めもおきば TechReport 2024.12

めもおきば
Aki @ nekoruri

AI とどう付き合っていくか－2024

ChatGPT(GPT-3.5)や Stable Diffusion が 2022 年に登場して以来、生成型 AI(Generative AI)、とりわけ大規模言語モデル(LLM: Large Language Model)と呼ばれる AI 技術が、社会の様々な場面で活躍するようになりました。

この記事では、具体的なツールの使い方はあえて触れず、幅広く AI 関係の技術と付き合っていくための考え方の手がかりを紹介します。

LLM の課題

LLM が大きな価値を生み出すようになり、いくつか課題がはっきりしてきました。

1. LLM は自分が計算機であることを忘れている
2. LLM は文脈を知らない
3. どのようなデータや文脈(コンテキスト)を与えるか
4. 法律やビジネスロジックなどのルールに従わせることが容易ではない

これらの課題を通じながら、LLM を中心とした AI 技術のすこし未来を創造する手助けとなれば幸いです。

LLM は自分が計算機であることを忘れている

LLM は大規模言語モデルと呼ばれるように、様々な言語知識を元に「学習」が行われ、その学習結果をもとに言語を使って「推論」を行います。本来 LLM を動かしているコンピューターも計算機であるはずですが、推論を行うプログラムでは特定の処理だけを「計算機の生の機能」を使った計算などをするようには実装されていないようです。そのため、「自分が計算機であることを忘れている」ように、数値計算が下手であったり、具体的なルールに厳格に従わせることが容易ではありません。

それを実現するためにこの数年で様々なテクニックが発明されてきていますが、結局のところ、LLM という基盤モデルの上で、どうやってそのずっと下にある「計算

機」の機能を直接使わせるかは、人間にも無い機能なので新しく考えていかなければならないでしょう。

よく言われるのが、「9.11 と 9.9 のどちらが大きいか」という問題に間違えてしまうというものです。Chain-of-Thought（CoT）などの手法でより深く考察させることによって正しい結果を返せる場合もありますが、そもそもの問題は、計算機がそれ自身の計算機を直接使った計算をしていないことです。これはけいさんだけでなく、ルールベースの〇×判定のようなものも、それっぽい文章の続きを考えるという現在の LLM においては、「深く考えすぎてしまう」という癖があります。

様々な手法で LLM の形を保ったまま解決しようとしています。たとえば LLM-as-a-Judge（生成 AI による自動評価）と呼ばれるテクニックでは、LLM の出力を、評価項目をプロンプロに仕込んだ別の LLM によって評価させるというものです。自動化が可能な最初のフィルターとして LLM-as-a-Judge を利用し、さらに厳格さが求められるユースケースであれば、人間による妥当性などの評価をはさむ Human-in-the-Loop という手法もあります。そしてそれらによる評価結果をもとに、LLM 自信を再学習させることで、より高精度な回答を得られるようにします。

また異なる手法としては、計算ができるプログラムを LLM の「外」に用意し、Function-Calling という手法で、LLM が計算プログラムへの入力までを考え、計算はその計算プログラムに任せる、というものです。

いずれにせよ、ポンと LLM のチャットボックスに投げるだけで解決できない問題はたくさんあり、それらを補助するようにワークフローエンジンや AI エージェントが登場しました。

LLM は文脈を知らない

LLM は膨大な知識によって学習していますが、その一方で実際に何かの仕事をするときに前提となる文脈情報をそのままでは持っていません。文脈情報とは、例えば

組織の中にある仕様書や手順書であったり、所属組織のビジネスモデル、AIを利用する人間が実現したい作業の目的や現在状況などです。

そういった文脈情報を収集し、ファインチューニングや、プロンプトに含める In-Context Learning などによってAIに伝えることで、より状況に適切な判断を出力してもらいます。

まず身近なところから、自分のコンテキストを AI のメモリー（記憶）として食わせていくために様々な試みが始められています。GitHub Copilot なんかは使ったことがある方も多いかも知れませんが、既存ソースコードを Copilot 君に文脈情報として食わせることで、適切なソースコード案を出してくれるようになります。

この分野は今まさに PC やスマホ業界で盛り上がり始めたところです。2024 年に発表された Apple Intelligence では、iPhone 上で行われた様々な行動データを元に、たとえば自分にとって重要な通知を優先して表示するなど、少しずつ利用者を手助けするような仕組みがはじまります。デバイス内部（オンデバイス）での処理を優先しつつクラウドを併用することでプライバシーに配慮しています。

