段か積み重ねられている．MoE では，この FFN 部分を多重化して学習し（この部分が Expert である），推論時には入力に合わせて適切な Expert を選択するので，総パラメータ数に対して推論効率がよい．2024 年の年末に発表された DeepSeek-V3 は，従来よりもはるかに多い 257 個の Expert を持ち，総パラメータ数 671B，実効パラメー

タ数 37B でありながら，GPT-4 に匹敵するベンチマーク性能を示して注目されている．

3.2 生成 AI のインパクト

急速に進展する生成 AI には，あらゆる学術分野，産業分野，そして社会全体に大きな影響を持つという包括性，将来的には人間と共存する知的レベルとなり得る革新性，さらに，それが加速度的に進展するという加速性などの特徴がある [1]．

科学技術においては，生成 AI の活用により仮説生成，仮説検証，論文による知識流通など科学技術の各ステップが大きく進展しようとしている．知識の細分化・専門化による学問領域の分断が指摘される中，生成 AI の活用による分野横断的な新たな知の創造も期待される．

産業界においても，生成 AI の活用による業務効率化・業務量削減が，労働力不足や長時間労働などの課題を解決する強力な一手となり得る．自動運転を含め，AI が自律的に判断を行う AI エージェントを社会がどのように受容していくか，慎重な議論・実験・設計が必要であろう．

生成 AI による社会への最大の波及効果として教育の変革が挙げられる．ROIS と NII の共催で毎月開催されている教育機関 DX シンポでは生成 AI の大学等での活用事例がよく紹介されている．2025 年 1 月には，生成 AI による線形代数の理解が OpenAI o1 などの出現によりこの 1 年で劇的に進化し，もはや優秀な学部生レベルであるとの報告があった [5]．生成 AI 技術は今後もまだまだ進展することが予想される．生成 AI が，授業内容について学生が抱える疑問にいつでも対話的に相談にのったり，大学院レベルの専門性のある議論の相手となる日も近いだろう．AI を批判的に活用し，課題を解決し，創造する能力を高める新たな教育のあり方を，慎重にそして早急に検討する必要がある．

3.3 生成 AI に関する課題

一方で，生成 AI に関する課題も多数存在する．ここではそれを 4 つの観点でまとめ，以下の節でそれぞれ簡潔に議論する．

3.3.1 生成 AI の能力に関する課題

2023 年当初は，人間と自然に対話するという AI の究極の目標をあっさりと実現してしまった ChatGPT に対して，逆に，あれもできないこれもできないという指摘が続いた．しかし，すでに紹介したとおり，RAG やプログラミングの利用，o1 的な推論技術によってその多くの問題は解決している．

一方で，いまだ解決されていない大きな問題として，事実と異なる内容を出力してしまうハルシネーション（hallucination）とよばれる問題がある．

そもそも，LLM が出力する言語表現が流暢でなかったり的外れであれば，人が信用することもない．流暢で，トピックを踏まえた応答であるため，確認することなく人がそれを受け入れてしまい，ビジネスの場面などで問題を引き起こす事例が報告されるようになってきている．現状では，ハルシネーションが一定程度発生することを理解した上で，LLM の用途を選ぶことが重要である．基本的には，ユーザが詳しい領域で，そのアシスタントとして LLM を用いるのであれば問題は少ないだろう．また，これは LLM 利用に限らず情報を利活用する際の基本的なリテラシーであるが，批判的思考の態度が極めて重要である．

3.3.2 生成 AI の悪用に関する課題

どのような道具にも善用と悪用がある．包丁も美味しい料理を作れるが人を傷つけることもできる．包丁は無許可で買えても，拳銃が無許可で買える社会には不安がある．生成 AI は拳銃のレベルであるとも言える．

生成 AI の悪用として，詐欺やフェイクニュース生成の補助，世論操作などの懸念がある．さらに深刻な問題として，コード生成 AI によるハッキングプログラムの生成や，CBRN（Chemical, Biological, Radiological, and Nuclear）兵器開発を容易にする懸念もある．

また，ユーザが悪意を持たずとも，非社会的な回答を生成したり，学習データにおけるバイアスを助長する発言を生成する可能性もあり，社会の分断などにつながる恐れがある．

技術的対策としては，悪用の想定質問とそれらに対する回答（理由を示しつつ情報提供を拒否する等）のデータセットを作ってファインチューニングを行うことや，LLM の外に入出力に対するフィルタリング機能（ガードレール）を設けること等が行われている．前者は，LLM の汎化能力により，データセットの想定質問そのものだけでなく，その種の質問全般に対して一定程度の回答拒否が実現できる．

