生成AIは、文章作成や要約、翻訳、資料作成、プログラミング支援など、さまざまな業務で活用が進んでおり、クラウドのLLM（大規模言語モデル）サービスを日常的に使っている人も多いだろう。ブラウザからすぐに利用でき、モデルの更新もサービス側で行なわれるため、手軽に導入できるのが大きな強みだ。一方で、企業利用、特に中小企業が業務にAIを組み込む場合、オンラインサービスならではの悩みが出てくる。機密性の高い情報をクラウドに送ることに対してのハードルだ。

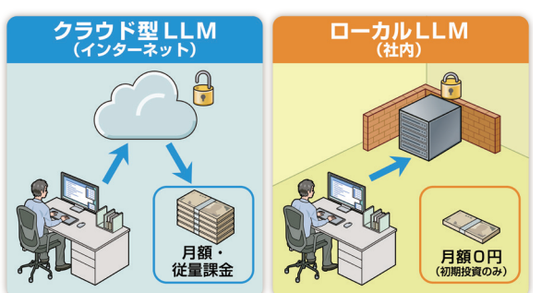
そこで候補として挙がるのが、社内にAIサーバーを用意し、自社のネットワーク内で運用する「ローカルAI」だ。本誌ではローカルLLMサーバーを利用することのメリットや、低コストに実現可能な中小企業向けローカルLLMサーバーの構築方法について解説する。

社内で解決!ローカルAI活用でコスト削減

ローカルAIは、一般的なオンラインのモデルと違い、文書検索や質問への回答生成を自社のPCやサーバー上で完結させられる。インターネット越しに外部サービスへデータを送る必要がないため、情報漏洩のリスクを抑えやすい。もちろん、社内ネットワークや端末の管理は必要だが、「AIを使いたいが、重要な情報を外に出したくない」という企業にとっては有力な解決策だ。

コスト面でもメリットはある。クラウド型LLMサービスは、個人利用なら月額数千円程度で済むことも多いが、業務で複数人が使うとなると話は変わる。人数分のライセンス費用やAPI利用料が積み上がり、毎月の固定費が増えていく。さらに、利用量が増えてくればトータルでかかるコストが読みにくくなる場合も……。ローカルAIの場合、初期投資としてPCやGPUなどのハードウェアを用意する必要はある。しかし、一度環境を構築してしまえば、モデルの利用そのものに月額課金は発生しない。もちろん電気代や保守の手間はかかるが、利用人数や利用頻度が増えても、クラウドサービスのように使った分だけ料金が増え続けるわけではない。社内で継続的にAIを使う前提なら、長期的にはコストを抑えやすい。

クラウド型LLMとローカルLLMの違い



クラウド型LLMはインターネットを介して、そのサービス提供業者にデータを送る必要がある。対して、ローカルLLMなら自社でサーバーを用意して、各社内PCからローカルネットワーク越しにアクセスできる

では、ローカルAIは実際にどのような業務で役立つのか。分かりやすい例が、社内文書の検索と要約だ。多くの企業では、業務マニュアル、製品仕様書、過去の報告書、議事録、規定などが社内に蓄積されている。しかし、必要な情報がどのファイルにあるのか分からない、検索しても該当箇所を見つけるまで時間がかかる、といった問題はよくある。

そこで、ローカルAIに社内文書を参照させれば、「この製品の保守手順を教えて」、「過去の類似トラブルの対応例をまとめて」といった自然文の質問に対して、関連文書を探しながら回答する仕組みを作れる。製造業や建設業では、技術伝承にも使いやすい。ベテランのノウハウがマニュアルや報告書、過去の作業記録に分散している場合、それらをAIで検索・要約できるようにしておけば、若手担当者が必要な情報にたどり着きやすくなる。現場で起きた不具合について、過去の事例や点検手順を確認するといった使い方も考えられる。

業種を問わず使える用途も多い。たとえば、社内Wikiのように社内規定やFAQを検索できるチャットボットを作る。開発部門やIT部門では、社内ルールや過去のコード、手順書を参照しながらコーディングやトラブルシュートを支援する。営業部門では、既存資料をもとに提案書やプレスリリースの下書きを作る。総務や人事では、社内規定に基づく問い合わせ対応や文書作成を効率化するといったものだ。

重要なのは、ローカルAIを「社内の情報を探し、整理し、下書きを作るための業務支援ツール」として位置付けること。最終判断は人間が行ないつつ、調査や整理、文章化にかかる時間を短縮する。そのための環境を自社内に持てることが、ローカルAIの大きな価値だ。

ローカルLLMの仕組み

ローカルAIの中心になるのが、「ローカルLLM」だ。LLMとは大規模言語モデル（Large Language Model）のことで、文章を理解し、質問に答えたり、要約したり、文章を生成したりするAIモデル。ChatGPTなどのクラウド型LLMサービス