Microsoft は Windows Recall として、デスクトップ画面に写っているものから何をしているのかを把握・記憶し、例えば「この前コピーしたけど保存先を忘れてしまったファイルの行き先」などを教えてくれそうです。ハードウェアベンダーでもある Apple と異なり、Microsoft は「すこし未来」のハードウェア環境を想定して機能開発を進めているため、現状においてはハードウェア要件（特に TPU の 40TOPS 以上）が高い割りにできないことも多く、伸びしろに期待がされている状態です。他にも、同社の LLM である Copilot 機能を Microsoft 365 の基本機能に含め、編集中のファイルを LLM に食べさせやすくする環境を全力で舗装しています。

Google は、Android や Chrome、Google Workspaces といった様々な環境に横断して Gemini 統合を進めています。個別の動きは上二社と似たような感じですが、強力なシェアを持つ Chrome の上で動く Project Mariner や、その他の拡張機能の存在感は大きいです。

この 2 年ほどで、RAG(Retrieval Augmented Generation)と呼ばれる、質問プロンプトに関係がありそうな知識を埋め込んで、LLM の回答に活用してもらう手法が当たり前になりました。

最新のニュースや特化ドメインの文献など、LLM の学習時点ではカバーされていない情報を動的に取得できます。最新の法改正情報や、最新のニュースや論文のサマリなど、簡単に使い始められる環境が整っています。

その一方で、「素朴な RAG」の厳しさも浮き彫りとなりました。今あるドキュメントを LLM によって成分に分解し、ベクトルデータベースを使って質問プロンプトに近い内容の文書を探し、質問プロンプトに付与して LLM に問いかけます。

最大の問題は、RAG の心臓部である「検索」というのが難しい技術的課題だということです。そもそもベクトルデータベースを使う場合にも、文書を一定のサイズに区切ってチャンク化するときのやり方や、ベクトル成分への分解方法、検索結果のランキングなど様々な技術的課題があります。

より範囲を拡げると、そもそもの文脈情報の知識をベクトルではなく知識グラフとして表現することで、ベクトルの近接よりさらに高度な検索を試みる GraphRAG なども挑戦が行われています。この領域には大きな期待がかかっていますが、知識を保存するグラフデータベース上におけるモデリングの作成ノウハウなど、専門家が必要とされる一方で彼らの数はまだ多くないでしょう。

別な方向性として、ロングコンテキストと呼ばれる、質問プロンプトに長いメッセージを直接投げられる LLM では、RAG に頼らずプロンプトに全部乗せてしまうという力業も可能です。 本一冊を丸投げしてもそれなりに動くと言うところまで来ていますが、ある意味コストの浪費とも言えるため、やはり RAG 系の技術を中心に、In-Context Learning としてプロンプトで検索結果を追加情報として与えるモデルはしばらくは主流だろうと思います。

こういった取り組みの少し先の方では、自分のコンテキストをすべて食わせた AI を自分の分身として個人でも企業でも活動する、BYOD ならぬ「BYOAI（Brind Your Own AI）」という時代がくるのかもしれません。

ビジネス AI

先進的な取り組みに寛容な企業や組織では、すでに AI を通常業務の一部としてどんどん取り入れています。お客さんに対するサービスに AI を組み込むのとはまた別に、社内の業務でも活躍します。

- 営業支援：メッセージ生成、案件掘り起こしをより賢く、短時間で大量に丁寧に
- 人では難しかった粒度での管理
- 文脈を踏まえたインサイト発見、不正検知
- 新入社員のオンボーディング（研修）
- レガシー機器の DX 化、デジタルツイン化

たとえば、AI カメラで異常を検知するには、これまでは機械学習やルールベースでいろいろ頑張る必要がありましたが、LLM であれば一定の精度でいきなり異常を検知することができるかも知れません。新入社員の質問先や壁打ち相手として、いくら質問を重ねても怒らない「仲間」には価値があります。

例えばソラコム社では、生成 AI を利用した IoT アプリケーションをローコードで開発する「SORACOM Flux」を提供しています。SORACOM Flux ではカメラ画像に対して自然言語でルールを記述し、別のシステム（例えば警告灯など）と連携できます。

サービスに AI を組み込む例としては、マルチモーダルを利用した AI 相手の英会話教室や個別指導塾といった教育が分かりやすい例ですが、古くからの「推薦アルゴリズム」といった様々な機械学習アルゴリズムにも LLM 的要素が組み込まれていくことでしょう。

デジタルツイン

実世界の話がでてきたところですが、実世界の情報をデジタル上でリアルタイムに再現し、シミュレーションや分析に活かすのがデジタルツインです。細かい 3D の「点群」から、GPS で取得したトラックのおおまかな位置など再現の粒度はさまざまです。

まずは SORACOM Flux のようにルールベースで記述しているのは、現時点ではビジネスロジックを実装するのにそれが一番早いからです。その一方で、製造業や自動運転などの領域では点群データなど細かい粒度のデジタルツインを構築し、日世界で起きることのシミュレーションを試みています。

