URL https://journal.ntt.co.jp/article/30175
DOI https://doi.org/10.60249/24113101

NTTコムウェアのエバンジェリストの目に映る最新AI動向と技術開発――生成AIの里　番外地

NTTコムウェアのエンタープライズソリューション事業本部 データマネジメントソリューション部　第二ソリューションコンサル部門は，巨大なNTTグループの中で，おそらくもっともコンパクトな担当の１つと思われます．現在，生成AI（人工知能）の分野を中心に最大のコストパフォーマンスを発揮できることを目標にしています．私たちの取り組みの一部としてNTTコムウェアの公式ホームページにて，「生成AIの里」を連載中であり，今回はその番外編の紹介となります．

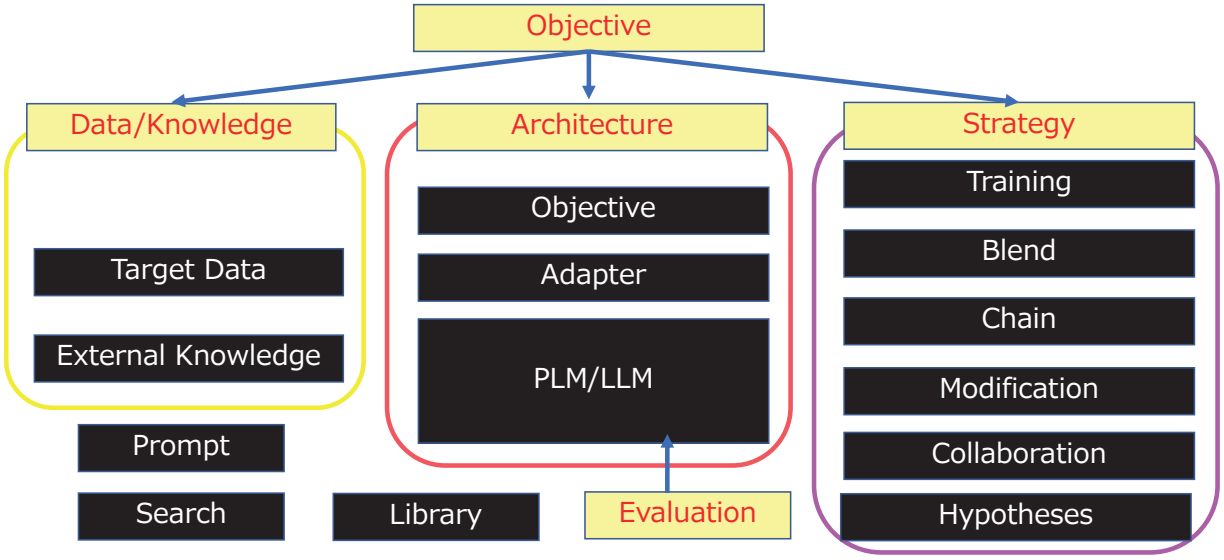
次世代生成AI技術の発展に対する研究・開発の位置付け

生成AI（人工知能）は大規模言語モデル（LLM：Large Language Model）だけではありませんし，最近突然出てきたわけではありません．情報検索やニューラルネットワークの分野の技術革新およびハードウェアの進化に伴い，一時期はビッグデータなどの看板を背負いながらも進化し続けた成果の１つです．**図１**に生成AIとその周辺技術を整理しています．生成AIは今でも進化を続けていて，情報検索と同じような道を駆け足でたどっていくと私たちは思っています．実際，両者はRAG（Retrieval-Augmented Generation）というフレームワークで遭遇しています．私たちは生成AIをAI，ML(Machine Learning)，DS(Data Science)などの分野から理論的解釈を行うとともに，新しい技術の研究および開発を進めています．

話題の分野であり，図１に示した周辺分野でも多くの研究者やエンジニアが日夜同じあるいは似たアイデアに基づいて，同じ目的に向かって進んでいます．したがって，後追いにならないように，少しでも先に課題を発見することが必要になります．事業会社として，研究対象は将来，顕在化する社会課題の解決につながると期待され，投資効果が説明できるものに注力しています．

