ローカルLLMの可能性
- オープンなモデルが拓くAI活用の展望 -

2025年8月4日

株式会社日本総合研究所

＜本資料に関するお問い合わせ＞
執筆者：先端技術ラボ 伊藤蓮、近藤浩史
本レポートに関するお問い合わせにつきましては、当社ホームページの 問い合わせフォーム よりご連絡ください。

はじめに

● 背景：

大規模言語モデル（LLM）が注目を浴び、各社で利活用が進んでいる。これまでの活用形態としては、OpenAI社などが提供するクラウド上のLLMをAPI経由で利用することが大半であった。一方、モデル自体が公開されているオープンなLLMの性能が向上してきたことで、自社環境上でLLMを活用する「ローカルLLM」も現実的になりつつあり、関心が高まっている。本レポートでは、「ローカルLLM」の最新の技術動向やビジネスにもたらす可能性について記載する。

● 本レポートの目的：

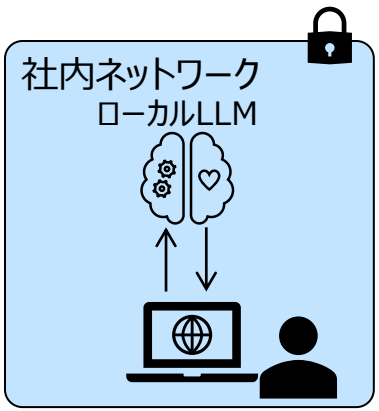
本レポートでは、ローカルLLMの概要・動向・想定ユースケース・導入事例・将来展望を整理した。読者がローカルLLMへの理解を深め、今後の技術投資やPoCの企画・実施、導入可否判断に活用することを目的としている。

● 対象読者：

本レポートの対象読者は、①LLM導入を進める担当者・意思決定者、②ローカルLLMについて概要を知りたい担当者・エンジニアとしている。LLMに関する基礎知識があると、より理解が深まるものと思われる。

従来のLLM活用に対する問題意識

ローカルLLMの可能性について考察



ローカルLLMのメリット*
モデルのカスタマイズ性
モデルの制御容易性
データセキュリティ・プライバシーの向上
リアルタイム処理・オフライン環境への対応

*詳細はP.11~13参照

エクゼクティブサマリ

日本総研
The Japan Research Institute, Limited

ローカルLLMとは？ そのトレンドは？(1章)	• ローカルLLMとは、クラウドで提供されているクローズドなLLMと対比して、自社でモデルを管理・運用できるLLMのことを指す。 • 従来は性能の観点から、ビジネスユースケースで用いられるLLMは一部のクローズドなLLMに集中していたが、高性能かつオープンなLLMが次々と登場し、利用するLLMの選択肢が増加している。 • 中国発のDeepSeekといった高性能なモデルがオープン化され、学習方法も公開されたことで、モデルのオープン化の流れ、および、ローカルLLMの活用が加速する可能性がある。
ローカルLLMのメリット・デメリットは？(2章)	• ローカルLLMを利用することで、クラウド提供の汎用的なLLMと比較して、①モデルのカスタマイズ性、②モデルの制御可能性、③データセキュリティ・プライバシーの向上、④リアルタイム処理・オフライン環境への対応といったメリットがある。 • デメリットとして、①クラウド提供のLLMと比較して十分な性能が出ない可能性、②初期導入費用の高さ、③モデルの保守運用コストの高さ、④専門知識を持った人材確保の困難さなどが挙げられる。
ローカルLLMの活用ユースケースや事例としてどのようなものがあるか？(3章)	• ローカルLLMのメリット・デメリットを考慮したうえで、目的やユースケースに応じた最適な導入パターンを検討していくべき。 • 顧客情報や高度な専門知識を活用することも多い金融業界、独自データやノウハウを有することが多い製造業、病歴情報などセンシティブな情報を扱うことも多い医療業界、組織ごとに独自のデータや業務フローを有し、かつ個人情報を取り扱うことも多い官公庁などで有効な事例が存在する。 • 複数社でのローカルLLM開発の取り組みや、ローカルLLM構築サービスの提供等の動きも見られる。
ローカルLLMにおける技術的な課題は？ 今後どのような変化がもたらされるか？(4章)	• ローカルLLMの開発・運用に向けて、①モデルの小型化と精度の両立、②推論速度の向上、③効率的なファインチューニング・学習手法の確立、④モデル管理手法・ツールの確立、といった技術課題があり、今後も研究開発が進んでいく。 • 今後、オープンなLLMの軽量化や高性能化が進み、ローカルLLMとして実用的なモデルが増加する。結果として、ローカルLLMを利用するモチベーションや意義が高まっていく可能性がある。 • クラウド提供のLLMとローカルLLM、エッジデバイスでのSLMのハイブリッド活用など、柔軟なシステム構成パターンが検討されていくものと予想される。

目次

日本総研 The Japan Research Institute, Limited

章	項目	ページ
1．ローカルLLMの概要・動向	• 本レポートで扱うローカルLLMの定義 • ローカルLLMを取り巻くトレンド • オープンなモデルの発展 • モデルの選定基準 • 利用可能なモデルの例 • 中国発のオープンなLLMの台頭とローカルLLMへの影響	pp.5-10
2．ローカルLLMのメリット・デメリット	• ローカルLLMのメリット・デメリット一覧 • ローカルLLMのメリット • ローカルLLMのデメリット	pp.11-15
3．ローカルLLMの活用パターンと事例	• LLMのシステム構成パターン • 業界ごとの主なローカルLLMの活用ユースケース • ローカルLLMの活用事例 • ローカルLLM構築サービス	pp.16-21
4．技術課題と今後の展望	• ローカルLLMの技術課題 • 今後の展望	pp.22-24