「知識ベースを増やすほど精度が上がる」AI 的なアプローチについてここまで話してきました。その一方で実世界では「物理法則を厳密に再現しようとすると無限にやることが増えてしまうので、どこかで割り切って近似する」というシミュレーション的アプローチが必要です。店頭に並んだキャベツの数を数えるのに素粒子レベルのシミュレーションは不要です。

大量の実データを取り込みながらも、計算コストと計算精度のトレードオフをどのように制御するか。さらに、自動運転や危険予測では社会インフラや人間の生命に直接関わるため、法規制やガイドラインとの整合性、サイバーセキュリティ、責任分担（保険の適用範囲など）といった広範な問題を考慮しなければなりません。

AI SaaS

自社の情報を文脈にするかわりに、お客さんの文脈情報を食わせることでさらなる発展が見込まれるのが AI SaaS 領域です。

現状の LLM では、ルールやコンプライアンスをはじめとしたドメインロジックを強制することが難しいので、ワークフロー上で様々なドメインロジックと LLM を組み合わせる事例が増えています。他にも、そのサービスの方向性や、重要視したいお客様理解の方向性、価値観など、そういったところが差別化ポイントになりそうです。

この後紹介する AI エージェントの話にもつながりますが、LLM にルールやワークフローを書かせて、それを LLM や人間に検証させることで、少しずつ複雑なドメインロジックを LLM に任せることが進んでいくでしょう。これは、優れたエンジニアがシステムの標準化・自動化を進め、最後には自分自身を不要にするところまで綺麗に仕上げていく、というのに似ていますね。

AI エージェント

2025 年はすでに「AI エージェントの年」と言われていますが、まさに LLM 領域でもっとも賑やかであることは間違いありません。

従来のワークフローエンジンが担当してきたタスク管理・オーケストレーションに「推論（reasoning）」の要素を加えることで、より高次の自律的行動が可能になりつつあります。たとえば企業の基幹システムや Web サービスに LLM が組み込まれ、ユーザーの入力内容を解析しながら、外部 API や内部データベースを駆使して高度なアクションを自動実行する例も増えています。

こうした AI エージェントが効率的に機能するためには、基盤となるナレッジベースやデータモデルの整理と再構築が欠かせません。RAG では外部のデータソースを参照して LLM の知識を補完しますが、その精度を高めるためには、情報資産を体系的に管理するデータカタログやメタデータ管理が重要です。セマンティックウェブ的な標準化や、グラフデータベース的なデータ間の関係性など、データ同士を関連づけておくことで、AI エージェントが必要な情報を漏れなく取り込める仕組みが整います。そういったことをやるには、組織が持つ広範な情報へのアクセスが必要で、逆に言えば、プライバシーやコンプライアンスに配慮したデータガバナンスの上に構築する必要があります。

現状の LLM だけでは「論理的な推論」や「厳密な計算」「ドメイン知識に基づくルール順守」といった部分が不十分です。すでに述べたとおり、LLM は自分が計算機であることを忘れ、プログラミングの変数や時間軸を追跡する「プログラムカウンタ」の概念が無いため、複雑な状態管理が苦手です。この弱点を補う動きとして、Dify など

のワークフローツールや外部推論エンジンとの連携が活発です。特定のビジネスルールを明示的に設定することで、LLM に対して「このルールは絶対に破ってはいけない」というガードレールを敷きつつ、必要に応じて厳密な計算処理や状態管理を他のシステムが担当する仕組みが検討されています。

また、AI エージェントが実世界と結びつく「デジタルツイン」の動向にも注目が集まります。NVIDIA の Cosmos のように、実環境の情報をリアルタイムで取り込み、仮想空間上でシミュレーションしながら制御を行う技術が進歩しつつあります。工場や都市などの現場データをつなぎ込むことで、AI エージェントが実際のロボットや生産設備に「手足」を与えられ、物理世界とのインタラクションがよりスムーズになります。これにより、自動化や効率化だけでなく、未然にリスクを察知して対策を打つなど、高度な意思決定が期待されます。

外部のサービスとの連携では、ChatGPT GPTs や Anthropic の computer use、ブラウザ操作を行う browser-use など、多様なプラグインや API コール手法が模索されています。ウェブブラウザを通じて他のサービスへアクセスしたり、クラウド環境のファイルシステムを操作したりと、AI エージェントが「自律的に使えるツー」が拡張されているわけです。

Anthropic が提案する MCP（Model Context Protocol）では、LLM から叩きやすいシンプルな RPC インターフェースを定め、様々な外部サービスとの連携ができる枠組みを作りました。awesome-mcp-servers[1]という GitHub リポジトリでは、様々なサービスなどと連携する MCP サーバーが紹介されています。