これらの研究で，私たちは主に表現学習とalignment（モデルから人が期待する挙動や効果を得るためにモデル内部のデータ表現の対応関係を修正すること）に着目しています．表現学習とは生成AIが対象とするテキストや画像などのデータを理解可能な表現とする学習で，一般的にこれらのデータをトークン単位に分割したときのembedding（言語や画像，音声データを数値のベクトル形式の表現）の学習に相当します．alignmentはこれら表現学習に用いられるプロセスです．

PLM(Pre-trained Language Model)やLLMも言語モデルの総称で，生成AIの中心に位置付けられ，人気を博した理由の１つに事前学習によりタスクへの適用が簡単であるということがあります．自然言語処理に取り組んだことのある方はご存じかと思いますが，処理対象となるテキストデータの収集や整形はその後の処理よりも大変なことがあります．さらにそれらのデータを用いて言語モデルを学習するとなると，今度はプログラムにも一工夫必要になります．PLMはこれらの前処理から私たちを解放するだけでなく，タスク適用へのハードルも下げました．LLMもPLMの１つであ



出典：「生成AIの里 第二回：生成モデルと検索モデル（前編）：Prêt-à-porter ou haute couture?」

図１　生成AIとその周辺技術

るのですが，Promptによる創発性を発揮する点が特徴です．名前のとおり巨大な言語モデルなので，embeddingのベクトルサイズも大きくなったこともありますが，それらは学習データが増加しても，alignmentによる結果を保持できるだけでなくトークンを超えinput-outputというより長い単位のalignmentの実行とその結果であるembeddingに反映できるようになったと解釈でき，分散表現の性能を高めていると考えられます．その結果，言語モデルはinstruction tuningや強化学習（RLHF（Reinforcement Learning from Human Feedback）等）による訓練を反映し，next token prediction以外のタスクでもcognitiveを示すようになりました．そのため，以前のモデルと区別してinstruction-followingなモデルという呼ばれ方もします．

生成AIのパフォーマンスを高める基礎研究への取り組み

すでに多くの優れたLLMが発表されており，今後も新たに登場するでしょう．そのため，ゼロからLLMを開発あるいはトレーニングするだけでなく，既存のLLMを特定のドメインやサービスに適用するために追加学習（ファインチューニング）を行うことも可能です．ファインチューニングでLLMの全パラメータを更新すると，LLMが元々持っていた有用な知識，特に内部ネットワークの重みやトークンの分散表現が上書きされてしまう可能性があります．そこで最近では，PEFT（Parameter-Efficient Fine Tuning）というファインチューニング手法の開発が活発に進められています．

■PEFT

LLMは巨大なネットワークであり，複数のネットワークの集合体とも見做せます．LLMが生成に必要な知識を持っている場合，それは宝くじ仮説やMixture of Expertsなどに基づき，LLM内部に，より小規模かつ知識に特化したサブネットワークが存在するとの仮定の下でモデル化できます．私たちのPEFTでは，これらのサブネットワークを発見し，それらの優先順位を動的に変化させ，パラメータの更新を最小限に抑えつつ，ネットワーク内の知識構造を動的に変化させる手法を検証しています．

■Prompt/RAG

PromptとはLLMなどの生成AIに対する指示のことですが，ここではPromptのチューニングやoptimizationも含みます．LLM内部に生成に必要な知識が十分でない場合，Promptを用いて外部から知識を補完することがあります．しかし，それを手動で行うのは大変で非効率的です．

そこで，外部データベースから必要な知識が含まれていそうなファイルを検索し，LLMに与えるアプローチがRAGです．

洗練されたPromptやその変更履歴は，自分だけでなく同じ目的を持つ他の人々にとっても貴重な情報です．私たちはこれらの貴重な情報を有効活用するために，Promptのチューニングや最適化，RAGのパーソナライズや協調フィルタリングを通じて，ユーザ間でPrompt（以降，RAGも含めて）の共有と利用を支援する研究を行っています．

