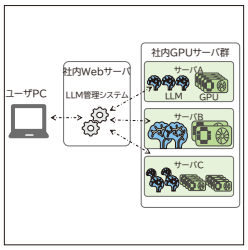

社内サーバ上で実行するローカルLLM統合プラットフォームの取り組み

Development of an Integrated Platform for Local LLMs on Internal Servers

田窪　伸哉*
Shinya Takubo

池田　龍生
Ryusei Ikeda

吉村　明展
Akinobu Yoshimura

平川　満
Mitsuru Hirakawa

近年、AI技術の急速な発展に伴い、企業や研究機関におけるAI活用が急増している。特にテキスト出力を行う大規模言語モデル（LLM; Large Language Model）の拡大が顕著である。当社では、顧客情報や、知的財産である発明アイデアなどの機密情報が社外に流出するリスクに対処するため、社内サーバ上で実行するローカルLLM統合プラットフォームの試作に取り組んでいる。この環境では、業務PCから社内のサーバへリクエストし、社内サーバ上の計算資源で推論を行うことで、完全に社内に閉じた安全なLLM活用を実現する。さらに、GPUを始めとする計算資源の共有やクライアント型ツールの構築、カスタムアプリケーション開発環境の整備により、安全性を維持しながら業務効率と研究開発サイクルの短縮に寄与できることを確認したので報告する。

In recent years, with the rapid development of AI technologies, the utilization of AI in companies and research institutions has increased significantly. In particular, the expansion of Large Language Models (LLMs) for text generation has been prominent. To mitigate the risk of leaking confidential information, such as customer data and intellectual property (e.g., invention ideas), we are developing an integrated platform for local LLMs that operate on internal servers. In this environment, requests are sent from user PCs to internal servers, and inference is executed using computing resources on the internal servers, thereby realizing secure use of LLMs entirely confined within the company. Furthermore, we report that by sharing computing resources including GPUs, building client-side tools, and implementing custom application development environments via APIs, we have contributed to improving business efficiency and shortening of research and development cycles while maintaining security.

キーワード：生成AI、ローカルLLM

1. 緒　　言

1－1　AI活用の現状と課題

近年、AI技術の急速な発展に伴い、企業や研究機関におけるAI活用は飛躍的に増加している。特に、テキスト生成を担う大規模言語モデル（LLM; Large Language Model）の拡大が顕著である。OpenAI社のChatGPTやGoogle社のGeminiなどに代表される「クラウドLLM」は、ユーザの入力情報を海外メーカが管理するサーバへ送信して演算を行い、その結果をユーザへ返信する方式であり、チャットシステムのみならず、多様なアプリケーションとの連携を通じて社会に浸透している。

これらクラウドLLMは利便性の高さと引き換えに、重大なリスクが潜んでいる。企業利用の観点では、機密情報や社内文書、顧客データなどが外部へ送信されることによる情報漏洩の懸念がある。また、特定ベンダーへの集中依存は、ベンダーの機能不全によって、事業が継続できなくなるリスクとなり得る。リスク低減のために、クラウド事業者と契約して入力内容の閲覧やAI学習への利用を制限する動きもあるが、クラウドLLMを利用する限り、データが社外へ送信されるという根本的なリスクを完全に排除することはできない。実際に情報処理推進機構の情報セキュリティ10大脅威[(1)]においても、2026年度版でAIのリスクがランク外から第3位に上昇するなど、その重要性が強く指摘されている。

このような背景を踏まえ、AI活用による企業競争力向上と、安全性の確保を両立させることが喫緊の課題となっている。

1－2　高性能ローカルLLMの進展

一方、インターネットからLLMをダウンロードし、ユーザのPCなどローカル環境で動作させることのできる高性能な「ローカルLLM」が次々と公開されている。これらはクラウドLLMに匹敵する自然言語／画像／音声などの認識能力および生成能力を有しており、技術の進展により、最新のクラウドLLMの性能にローカルLLMが追いつくまでの期間が短縮されてきている。**図1**に、クラウドLLMとローカルLLMの主要なベンチマーク性能比較を示す。

また、Hugging Face[(2)]をはじめとする、LLMやデータが公開されているプラットフォームの整備も進んでおり、ローカルLLMを扱う技術的ハードルが低下し、その活用が容易となっている。