しかし、なんでもできるようになる以上、権限管理のあり方が大きな課題です。「端末ユーザーができる操作を、そのまま AI にフルアクセスさせてよいのか」という疑問はセキュリティ上見逃せません。各プラグインやエージェントには、明確な認可（OAuth など）や操作ログの記録、一定の利用制限が必須となるでしょう。

[1] https://github.com/punkpeye/awesome-mcp-servers

このように、AI エージェントは単に会話を生成するだけでなく、外部の知識基盤やサービスと連携し、場合によっては物理世界まで踏み込む「総合的な知的エージェント」へと進化しています。一方で、ビジネスロジックやセキュリティ、ドメイン知識の厳守、そしてプライバシー保護など、解決すべき課題も山積みです。今後はデジタルツインの高度化とともに、AI エージェントがより柔軟かつ安全に「リアル世界」へ作用するための基盤づくりが一層求められるでしょう。

AI 駆動開発

あまり具体的な話を書くとすぐに状況が変わってしまうので今まさに最も書きづらい新しい領域が AI 駆動開発です。LLM を強く利用してソフトウェアの開発をするのは、2021 年の GitHub Copilot からです。現在は単なるコード補完の候補を出すだけでなく、チャットでコード変更の指示を出したり、全体のアーキテクチャの壁打ちをしたり、そもそも自然言語で渡した仕様を実装してもらったりと、できることが毎週ペースで大幅に増えています。

さらに、「同僚」もしくは「後輩」のように AI の持つ文脈情報を育てていく Cline とのペアプログラミングのような形もだいぶ現実化してきました。みんな誰もが自分の Cline を育成している時代がすぐに来るでしょう。

その一方で、大規模なソフトウェアリポジトリをまるまる全部 LLM に食わせるにはまだ課題が大きく、Cursor など業界を牽引しているサービスには、そのあたりをうまく RAG などに落とし込むようなノウハウがあるのでしょう。

AI 常時接続時代の到来

私たちの社会は今、24 時間絶えず AI と接続できる「AI 常時接続時代」へと近づいています。いつでもどこでも“電脳化”された知能との対話・連携が可能になってきました。こうした変化は私たちの生活スタイルだけでなく、ビジネスや社会インフラのあり方にも大きなインパクトを与えています。

このように、人間と AI が常時接続され、経験を共有し合う時代には、ユーザーや機器、そしてリアルワールドからのフィードバックが断続的に AI へ流れ込み、AI 自身がリアルタイムで学習やアップデートを行うことが可能になります。AI もまた「経験学習モデル」のように経験から成長する可能性が高まります。

ユーザーの行動や嗜好を AI が学び続けることで、より「その人らしい」応答や提案を返せるようになるでしょう。この流れの中で、「マイ Gemini」という個人専用の AI エージェントを所有するコンセプトが注目されています。

これは、ユーザーごとのニーズや履歴を深く学習し、最適化された回答・サポートを行うだけでなく、ユーザーの同意のもとでプライベートなデータや外部サービスへも直接アクセスできるものです。まさに「自分専用の AI パートナー」として機能することで、生産性やクリエイティビティを大きく飛躍させます。

一方で、こうした AI がアクセスし得る情報量や権限が拡大するにつれ、データガバナンスやプライバシー、セキュリティ管理がこれまで以上に重要になります。企業のシステムに接続する場合には、厳密な認証・認可の仕組みや操作ログの管理が必要となるでしょう。人間であれば慣習的に見えていても見えなかったことにするような情報はまれによくありますが、AI に同じ事を強いるのは難しいかも知れません。

AI 常時接続時代には、個人をそのまま模倣する「デジタルクローン」と、物理空間を仮想空間に再現する「デジタルツイン」の概念がさらに進化していきます。

個人が発信するメッセージや行動履歴、さらには思考プロセスを膨大な文脈情報として学習することで、その人の「分身」として振る舞えるような AI がデジタルクローンとして活用されていくでしょう。だいぶ遠く見えますが、たとえば営業メールの下書きを LLM に書かせているような、さらにその先に続いています。

製造業や都市開発、物流などで、現実世界の構造をデジタル空間上に忠実に再現し、シミュレーションや予測、最適化に利用するデジタルツイン技術が、AI 常時接続環境と組み合わさることで、リアルタイムに変化する物理世界のデータを即座に反映

し、工場のライン制御や交通渋滞の緩和、災害対策まで幅広く応用できるようになります。

将来は、個人や組織が必要とする機能・サービスを AI エージェントが自律的に見つけ出し、最適なワークフローを組み立てるようになるでしょう。データ基盤を整備し、権限管理やプライバシー保護を徹底すれば、誰もが手軽に「電脳化」された知能をすぐ横に置いて活用できる時代が来ます。