LLMは数十億のパラメータを持つネットワークであり，その巨大さが驚異的な性能をもたらす一方，学習や運用コストも増大させています．学習はともかく，LLMを凍結して推論だけを実施するならばこれほど巨大なネットワークは必要ないのかもしれません．そのため，蒸留，刈込み，量子化などのテクニックがあります．

今後は，大多数のLLMが採用しているTransformerの構造を部分的に数世代前のネットワーク構造であるMLPにより代替するかもしれませんし，ネットワークをユークリッド空間でなく双曲空間にすることで，パラメータ当りの情報表現が豊かになり，ネットワークのサイズをコンパクトにすることが主流になる可能性もあります．またembeddingだけで十分なこともあるでしょう．そのような将来の技術動向や可能性を予測しながら基礎的な研究も進めています．

マルチモーダルAIへの進化と応用研究への取り組み

■マルチモーダル

生成AIにおけるマルチモーダルとは，言語，画像，音声，系列データなど，異なる種類のデータを同時に理解することを指します．LLMは言語のみを扱うため，シングルモーダルのモデルです．しかし，生成AIでのマルチモーダルは，人間が利用する際の解釈（可読）や利便性の向上，そして大量のデータから学習するために，内部でLLMを利用するモデル（VLM：Vision Lange Model：視覚言語モデル）が存在します．LLMが画像を言語と同時に扱うVLMとなるためには，embeddingの学習とalignmentの処理が重要です．LLMは言語に対応するエンコーダを持っているため，VLMはLLMと画像に対応したエンコーダを併用することで，画像も言語と同様にトークンという単位に分割し，そのembeddingを獲得できます．

しかし，このままでは画像から得られたembeddingは言語の分散表現と数値ベクトルという形状は同じでも，対応が取れていません．例えば，りんごの画像トークンの分散表現と「りんご」という単語のトークンのembeddingの距離が，他の単語のトークンのembeddingとの距離よりも近くなるようにはなっていません．

そこでalignmentの処理が必要となります．多くの場合，LLMのパラメータは言語のembeddingを含めて固定されています．したがって，**図2**に示すようにVLMの学習，ファインチューニングではalignmentにより画像から得られるembeddingと，言語から得られるembeddingの相対的な距離を近づけるために，画像をembeddingの空間に射影するネットワーク，例えば，エンコーダからの出力をembeddingに変換するネットワーク，そ

の重みを更新します．その結果，VLMは画像も言語もトークン単位で意味の対応が取れた分散表現を獲得できます．これでVLMに画像を入力すると，その画像についての説明（キャプション）を出力するタスクが実行できます．これはPLMの持つ「入力したテキストに後続するテキストを出力する」というタスク，Next token predictionにおける入力を画像にしたタスクと考えられ，PLMの応用としても存在していました．LLMになり「Promptによるcognitiveな性能の抽出」という性能が向上すると，VLMにこれらのLLMを採用すれば，VQA（Visual Question Answering）などのタスクも実行できます．

このようなcognitiveな性能をVLMで発揮させるために，VLMのファインチューニングにinstruction tuningとそれに対応した学習用のデータを採用することで，alignmentがトークンレベルから入出力のレベルになり，それに伴いembeddingも更新されます．その結果，VLMに画像とそれに関する質問，例えば「画像に写っている場所はどこですか？」を与えると，「●●です」と出力できるようになります．

私たちは前述のPEFTやPrompt/RAGで得られた知見をベースに，VLMの性能向上をめざしています．具体的にはモーダル間の知識のalignmentをトークンや入出力単位で実行することで，分散表現を差分更新し，VLM内部のLLMに対するPromptの解釈の性能向上について検証しています．

マルチモーダルは画像とテキストだけに限定されるものではありません．私たちは非構造・構造データとテキストの関係もマルチモーダルにおける関係と解釈し，入力データに対し分析や解釈を行うデータ分析ツールとしての応用についても検証中です．