このような課題に対する国際的関心は高く，2023 年 5 月の G7 広島サミットにおいて AI に関する国際的なルール作りを推進する広島 AI プロセスが立ち上がった [6]．2024 年 2 月には我が国において AI Safety Institute（AISI）が設立され [7]，2024 年 5 月には欧州連合で AI 法（AI Act）が採択されている．今後，生成 AI のメリット・デメリットを勘案しながら，ハードロー（法規制）とソフトロー（ガイドライン等）の設計・運用が進んでいくものと思われる．

3.3.3 生成 AI の学習データに関する課題

LLM の学習には大規模かつ一定程度良質なコーパスが必要であり，基本的にウェブコーパスを活用することになる．ウェブ上の情報はまさに玉石混交であり，嘘も有害情報も含まれるため，様々なフィルタリングを行う必要があるが，これはそれほど簡単ではない．

また，良質なデータの多くはコストをかけて作成された著作物である．生成AIの急激な技術革新にともない，生成 AI の学習データについて著作者の権利保護が大きな課題となってきている．米紙ニューヨーク・タイムズは，無断の記事使用が著作権侵害であるとして OpenAI と Microsoft を提訴している．

我が国では著作権法 30 条の 4 において，AI の開発・学習段階では，著作権者の許諾なく著作物を利用することが認められている．一方で，生成・利用段階では，その類似性・依拠性に従って著作権侵害となる可能性があるとされており，我が国でも著作者団体からの懸念の声が上がっている．2024 年 3 月に文化審議会著作権分科会法制度小委員会から公表された「AI と著作権に関する考え方について」は，現状の法律のもとで，権利の保護と AI の活用のバランスを踏まえ問題点を整理する内容であった [8]．

生成 AI 学習におけるデータの取り扱い，特に著作権者との関係について私見を述べたい．一旦オープンになっているデータは，社会の発展のために積極的に AI 学習に活用すべきであると考える．一旦オープンになっているデータとは，ウェブ上のコンテンツ，論文，放送映像等である．たとえばフランスでは，国家戦略として AI 学習のためのデータ整備が進められており，国立視聴覚研究所（INA）とフランス国立図書館（BNF）の公開データ活用の検討状況は参考にすべきであろう [9]．

一方で，著作者の権利への配慮が必要なデータについては，AI 学習に供するデータの割合をコントロールするとともに，適切なリファレンスやメタデータを提示することが一つの解決策である．ウェブ上のニュース記事，雑誌記事の多くは，1/3 ～ 1/2 程度が無料で閲覧でき，記事全体を読むには有料プランに入る必要がある．そのような場合には，無料で閲読できる範囲のデータが AI 学習に利用できることが望ましい．

その上で，生成 AI の出力に対して，その根拠情報のリファレンス（もとのウェブページへのリンクなど）を提示すべきである．生成 AI の巨大なニューラルネットワークにおいて，出力の根拠をその生成過程から特定することは原理的に困難であるが，出力に類似した学習テキストを検索によって見つけ出すことは難しくない．類似度が最も高い幾つかのテキストをリファレンスとして表示すれば，ユーザがそのサイトに訪れることでウェブの広告モデル等とも整合し，それが興味を引く内容であればその書籍，雑誌，論文等の購入にもつながる．著作者からすれば，生成 AI 学習にコンテンツを提供することが宣伝にもなるのである．

著作権法の目的が「文化の発展に寄与すること」であることに立ち返り，著作権と生成 AI の関係を再考することが必要である．

3.3.4 生成 AI の開発体制における課題

これまでに議論した 3 つの課題とも関係して，生成 AI の開発体制における課題がある．高性能な超大規模 LLM の学習には膨大な計算コストがかかるため，一部の組織の寡占状態となっており，クローズドに研究開発が進められ，どのようなモデルをどのようなデータで学習したかなどの情報はオープンになっていない．このことが，安全性や著作権問題を含め，人々のさまざまな不安の元凶である．

最近では「オープン」を謳ったモデルも公開され始めているが，多くは限定的なオープンである．たとえば，多くの派生モデルの元となっている Llama 2，Llama 3 は，モデルの構成やパラメータはオープンであるが，学習コーパスは公開されていない．すべてが公開されているのは 100 億パラメータ程度の英語のモデルに限られてきた．

このような問題意識のもと，2023 年 5 月から，オープンかつ日本語に強い LLM の研究開発を目指す LLM-jp という活動を始めた [10]．モデル・データ・ツール・技術資料等を議論の過程・失敗を含めすべて公開することを基本的な考え方とし，この趣旨に賛同すれば誰でも参加可とした．当初は 30 名程度の自然言語処理の研究者の勉強会としてスタートしたが，HPC を含む様々な分野の研究者，産業界の AI 開発者・利用者，医師，法律家などが順次加わり，現在では約 2,000 名が参加している．