同時に、技術的・倫理的なハードルも高くなります。常時接続によるプライベート空間の侵食や、AI が果たすべき責任範囲など、新たな社会的ルールづくりが急務です。デジタルクローンやデジタルツインが広範囲に普及する中、私たちは AI が生成する膨大な情報をどう選別し、どのような意思決定に役立てるか、組織やコミュニティ全体で考え続ける必要があります。

その先には、人間と AI が高度に融合し、より創造的で効率的な社会が待っているかもしれません。常時接続時代の AI は、私たち一人ひとりにとって新たな「知的空間」を切り拓きつつあるのです。

AI と電力　クラウド編

ここまで夢の話を色々してきましたが、ここからもう少し現実的な AI を支える技術について触れていきます。

LLM の高度化にともない、計算機資源の需要が急激に拡大しています。その結果、GPU を中心とした専用チップの TDP（Thermal Design Power）――すなわち消費電力と発熱量の増大傾向が著しく、これを支えるクラウド環境やデータセンターの整備が急務となっています。

かつての大規模分散処理（Hadoop など）では、ラック全体を広く使う「低密度・多数ノード構成」が主流でしたが、AI トレーニングにおいては GPU 間のインターコネクト速度やメモリ帯域が極めて重要になったため、「高密度で GPU を詰め込み、NVLink などの高速な接続で性能を最大化する」形態が求められています。ところ

が、高性能 GPU を高密度で設置すれば、消費電力と発熱量、そしてその熱密度が膨れ上がり、通常の空冷では冷却能力が追いつかないという課題に直面します。

こうしたニーズに対応するため、NVIDIA は「DGX-Ready Data Center プログラム」を展開し、DGX H100 やその後継 GPU などを運用できる十分な電力量・冷却能力を備えたコロケーション（データセンター）を認定・紹介しています。日本国内でも 2024 年 10 月時点で複数の施設が整備されており、PFN（Preferred Networks）やチューリング（Turing）など AI ベンチャーや研究機関が高性能 GPU リソースを積極的に活用している事例が報じられています。特に次世代の NVIDIA GB200 世代を 72 基搭載した「NVL72」構成では、1 ラックで 480V 入力かつ直接液冷（DLC：Direct Liquid Cooling）を前提とする設計が求められ、データセンターの基盤から変えていかないと物理的にも運用が難しい状況です。

冷却技術についても大きな転換が進んでいます。空冷ではラックあたり数十 kW が限界とされてきましたが、液冷、とくに冷却水を直接サーバー内部のプレートに流す DLC 方式への対応が、最新のハイエンド GPU を運用する現場で続々と導入されています。

GPU 同士をつなぐ NVLink ケーブルや高帯域メモリとの近傍性を高めるには、狭い空間に高性能チップをギュッと詰め込む必要があり、その結果、熱密度も跳ね上がるからです。すでに欧米の大規模クラウド事業者や日本の主要データセンター事業者も、液冷対応設備やグリーン電力の確保を急ピッチで進めています。

こうした背景には、環境面の要請も大きく関わっています。カーボンニュートラルが世界的な目標となる中、膨大な電力を要する AI トレーニングを「低炭素」かつ効率良く実施することは喫緊の課題です。総発電量に限りがある以上、従来のように電力を潤沢に使うわけにはいかず、高効率冷却や再生可能エネルギー活用の取り組みを強化しなければ、まもなく需要に供給が追いつかなくなる可能性があります。また、クラウド利用者にとっても、GPU インスタンスの大規模利用は大きなコスト負担につながるため、FinOps（クラウド費用の最適化管理）を徹底するうえでも「省電力で高い演算性能」を目指すことは非常に重要です。

クラウドの世界ではかねてより「統計多重効果」がサービス効率を高める基本原則でした。複数ユーザーがリソースをシェアすることで、ピーク時とアイドル時の無駄を補い合う仕組みです。AI トレーニングにおいても、GPU リソースの需要タイミングが異なる組織同士でうまくシェアできれば、電力ピークを分散して環境負荷を下げつつコストも削減できるでしょう。将来的には、グローバルなタイムゾーンの差を利用して AI ジョブを動かす地域を動的に切り替えたり、再生可能エネルギーの発電量が多い時間帯を狙ったりといったシナリオも考えられます。

AI 分野でさらなるイノベーションが起こるには、高性能な GPU や AI チップが必要不可欠です。一方、その実装・運用環境であるデータセンター側も、電力・冷却技術・高密度設計へのシフトを進めなければ「発展のボトルネック」になりかねません。ハードウェアとインフラがかみ合うことでようやく実現しつつある「超高密度・超高性能」の AI トレーニング環境は、クラウド事業者と利用者の双方に大きなメリットをもたらすと同時に、エネルギー効率やカーボンフットプリントへの配慮が欠かせない時代へと突入しているのです。

