# 4. 医療エージェントAI，フィジカルAI時代のNVIDIAの取り組み

山田 泰永 エヌビディア（同）エンタープライズ事業部

## エージェントAIの登場

直近のわずか2，3年の間に，「ChatGPT」（OpenAI社）を始めとする大規模言語モデル（LLM）サービスは，当たり前のツールとして社会に幅広く浸透し，活用されるようになった。そして，従来のチャット型LLMサービスにとどまらずに，自律的に複雑な工程を実行するエージェントAIが，今まさに爆発的な普及の時期を迎えている。

現在のエージェントAIは，LLMによる自然言語をインターフェイスとして，与えられた目的に向けて自ら計画を立て，一連のワークフローを構成し，必要なツールを選択して実行することが可能である。例えば，ビジネス現場であれば，「顧客から来たクレームのメールを解析し，必要に応じて社内の配送システムや顧客管理システムにアクセスして変更や関係部署への指示を出し，顧客への返信メールを自動生成して発信し，全体のアクションについての進捗状況やサマリを管理者に逐次メッセージする」といった一連のワークフローを，非エンジニアの現場の従業員が作成し，実行することが可能になる。

現時点でのエージェントAIの一つの本質は，こうした「プログラミングができない，あるいはプログラミングが専門ではない人間が自然言語だけを利用して，頭の中のイメージを具体的なアプリケーションに落とし込んだり，解析や処理のワークフローを構成して実行できる」という点ではないだろうか。

これを医療の分野で考えると，例えば，「電子カルテデータを検索して，患者Aに類似した症例を10件ピックアップしてサマリを作成，さらに，それら症例のCTデータから肝臓だけを3Dセグメンテーションしてビューワに表示」や，「臨床試験に関する複数のドキュメントを参照し，大量の患者データの中から特定の条件に該当する候補をリストアップしてランク付け，ランク上位の患者の個別のサマリを作成」というような自然言語による指示で，これまでは非常に手間が掛かったり，専門家に依頼してコストと時間を掛けなければ実現できなかったような処理が，簡単に実現する可能性がある。そして，もちろん個人レベルだけではなく，組織としてさまざまなワークフローを処理するエージェントAIを作成して，それらを「ワーカー」として常時大量に運用することも可能である。すでに，先進各社のマネージドサービスを活用すれば（そして，セキュリティやプライバシーの問題を度外視すれば），技術的にはこのような機能の実現は十分に可能性があり，特にエンタープライズビジネスの領域では活用が始まっている。しかしながら，こうしたエージェントAIの本格活用が進むにつれて，加速度的にAI計算量が増える「トークン爆発」が懸念されている。

## エージェントAI時代におけるNVIDIAの取り組み

こうした情勢の中でハードウエア面では，NVIDIAは，エージェントAI時代のCPUとして高性能かつ高効率な「Vera」CPUと，次世代GPUである「Rubin」を発表した。推論処理のスループットを最大10倍に高めるものであり，また，推論処理の中でもデコード部分に特化させたGroq 3 LPUを活用することで，さらに効率的な大規模推論環境のオプションも提供される。これらラック規模のスーパーコンピュータだけではなく，デスクサイドでも最大20PFLOPSの性能を発揮する「DGX Station」や，手のひらサイズながらも最大1PFLOPSの性能と128GBのメモリ容量でLLMやエージェントAIが実行可能な「DGX Spark」も取りそろえて，組織や個人それぞれの要求規模に見合った「AI Factory」を実現する基盤としている。

NVIDIAは，これまでのLLMを中心とする生成AIの発展には，いわば縁の下の力持ちとして，ハードウエアとさまざまなソフトウエアで貢献してきた。ハードウエア面では高速な演算処理を行うGPUをベースとして，GPU間やGPU-CPU間をつなぐ高速なインターコネクト，それら基本ハードウエアを組み合わせた「DGX」サーバなどが挙げられる。ソフトウエア面では分散並列学習を支える低レベルのMegatronライブラリ，LLMの学習を効率化するNeMoフレームワー


〈0913-8919/26/￥300/論文/JCOPY〉