まだ始めていない方、ご検討中の方、急ぎましょう！

今すぐ始めたい“ローカルLLM”

オンプレミス環境でLLM（大規模言語モデル）を運用する際には、
ニーズや要件、リソース、セキュリティポリシーなどに応じて検討する必要があります。
ローカルLLMに取り組む際のポイントをまとめました。

LLMとは？

膨大なテキストデータを用いてトレーニングされ、言語のパターンや文脈を学習し、言語理解や生成のタスクを行うことができるモデル。多層のニューラルネットワークからなり、膨大な量のデータを処理し、高度な言語理解が可能です。

文章生成

与えられた文脈に基づいて自然な文章を生成することができます。例えば、文章の翻訳、要約、クリエイティブライティングなど。

言語理解

テキストデータから意味や文脈を理解し、質問応答、文章分類、感情分析などのタスクを実行するのに役立ちます。

対話システム

LLMは、対話システムやバーチャルアシスタントに使用され、ユーザとの自然な対話を実現します。

情報検索

検索エンジンの質問やクエリ※を理解し、適切な情報を提供するのに活用されます。

※ クエリ：データベースや情報検索システムなどに対する、データの取得や操作を目的とした問い合わせのこと

ローカルLLMとは？

オフラインで行えるLLM。秘匿性の高いデータやクラウドに乗せたくないオリジナルデータを取り扱う場合、AIに外のデータから学ばせたくない場合などにクローズドな環境で行うLLMです。

限られた環境で行うため
データの秘匿性が守られる

さまざまな事業やサービスで
活用が始まっています

オンプレミス環境で行う理由、メリット、注意点

理　由	1 セキュリティとプライバシーの制御	2 コスト管理	3 カスタマイズと制御
メリット	1 完全なデータ管理	2 パフォーマンスの最適化	3 運用の柔軟性
注 意 点	1 インフラストラクチャの管理負担	2 スケーラビリティの制限	3 技術の更新とアップデート

ローカルLLMに取り組む際のポイント

計算リソース

LLMは計算リソースを大量に必要とするため、十分なCPUやGPUの性能を持ったマシンが必要。リソース不足は、全てが遅くなる原因に。

GPUメモリ要件

モデルのサイズに応じて、十分なメモリ確保が必要。メモリが足りないと、大規模なモデルのロードはできません。

データセキュリティ

ローカル環境でLLMを運用する場合でも、セキュリティは重要。機密情報や個人情報が含まれる時は、アクセス制御や暗号化の実装を。

モデルの管理

ローカル環境でのLLM運用は、モデル管理が重要。定期的なバックアップやバージョン管理で、誤った変更や損失を防ぐことが必要。

オンライン/オフライン運用の考慮

オンラインでリアルタイムの推論が必要な場合と、オフラインでバッチ処理を行う場合の両方を考慮することが必要。それぞれの運用に合わせたシステム設計が重要。

監視とトラブルシューティング

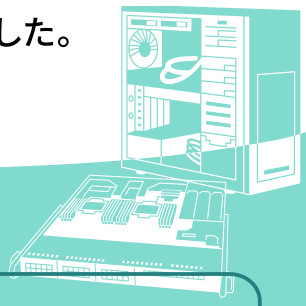
ローカルLLM運用では、システムの監視とトラブルシューティングが重要。システムのパフォーマンスやエラーをモニタリングし、問題が発生した場合は迅速な対処が必要。

リソースの最適化

計算リソースやメモリなどのリソースが限られている場合も。リソースの最適化を行い、効率的にモデルを運用するための努力が必要。

VISUAL TECHNOLOGY

実際にオンプレ環境を用意する時に、留意する点をまとめました。
ハードウェアを購入する時に、参考にしてみてください。



LLMのための

オンプレ購入の際に気をつけるポイント

計算リソースの確保

LLMは計算リソースを大量に必要とするため、CPUやGPU、NPU※などが高性能なハードウェアの選定が必要。モデルサイズや処理量に応じて、十分な計算リソースの確保が重要。

※ Neural Processing Unit

メモリとストレージの容量

大規模なデータセットやモデルをロードするための十分なメモリとストレージ容量が必要。モデルサイズや使用するデータ量を考慮し、適切なメモリとストレージの選定を。

ネットワークインフラの強化

高速で安定したネットワークインフラが必要。データの入出力やモデルのアップデートには大量のデータ転送が必要となるため、高性能なネットワークを導入することが重要。

セキュリティとコンプライアンス

LLMの運用には、セキュリティとコンプライアンスへの配慮が不可欠。特に機密情報や個人情報を扱う場合は、適切なセキュリティ対策とコンプライアンス要件の遵守が必要。

拡張性とフレキシビリティ

LLM運用は時間とともに変化する可能性があるため、拡張性とフレキシビリティが重要。将来的な需要増加や新しいハードウェア導入に対応できるような、システムの設計を。

サポートとメンテナンス

適切なサポートやメンテナンスサービスを提供しているベンダーを選定し、運用中に問題が発生した場合に迅速かつ効果的に対処できる体制を整えることが重要。

LLMではGPUが特に重要！

LLMで、GPUが重要な役割を果たす理由

並列処理能力

LLMは、数十億から兆単位ともいわれるパラメータを持つ非常に大規模なモデルです。パラメータを効率的にトレーニングするためには、数千個ものコアを持つGPUの並列処理能力が不可欠です。

高速な演算

GPUは、LLMのトレーニングや推論に必要な、多くの行列演算やテンソル演算を高速に処理することが可能。最新のGPUは、搭載メモリの増加、機械学習/AI/LLMなどに必要な演算機能の強化が進んでいます。

マルチGPU

接続バスの高速化、GPUメモリの大容量高速化へと進化中のGPUは、RDMA（GPUDirect）をサポートするInfiniBandの使用で、マルチノード化し、より大規模な計算処理を高速に行うことができます。

実験の迅速化

LLMの開発や実験では、さまざまなハイパーパラメータの組み合わせを試すことが一般的です。GPUで、これらの実験を迅速に実行し、効率的に最適なモデルの構築や調整を行うことができます。

上記の内容を踏まえて、ビジュアルテクノロジーにお任せください！

システムのご提案からサポート・保守までご提供

マルチベンダ体制

マルチベンダだからこそ、さまざまなメーカーの取り扱いがあります。ユーザ様の利用環境やご予算に合わせて、システムのご提案と製品のご提供をいたします。

SEサービス

ハードウェアの販売だけではなく、セミオーダーのシステム構築をはじめ、ソフトウェアのインストール、保守や設置作業などのエンジニアリングサービスも行います。

お客様サポート

これまで、多くの研究開発機関や官公庁、大学、企業にハードウェアを納品しております。その実績を活かし、皆さまの事業や研究開発をサポートいたします。

HPC分野での実績

GPUメモリ量が重要です。マルチノード化の検討やGPU間のインターコネクト、InfiniBandなども最適な選択が必要です。HPC分野で実績のある弊社にお任せください！

お問い合わせ

どんなことでもお気軽にご質問・ご相談ください！
弊社サイトのお問い合わせフォームへご入力いただくか、下記までご連絡ください

メール vt-sales@v-t.co.jp　TEL 03-6823-6789（受付時間 平日10:00～17:00）

お問い合わせフォーム ▶

ビジュアルテクノロジー株式会社
〒111-0052　東京都台東区柳橋2-1-10

VISUAL TECHNOLOGY

www.v-t.co.jp