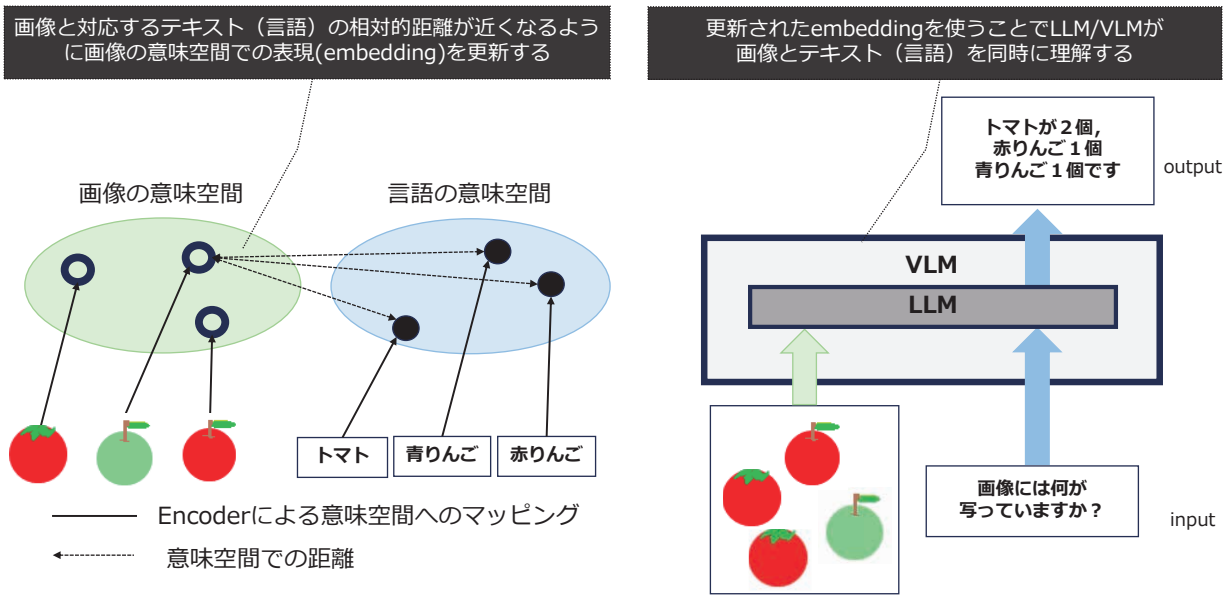
■パーソナライズ・レコメンド

ここでパーソナライズとは，生成AIの出力を利用者の好みやニーズに合わせるために，PromptやPEFTの最適化を含むものです．パーソナライズは，以前からあるレコメンド技術やサービスにも導入されています．生成AI，特にLLMを取り入れることで，レコメンドはコンテンツ理解だけでなく，コンテキスト理解の性能も向上します．その結果，Knowledge Graphなどのテキストデータやユーザレビュー，コメントなどのコンテンツから，アイテムの特性やユーザの嗜好，ニーズ，およびそれらの関係を把握できるようになります．これにより，データのスパース性やCold start problemなどの課題解決だけでなく，ドメイン横断のレコメンドも可能になります．

さらに，レコメンドにおいて商品説明や推薦理由などのテキスト自動生成や，ユーザとの対話によるインタラクティブな支援などの機能も強化され，ユーザ体験やロイヤリティの向上も期待されています．

ここでも，前述のPromptやPEFTが重要な役割を果たし，パーソナライズに向けたLLMのPromptデザインやPEFTと組み合わせた最適化の検証を行っています．具体的には，ユーザの嗜好やニーズに合わせたテキストや対話生成をめざしています．

さらに，マルチモーダルと組み合わせたMRS（Multimodal Recommender Systems）も生成AIのレコメンドへの応用です．MRSは，複数のモダリティを統合し，従来の課題であるデータのスパース性やCold start problemを解決しつつ，ユーザやアイテムの表現のalignmentを強化します．具体的には，例えばアイテムのテキストに対応する画像を学習データとして与えた場合，



(a) マルチモーダルにおける意味空間の更新　　(b) マルチモーダルにおけるVQAの概要

図２　ファインチューニングにより更新されたembeddingによりLLM/VLMが分散表現を獲得するイメージ

MRSはそれらの関連を反映してそれぞれの表現についてalignmentを通じて学習します．学習した結果，入力がテキストあるいは画像のみであっても，MRSはその内容を理解して対応します．