AI と電力　デバイス編

AI 技術を前提として、ハードウェア各社はデバイスレベルでの推論能力を飛躍的に高めています。従来のようにクラウドにデータを送って集中処理するだけでなく、ユーザーの手元の端末で局所的に高度な推論を行えるようになってきました。

GPU を大量に詰め込んで統計多重効果を狙えるデータセンターへの集中と、省電力でそこそこの性能を持ったローカルデバイスの両方を活用し、データの性質や利用シーンに応じて、集中処理と分散処理を組み合わせるアーキテクチャを考えていかなくてはいけません。

マイクロソフトが提供する「Copilot」などの AI アシスタント機能が、PC の標準機能として搭載されはじめています。OS レベルでの最適化により、Web 接続を維持しながらもローカルでの推論がスムーズに行われるようになると、ユーザー体験が格段に向上します。Apple Intelligence は、チップ設計段階から機械学習に最適化

されたコアを内蔵し、省電力かつハイパフォーマンスを狙う戦略です。加えて「Apple PCC（Private Cloud Computing）」のように、企業や教育機関向けにプライベートなデータ処理環境を提供しつつ、ローカルとクラウドのいいとこ取りを目指す動きも注目されています。

巨大で大賢者な LLM を端末に搭載するのは、モデルサイズや学習済みパラメータの量から、無謀な場合も多いでしょう。しかし、必要最小限の常識的知識や論理的思考力をローカルモデルに持たせ、ドメイン固有の情報は都度プロンプトや RAG で与える方式であれば、デバイスのリソースでも必要な精度の応答ができるようになるでしょう。

Google の軽量版 LLM「Gemini Nano」や、マイクロソフトの「Phi-4」といったモデルも話題です。これらはエッジ端末でも推論が可能なように最適化されており、特定のドメイン知識はクラウドとの連携で随時補完する形を想定しています。国産 LLM でも、端末レベルで動く軽量モデルが開発中との報道があり、今後の普及が期待されます。

ローカル推論の可能性は障害者支援技術にも広がります。たとえば視覚障害のあるユーザーがスマートグラスやスマートフォンを通じて「目の代わり」として情報を取得する場合、遅延のないリアルタイム解析が不可欠です。クラウドだけに頼らず、端末がある程度の画像解析や音声フィードバックを行えれば、屋外やネットワーク環境が不安定な状況でもユーザーをサポートできます。ロボット分野でも、緊急の意思決定を端末側で行い、クラウドは大局的な情報収集や学習アップデートに専念する構成が有効です。

エッジ AI やローカル推論は、一時的なブームを超えて定着しつつあります。デバイス側で高精度の AI 推論を行うためのチップ設計や軽量化モデルが続々と登場し、Apple、Google、マイクロソフトといった大手はもちろん、国産 LLM や研究機関も競争を加速しています。

今後は、利用シーンやデータの性質に合わせて、クラウドとローカルのどちらでどの程度の処理を行うかを柔軟に選択する「ハイブリッド AI アーキテクチャ」が当たり前となるでしょう。熱密度が高いクラウドデータセンターだけに頼るのではなく、ユーザー端末に特化したハードウェアや軽量モデルを駆使し、必要なドメイン知識は後付けのコンテキストとして与え、知能の地産地消を行う。こうした設計思想が、スマートカメラや AR グラス、障害者支援ロボットなど多様なデバイスに広がることで、より柔軟でプライバシーにも配慮した新しい「人間中心の AI 体験」が実現していくはずです。

推論と知識

LLM における「推論」は、たくさんデータを集めたらそれっぽく動いてしまったという段階から、もう少し内部でなにが起きているかの解明と、それを踏まえた改善が進んできました。プロンプトやチャット履歴などの「コンテキスト情報」を活用しながら、複雑な課題に対して動的に回答・推論します。その核となるのが、ReAct、Chain-of-Thought、Tree-of-Thoughts などの推論フレームワークです。r

ReAct は、Reason（考える）と Act（行動）のフェーズを分離するフレームワーク。外部ツールを呼び出すなど“行動”が明確に伴うタスクで、AI がどのような根拠でステップを踏んだかを可視化する狙いがあります。

Chain-of-Thought（CoT）は複雑な思考過程をテキストとして段階的に吐き出すことで、問題解決や推論精度を高める手法です。In-context Learning の応用形態ともいえ、数学的な推論や論理的議論など、深い考察を要するタスクで特に有効です。いわゆる“長く考える賢さ”を獲得する手段として注目されています。