このように「ローカル環境で安全に実行できる高性能なLLM」が現実のものとなりつつあるため、当社はクラウドLLMへの依存を脱却し、社内環境で安全にLLMを活用できる「ローカルLLM統合プラットフォーム」の構築に取り組

んでいる。

本稿では、2章で社内サーバ型LLMプラットフォーム、3章でクライアント型LLMプラットフォーム、4章でカスタムアプリケーション開発環境についてそれぞれ説明する。

2. 社内サーバ型LLMプラットフォーム

2−1 サーバ型システム構成

システム実現に向けた検証において、社内のサーバ上にローカルLLMツールを構築し、Webアクセスを通じて対話型のチャットやAPI連携を実現した。

サーバ型LLMプラットフォームのシステム全体構成を**図2**に示す。

ユーザは自身のPCからローカルLLMツールにWebアクセスし、チャット画面あるいはAPI経由でリクエストを投げる。Webサーバ側で受け取ったリクエストは社内GPUサーバに振り分け、LLMの推論結果を逆の順序をたどってユーザ側に返答する。

本方式の主な特徴は下記である。

(1) 多用なLLM選択

LLMはアーキテクチャや学習データによって特性が異なる。そこでこのプラットフォームでは、十数種類のローカルLLMを搭載し、ユーザが自由に選択できるようなインタフェースを具備する。さらに、単一のリクエストを複数のLLMに同時に送信し、各回答を並べて提示することもできる。これにより、ユーザは複数結果を比較し、それらから最適な回答を抽出する作業を効率的に行うことができる。

(2) ユーザ認証とLLM管理

未登録ユーザによる不正利用を防止するため、ユーザ認証機能を実装している。これにより、ユーザ情報を適正に

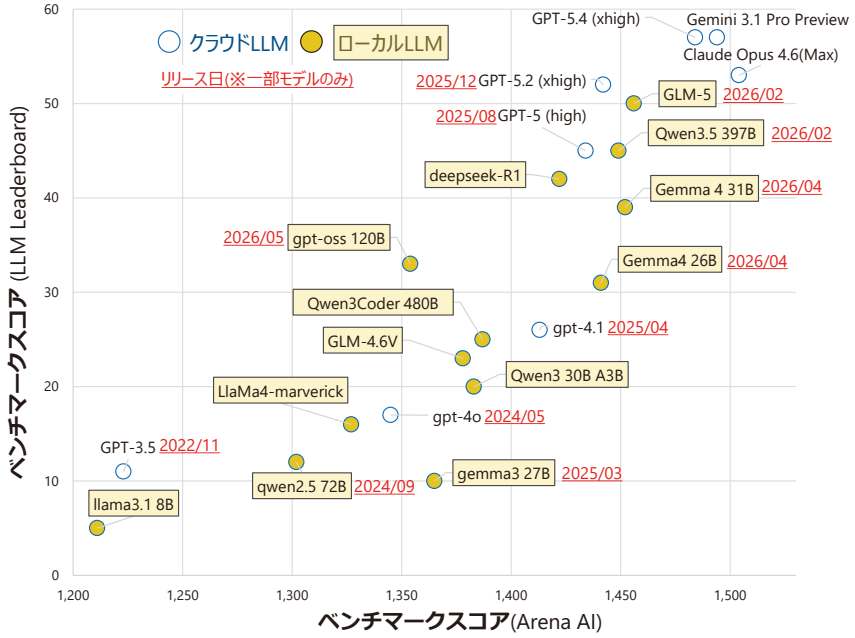

図1　クラウドLLMとローカルLLMの性能比較（2026年4月時点）
（主要なベンチマークであるLLM Leaderboard(3)とArena AI(4)で比較）