このようにマルチモーダルによりレコメンドは入力データの理解を深め，機能の性能を向上させることができます．MRSでもPromptやRAGやPEFTが大活躍するのはいうまでもありません．

一方，マルチモーダル同様，モダリティ間のギャップの課題もあり，その課題解決に向けてマルチモーダルとMRSの研究も同時並行で進めています．

システム実装から考える生成AIの将来展望

事業会社では生成AIについて，理論だけでなく，そのシステム実装まで検討する必要があります．私たちはオープンソースをベースに，生成AIを構成するモデルの訓練や適用するドメインやタスクに向けての最適化のみならず，それらの運用を低コストで実現するシステム構成や必要なライブラリも調査検証をしています．LLMでも複数のオープンソースを選択できるようにHugging faceだけでなく，最近はvLLMなども利用しています．

また複数のLLMを組み合わせるBlendについても調査中です．「蒸留」，「量子化」，「刈り込み」といった軽量化だけでなく，ライブラリとしてdeep speed flash attentionなどの利用にも取り組んでいます．最近のLLMの学習には強化学習やインストラクションチューニング用のデータの準備も欠かせないので，幅広くライブラリやデータセットの調査も継続しています．技術の新陳代謝が速いので，「飛鳥尽良弓蔵狡兎死走狗烹*」を痛感せずにはいられません．次に紹介する学会活動に参加して痛感するのは「優秀な研究者は同時に優秀なエンジニア」であるということです．これは生成AIに限った話でなく，ビッグデータが話題となっていたときから同様です．サーチエンジンが検索だけでなくブックマーク代わりに使われることが多いように，生成AIも想定外の使われ方をするかもしれません．気が付いたら四面楚歌になっていないように，まずはシステムに通じたエンジニアをめざして研究，調査，検証等に取り組んでいます．

学会参加や大学講演を通じた外部向け活動と展開

■学会

調査，研究にあたっては社外のさまざまな団体，特にアカデミックやOSS（Open Source Software）などのコミュニティとも連携しています．具体的には国際会議により名称が異なりますが，

表　主要国際会議への参画状況

会議名称	2021	2022	2023	2024	2025
ICLR		→	→	→	→
ICML		→	→	→	
NeurIPS	→	→	→	→	
KDD	→	→	→	→	→
AAAI				→	→
WSDM				→	→

ICLR: International Conference on Learning Representations
ICML: International Conference on Machine Learning
NeurIPS: Neural Information Processing Systems
KDD: Knowledge Discovery and Data Mining
AAAI: The Association for the Advancement of Artificial Intelligence
WSDM: Web Search and Data Mining

表のようにProgram Committee（PC）やReviewerなどで貢献しています．

会議等においては「生成AIに査読してもらったほうがましだったよ」と言われないような査読を心掛けています．またACL2023で，KDD2024といった国際会議にも参加し，その報告を外部に公開しています．

■大学

大学で生成AIに関する講義の機会をつくっています．2023年は，九州大学にて「生成AIの数理」というテーマで講演を行い，上智大学では非常勤講師として，「自然言語処理と言語モデル」の連続講座を開始し，2024年度も継続しています．講義は，自然言語処理やLLMを重要度に応じ解説および演習するだけでなく，受講生のリクエストに応じて最新の技術やトレンドも解説しています．インタラクティブな講義になるように，受講生の質問に答える時間を必ず確保し，こちらも「生成AIに教えてもらったほうがましだったよ」と言われないように，受講生の皆様の理解のお手伝いができればという思いで取り組んでいます．

今後はイベント等で，生成AIのデモ展示などで私たちの取り組みを紹介する企画も検討していきます．

◆問い合わせ先

NTTコムウェア
エンタープライズソリューション事業本部
データマネジメントソリューション部
E-mail　es-bz-promotion@srv.cc.nttcom.co.jp

* 飛鳥尽良弓蔵狡兎死走狗烹：とらえる鳥がいなくなると良い弓も不要となり使われなくなる旨の故事成語．用のあるときは使われるが用がなくなると使われなくなることの意．