CoT の応用として、多段階推論の途中の思考が非常に長くなることがあります。これを途中でキャッシュしておくことで、再利用可能にするアイデアも研究されています。いわゆる「思考の部分保存」によって、長時間の思考プロセスを段階的に処理したり、同様の問題で推論を早めたりすることが可能です。

さらに途中の「中間プロンプト」を経由せず、LLM の内部ベクトル表現や Attention マップを、次のステップに直接受け渡すアーキテクチャも検討されています。外部から見ればブラックボックスに見える推論過程を、より明らかにしていくことで推論能力と計算機コストの両立が可能になっていって欲しいところです。

多くの専門家が「2025 年に向けて、LLM の内部構造がより明らかになる」と予想するのは、この可視化・解読技術が新たな知見と高精度な制御手段をもたらすからです。AI がなぜその答えにたどり着いたのかを説明できることで、企業やユーザーに安心感を与えることができます。新たな学習手法や推論アルゴリズムが発見されれば、さらなる性能向上が期待できるでしょう。

チャットで大きく花開いた LLM ですが、チャット以外のあらゆる場所で活用されるようになり、より複雑な課題を対応するために、LLM の推論技術や、その推論に追加の文脈を与える技術の発展が期待されています。2025 年以後の大きなブレイクスルー（たとえば「AGI」のような）は、こうした積み重ねの上で登場するのでしょう。

2025 年 技術トレンド予想

今年個人的に心に残った技術や、来年盛り上がりそうなキーワードをつらつら書きます。

「2024 年予想」の振り返り

前回「TechReport 2023.12」での 2024 年予想をふりかえってみます。

メガクラウドと特化型クラウド

引き続きメガクラウド上位 3 社によるシェア寡占状況に大きな変化はありませんが、少しずつ特化型クラウドサービスの事例を目にすることが増えました。

1 年前に紹介した Cloudflare や TiDB、Snowflake といったサービスは着実に伸びていますし、今年は Firebase 代替を名乗る Supabase が正式サービスになりました。

Supabase は PostgreSQL とその上に構築された BaaS（Backend-as-a-Service）機能群を提供しています。データベースが PostgreSQL であることや開発がオープンソースで行われ、自分のサーバーを使ったセルフホストも可能であることも、シェア拡大を後押ししていそうです。

ハイパーバイザーの SoC 化

各社の専用デバイスチップは、新しい世代のものが出たり、同じような設計方針に収斂進化していきそうです。前の記事でも紹介したとおり、LLM 学習用の GPU クラスタではノード間のネットワーク帯域への要求が強く、直接液冷方式が前提となったり、データセンターにとっても大変な今後数年になっていきそうです。

ライセンスとクラウドベンダー

オープンソースプロジェクトだけでは開発者たちの利益を確保できない、というクラウド時代の問題を受けてオープンソースであることを辞めてしまうプロジェクトが増えたこの数年でしたが、良いニュースとしては Elasticsearch がオープンソースライセンスに復帰しました。これは AWS が OpenSearch として独自にプロジェクトをフォークした事を受けたものです。

新しい動きとしては、いくつか乱立した非オープンソースなライセンスを整理する形で、「Fair Source」が提唱されました。また、オープンソースプロジェクトにお金を再配分するための仕組みを作ろうという「ポストオープン」という提案もありました。

逆に今回の件を経た成功例としては、Redis から AWS が中心となってフォークした「Valkey」は、開発が盛んになったことで大幅な性能向上が実現しました。似た動きはこれまでも MySQL と MariaDB のような事例がありましたが、ここまで大きな成果を出した例はめずらしいでしょう。

イベント駆動型 API

これは正直 2024 年はあまり来ませんでした。

AWS AppSync Events など WebSocket API のためのフルマネージドサービスが登場したりと、フロントエンドを含んだイベントドリブンアーキテクチャのための道具は出てきていますが、やはり API エコシステムという責任分界点をイベント駆動にするのは、まだ時期尚早ということかもしれません。

LLM 時代の AI ペアプログラミング力

これは AI の記事のところでたくさん書かせていただきました。

間違いなく開発の主流になっていくことでしょう。

生活必需品としての GPU、そして NPU

Microsoft による Copilot+PC など、力押しでニワトリとタマゴの問題を解決しにいっているため、環境が無いのでローカル LLM を使えないという状況は少しずつ良くなっていきそうです。

その一方で、ハイエンド GPU の電力消費量はどんどん上がっていて、こちらは何かブレイクスルーが必要そうです。

そのほかの「Passkey」、「ウェブアクセシビリティ」、「リアルイベント、カンファレンス、勉強会の再開」については、個人的に大きな動きもなかったので割愛します。