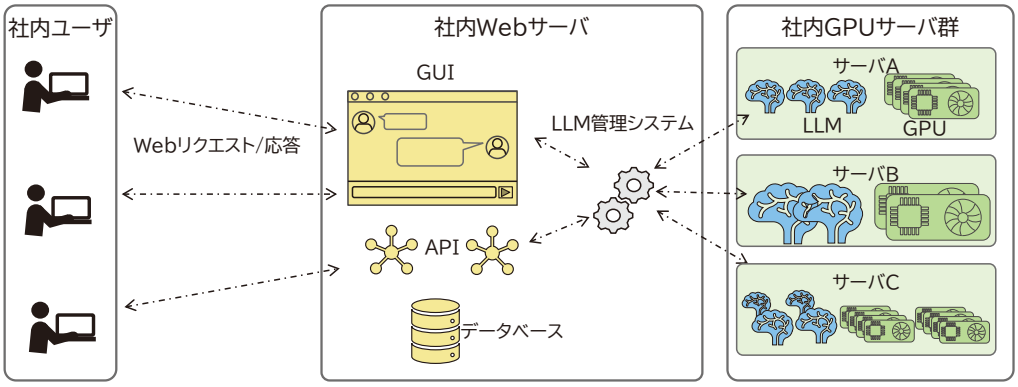

図2　社内サーバ型LLMプラットフォーム全体構成

管理するだけでなく、部門の業務特性や活用シーンに適したLLMをカスタマイズ・配備し、社内での安全な活用促進を図っている。

(3) 計算資源

このプラットフォームでは単一のGPUサーバではなく、複数のサーバに分散して構築されており、搭載するLLMのパラメータ規模やメモリの要件、各GPUのスペックに応じて、推論性能を最大限引き出せるよう、最適なサーバへの配置及びリソース割り当てを行っている。

2−2　計算資源の管理方式

LLM推論時の計算資源は集中管理方式を採用している。

これは、複数アプリケーション間でGPUリソースを動的に共有し、リソース利用率を最大化しながら開発者の環境構築負担を軽減する管理手法である。

本方式の主な利点は下記である。

(1) 他アプリケーションとのリソース共有

当社はLLM以外にも、画像認識向けのAIツールや科学計算シミュレーション向けのツール等、GPUによる計算加速が必要なアプリケーションを多数開発し社内で共有している。そのため、アプリごとにGPU割り当てを固定化せず、リクエストに応じて動的に割り当てることで、単独用途としてGPUを割り当てた場合の非効率性やコスト増大を防ぎ、全体のリソース利用率を最大化している。

(2) 計算環境共有によるメリット

我々が利用者として想定する社内の事業部や研究所の技術者は非常に高いプログラミングスキルを持つ一方、情報処理の専門家ではないため、自力で高度な社内計算環境（高額なサーバの決裁手続き、電源確保、ソフトウェアのバージョン管理、依存ライブラリの調整など）を構築することは困難で、多大な労力が必要であった。そこで、本システムでは、GPUの保守や推論サーバの基盤を一元管理することで、各技術者が生成AIをはじめとする様々なアプリケーションの活用に専念し、環境構築に対する労力を軽減することを狙いとする。

2−3　検証から得られた改善点

以上の機能を持つ、社内サーバ型LLMプラットフォームを数百人のユーザに公開しトライアルを進めたところ、1日あたりの利用は数万リクエストに達した。

同時に、システムに対する新たな要望として、下記の点が挙がってきた。

(1) 個人ナレッジとのシームレスな連携

LLMの業務活用を拡大するためには、過去の開発履歴や個人の蓄積したナレッジを参照・再活用する仕組みが不可欠である。そこで、ユーザが日常的に扱う個人データとLLMをより密に連携させ、最大限に活用できる環境を構築する。

(2) カスタムアプリケーション開発環境の提供

図2に示したサーバ型の構成では開発環境がシステム側に限定されており、ユーザが自ら機能を実装・検証することが困難であった。

そこで、Obsidian[5]等の汎用的なナレッジ管理ツールのベンチマーク検証と並行して、ユーザごとに最適化した柔軟なアプリケーションを構築し、かつユーザ自身のPC上で迅速にデバッグ・改善を回せる環境の構築を進めることとした。これにより、サーバ型の利便性を維持しつつ、個人のデータ活用とアプリケーション開発の自由度を最大化することが可能となる。

以上の検証結果から、業務PCにアプリケーションをインストールして利用する「クライアント型LLMプラットフォーム」の構築を行ったので次章で説明する。

3.　クライアント型LLMプラットフォーム

3−1　クライアント型システム構成

クライアント型LLMプラットフォームのシステム全体構成を**図3**に示す。

ユーザはまず、自身のPCに軽量なクライアントアプリケーションをインストールする。アプリケーションはサーバ型と同じくWeb経由でアクセスするツールが提供されており、以降はサーバ型同様、チャット画面やAPIからリクエストを投げる構成である。LLM推論時の計算資源自体は2−2で示した集中管理方式を踏襲することで、推論速度を維持している。