この活動をベースに，文部科学省において「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」（事業期間：2024 年度～ 2028 年度）が始まり，2024 年 4 月に国立情報学研究所に大規模言語モデル研究開発センター（LLMC）を設置した．また，LLM-jp および LLMC の成果として，学習データまで含めてすべてオープンな LLM としては世界最大規模の 172B パラメータのモデルを 2024 年 12 月に公開した．このモデルは日本語の標準的な評価プラットフォーム llm-leaderboard において GPT-3.5 を超える精度を達成す

るとともに，安全性の観点でも現在のトップクラスのLLM に匹敵する性能を有している．さらに，LLM の入出力に類似するテキストとそのメタデータ（ウェブコーパス由来の場合の URL 等）を表示する機能を備えている．今後も，モデルの高度化とともに，LLM の透明性・信頼性に関する研究開発を進めていく予定である．

4. おわりに

元国立国会図書館館長の長尾真は，「進歩発展」という概念が有限の地球上ではもはや成立しないと考え，経済的な豊かさよりも文化的・人間的豊かさこそが我々がこれから追求すべきことであると指摘し，「知識はわれらを豊かにする」というスローガンを掲げた [11]．そして，2008 年 4 月の日本出版学会講演において，国立国会図書館が収集・電子化した蔵書を公開し，最寄りの図書館へ行く交通費程度で借りられるようにし，その料金を出版社や著者に支払うという仕組み，いわゆる「長尾構想」を提唱した．この根底には，知識は万人の共有財産であるという思想がある．しかし，15 年以上たつ現在もこの構想は実現に至っていない．

学術分野における論文・データについては，DX 化の流れとともにオープンアクセスが実現されつつある．しかし，人類の知の集積は学術分野にとどまらず，はるかに大きなスケールで存在している．それらが活用されて始めて生成 AI は真に人類社会に大きく貢献するものになるだろう．たとえば，文化的，言語的な多様性の理解の上にたった従来とは比べものにならないレベルの自動翻訳を実現し，人々の異文化への関心を高め，理解を促進する．これは，現在の社会課題の元凶ともいえる国際社会の分断を解決する突破口となりえる．オープンなデータと知識が，人類と AI の共存社会を豊かにすることを一日も早く実現したいものである．

2025 年 1 月 22 日

注釈

1） 生成 AI の課題と波及効果，我が国として進める事項について，日本学術会議情報学委員会を中心に審議を行い，提言「生成 AI を受容・活用する社会の実現に向けて」を 2025 年 4 月に発出しているので参考にして頂きたい [12]．

参考文献

[1] 学術論文等の即時オープンアクセスの実現に向けた基本方針．https://www8.cao.go.jp/cstp/oa_240216.pdf（2025 年 1 月 21 日参照）
[2] AI 等の活用を推進する研究データエコシステム構築事業：トップページ．https://www.nii.ac.jp/creded/nii_ac_jp_creded.html（2025 年 1 月 21 日参照）
[3] EOSC: トップページ．https://eosc.eu/（2025 年 1 月 21 日参照）
[4] ChatGPT．https://openai.com/index/chatgpt/（2025 年 1 月 21 日参照）
[5] 降簱大介：大学 1 年生の数学（線形代数）は生成 AI に理解できるのか，第 84 回教育機関 DX シンポ，https://www.nii.ac.jp/event/upload/20250114-7_furihata.pdf
[6] 広島 AI プロセス．https://www.soumu.go.jp/hiroshimaaiprocess/（2025 年 1 月 21 日参照）
[7] Japan AISI：トップページ．https://aisi.go.jp/（2025 年 1 月 21 日参照）
[8] AI と著作権に関する考え方について．https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf（2025 年 1 月 21 日参照）
[9] POLITICO：フランスの戦略．https://www.politico.eu/article/la-bnf-prete-a-ouvrir-ses-archives-pour-franciser-des-modeles-dia/（2025 年 1 月 21 日参照）
[10] LLM 勉強会：トップページ．https://llm-jp.nii.ac.jp/（2025 年 1 月 21 日参照）
[11] 長尾真：『楽天知命』，イースト株式会社，2019.
[12] 提言「生成 AI を受容・活用する社会の実現に向けて」．https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-26-t381.pdf

【著者略歴】

黒橋 禎夫

国立情報学研究所所長／京都大学特定教授 .1994 年京都大学大学院工学研究科博士課程修了. 博士(工学). 2006 年 4 月より京都大学大学院情報学研究科教授．2023 年 4 月より同特定教授および国立情報学研究所長を併任．自然言語処理，知識情報処理の研究に従事．JST さきがけ「社会システムデザイン」研究総括（2016 ~ 2022），文部科学官（2020 ~ 2022）等を歴任．言語処理学会 10 周年記念論文賞，同 20 周年記念論文賞，第 8 回船井情報科学振興賞，2009 IBM Faculty Award，文部科学大臣表彰科学技術賞等を受賞．