というわけで、ここからは来年注目したいキーワードを取り上げていきます。

パブリッククラウド動向

全体感について、最初に毎回取り上げることにしました。

すでに書いた通りメガクラウド 3 社についてはそこまで大きな変化はないのですが、今年の大きな動きとしては国内クラウド事業者が本格的に LLM 向けデータセンターの構築に踏みきったと言うことでしょう。

現状では NVIDIA カードを国際的に取り合っている状況ですが、戦略物資として輸出が制限される国などもある一方で、熊本工場が本格稼働し始めた TSMC の動きなども注目したいところです。

LLM によって加速しているデータセンターによる電力消費量ですが、グローバルでは 4 年で 2 倍になるだろうという予測がでているようです。市場規模でも年 1.2 倍程度の伸びがあるようです。

このように消費電力量が LLM「思考能力」のボトルネックとなりつつある時代において、すでに発電能力が限界に近い日本国内は、必然的にこの需要に乗り遅れる形と

なります。使っていない農地の太陽光発電への転用といったグリーン電力の積み増しと、比較的新しく安全な原子力発電所の再稼働、CO2 排出量を削減した火力発電所の増設などの、全方面全力全開のエネルギー政策が求められます。

データエンジニアリング

これまで、非常に先進的な企業のみが活用できていたデータ基盤ですが、データエンジニアリング向けのクラウドサービスが増えてきたこと設けて、2025 年は規模の大小はあっても、多くの組織で「きちんと」やっていくことを求められる時代になったと言えます。

いくつかの特徴的な動きを紹介します。

Apache Iceberg のような「Open Table Format」と呼ばれる技術が注目されています。古くから使われてきた Apache Hadoop の Hive テーブルなどを下地に、テーブルのメタデータの保存方法などを細かく定めたものです。Hive では、Hive 側の DDL としてデータに含まれるカラムの型などを定義する必要がありましたが、Iceberg ではデータ側のメタデータとして管理されるようになり、複数のソフトウェアからの相互接続性も保たれます。たとえば AWS であれば Amazon Athena から Iceberg テーブルに直接クエリができるほか、ストレージ側も Amazon S3 Tables として Iceberg をネイティブにサポートしました。

シングルバイナリで動くデータベースエンジンとして話題の「DuckDB」も、2024 年 6 月に正式リリースを迎えたことで、目にする機会が増えました。上記の Iceberg テーブルへのアクセスも拡張がありできますし、duckdb-wasm という Wasm 版があることで、エンジンを利用者のブラウザ上で動かし、S3 等にあるストレージから直接クエリの結果を取得するといったまったく新しいパターンが現実的になりました。

Snowflake や dbt など気になるプロダクトの存在感が圧倒的で、個人的にかなり注目している分野です。

2024 年は大型インシデントが続きました。

過去最悪級のサプライチェーン攻撃であった「XZ Utils バックドア」は、多くのソフトウェアエンジニアの背筋を凍らせました。これはオープンソースプロジェクトに入り込んだ攻撃者、ある種の内部犯行でもありますが、バックドアの設置も巧妙に行われており、丁寧なソースコードレビューをしていても発見が難しい類のものでした。

クラウドストライク社による分析から、標的組織への侵入に成功してから、本格的に侵入範囲を拡げるまでは平均 62 分という報告[2]がありました。つまり、ひとりでもセキュアでないユーザーを出さないこと、そしてすぐに検知して止めるまでを自動化する、どちらもやらなくてはいけません。

このあたりは、情報セキュリティの定石通り、足回りからきちんとやっていく必要があります。つまり、穴が残っていればそこからきっちりやられる、そういう時代に進んでしまったと言うことを認め、ユーザーの認証、デバイスや情報資産の特定・管理、行動の監査、といったことを、丁寧に実行していくしかありません。

[2] https://japan.zdnet.com/article/35216900/

あとがき

年末の冬コミで出せなかった TechReport 2024.12 ですが、技術トレンドふりかえりは絶対に書きたいという事で、技書博の機会でなんとか出させていただくことができそうです。

前回の TechReport 2024.05 では Vivliostyle をためしてみたのですが、個人的にそこまでしっくりこなく、今回は一旦 Word に逆戻りです。ただ最近注目を浴びている Typst が、自分がやりたいと思っていたことが一通りで来そうで、うまく検証ができたら次は Typst でいろいろやってみたいところです。

電子版は以下の URL でダウンロードできます。

`https://files.nekoruri.jp/techreport_202412_ggtw.pdf`

感想を以下の URL でお待ちしています。

`https://forms.gle/zSJ2tpebSn1sirSF6`

めもおきば TechReport 2024.12

発行日	2025 年 1 月 25 日　初版 第十一回 技術書同人誌博覧会
著者	Aki @nekoruri aki@nekoruri.jp
発行	めもおきば https://d.nekoruri.jp/
印刷	めもおきば