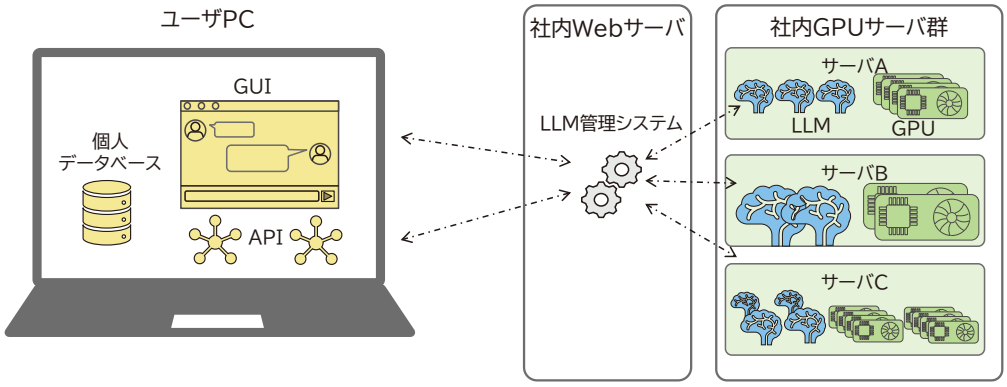

図3　クライアント型LLMプラットフォーム全体構成

3−2　システムの利点

本システムの最大の利点は、自身のPC上のデータとのシームレスな連携である。

例えば、アクセス権限のある社内サーバのフォルダパスを入力するだけで、フォルダ内にあるドキュメントをすべて走査し、調査や分析をLLMに実行させることができる。

また、デバッグ環境も同時に提供することで、部門やユーザに特化した自由なアプリケーションをユーザ側でツールに搭載できるようになる。

また、会話履歴や添付したファイルは自身の業務用PCにのみ分散して保存されるため、ネットワークが一時的に切れた場合でも、業務用PCに保存された会話履歴は閲覧でき、再接続時に自動で同期される。

上記で述べたサーバ型およびクライアント型の特徴を**表1**に比較する。

表1　サーバ型とクライアント型の比較

特徴	サーバ型	クライアント型
実行環境	社内サーバ	ユーザPC
ファイル連携	△ (必要なファイルを添付)	◎ (参照フォルダを指定)
デバッグ環境	なし	あり
導入コスト	低 (クライアント側設定不要)	中 (インストールが必要)
計算資源	集中管理	集中管理

ユーザは、ファイル連携の必要性、デバッグ環境の有無といったユースケースに応じて、サーバ型・クライアント型を選択できる。

3−3　アプリケーション開発環境の提供

クライアント型LLMプラットフォームは、単なるチャットインターフェースの提供にとどまらず、ユーザ自身がカスタムアプリケーションを開発・搭載可能な環境を同時に提供する。

本プラットフォームのクライアントアプリケーションには、LLMと連携するための APIライブラリ、および開発ツールキットが標準で搭載されている。これにより、ユーザは自身のPC上で、個人や部門特有の業務フローに最適化したアプリケーションを自由に開発・搭載することが可能となる。**図4**に、アプリケーション開発環境の全体像を示す。

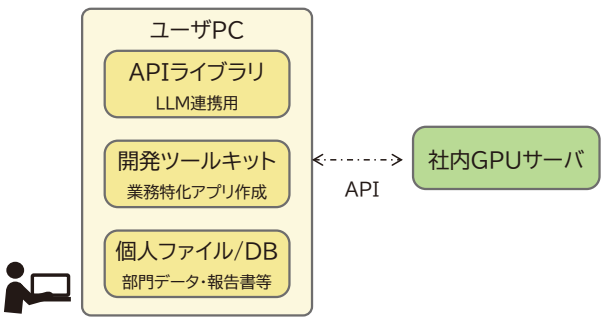

図4　アプリケーション開発環境全体像

開発環境はシンプルに設計されており、Python言語による数十行程度のコード記述で、実用的なツールを構築できる。例えば、複数のLLMを組み合わせて、1つ目のLLMに個人フォルダ内の議事録をすべて読み込ませ、結果を2つ目のLLMに渡して特定形式の報告書へ変換し、部内の社員にメールを送信するといった処理を、自身のPC上で容易に実装できる（**図5**）。

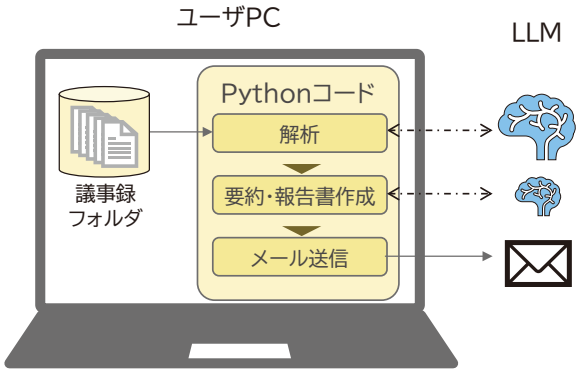

図5　カスタムアプリケーションの例

なお、推論処理自体は**図3**の構成に示す通り、社内サーバのGPU資源を利用するため、業務PCの性能に依存せず高速な応答を維持できる。

3−4　アプリケーション共有ハブの仕組み

個別に開発されたアプリケーションを組織全体で活用するため、本プラットフォームでは「アプリ登録ハブ」機能を備えている。開発したアプリケーションを他ユーザに展開したい場合、ユーザはアプリケーションをハブに登録する。登録されたアプリは、管理者によるセキュリティおよび動作検証を経て、プラットフォーム上に公開される。

他のユーザは、ハブから必要なアプリケーションを検索しダウンロードすることで、自身のPC環境に簡単に導入することができる。**図6**に、アプリケーションの開発から

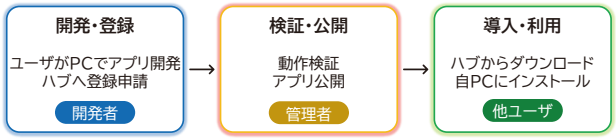

図6　アプリケーション開発・共有フロー

共有、導入までのフローを示す。

この仕組みにより、優れたアプリケーションの横展開が容易となり、社内全体の業務効率化を加速させることができる。また、アプリのバージョン管理もハブ上で一元管理されるため、ユーザは常に最新かつ安定したバージョンを利用可能である。

4. 結　　言

企業競争力向上と安全性を両立させるため、社内サーバ上で実行する「ローカルLLM統合プラットフォーム」を構築した。GPU計算資源の集中管理と共有を実現することで、開発者の環境構築負担を軽減するとともに、限られた計算資源の利用率を最大化した。さらに、クライアント型プラットフォームの導入により、業務PC上のローカルファイルとのシームレスな連携や、個々のユーザに最適化したデバッグ環境を提供し、LLM活用のハードルを低減させた。加えて、カスタムアプリケーション開発環境とアプリ共有ハブを整備することで、部門や個人特有の業務フローに最適化したツールの横展開を可能とした。

現在、本プラットフォームは社内の複数部門に展開し、実務におけるLLM活用が加速している。多様なユースケースを検証し、業務価値の高いアプリケーションやシステムを見極めていくことが重要となる。今後は、情報システム部門と連携し、社内の標準的な基盤としての実装・展開を推進する。これにより、機密情報の社外流出リスクを排除しつつ、業務効率化と研究開発サイクル短縮を実現し、持続的な競争優位性の確立を目指す。

参 考 文 献

(1) 情報処理推進機構、情報セキュリティ10大脅威2026
https://www.ipa.go.jp/security/10threats/10threats2026.html
(2) Hugging Face
https://huggingface.co/
(3) LLM Leader Board
https://artificialanalysis.ai/leaderboards/models
(4) Arena AI
https://arena.ai/leaderboard
(5) Obsidian
https://obsidian.md/

執 筆 者

田窪　伸哉*：DX技術研究所　主席

池田　龍生：DX技術研究所

吉村　明展：DX技術研究所　グループ長

平川　満：DX技術研究所　部長

＊主執筆者