本レポートで扱うローカルLLMの定義

●「ローカルLLM」という用語は様々な文脈で利用されることがあるが、下記に示す**「システム構成の定義」と「モデルの定義」の両方**を満たすものを、本レポートで扱う**「ローカルLLM」の対象**とする。

システム構成の定義

- 本レポートの「ローカルLLM」とは、クラウド提供のLLMと対比し、**モデルを自社で管理・運用できるもの**を対象とする。

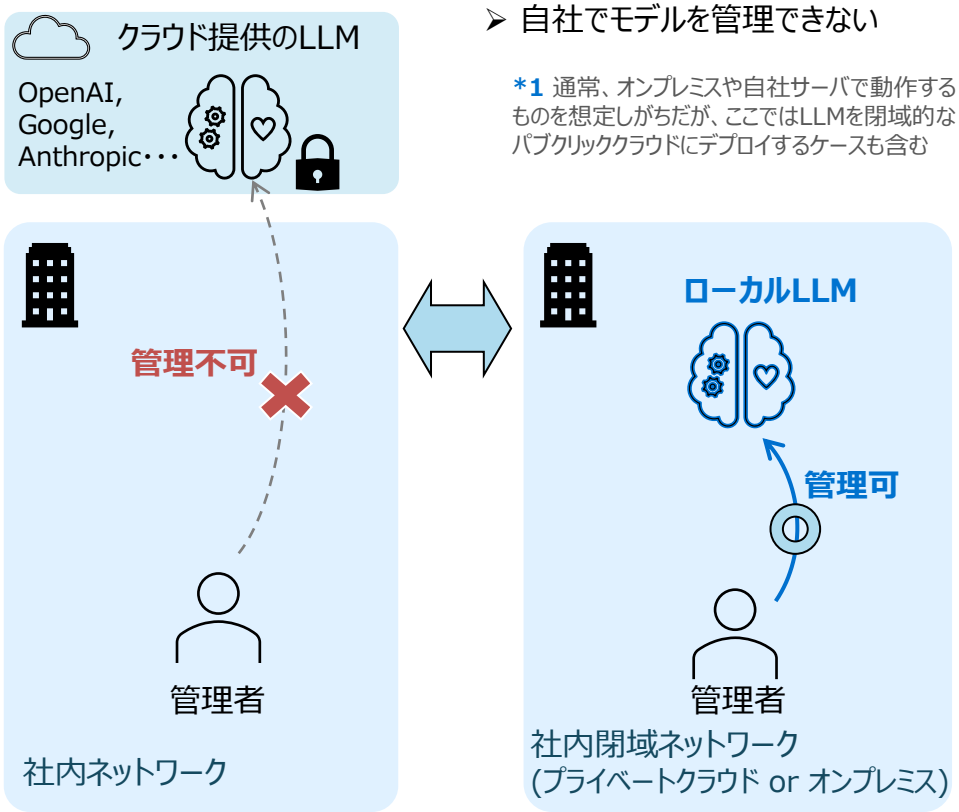
クラウド提供のLLM

- API経由でモデルを利用するのみ
- 自社でモデルを管理できない

ローカルLLM

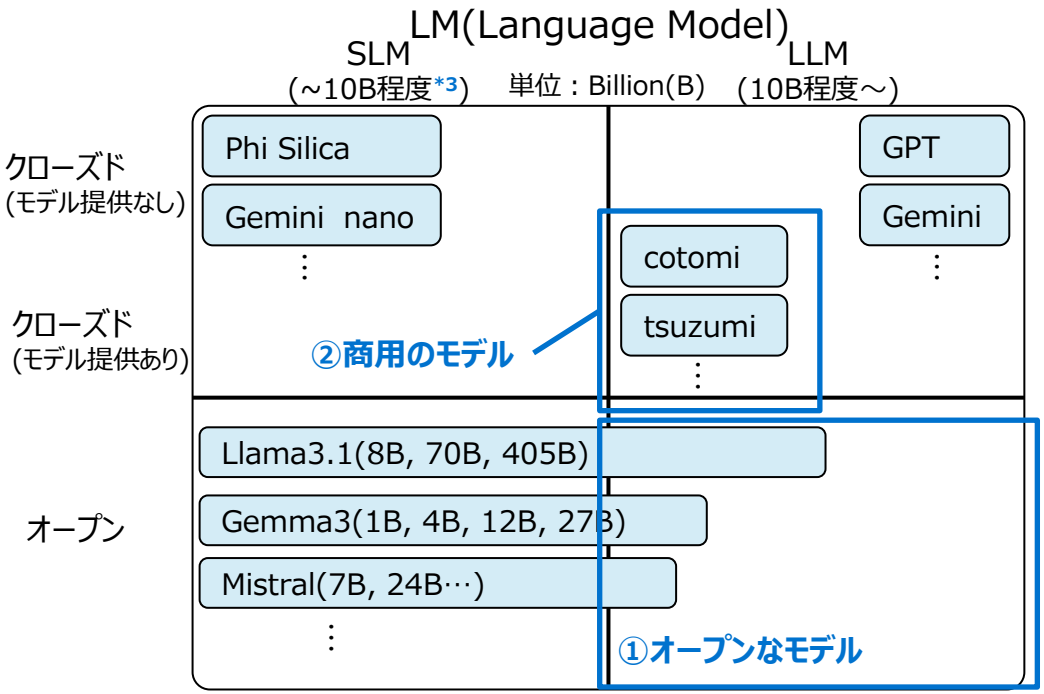
- 閉域ネットワーク内に必要なシステムを構築・モデルをデプロイして利用*1
- 自社でモデルを管理できない

*1 通常、オンプレミスや自社サーバで動作するものを想定しがちだが、ここではLLMを閉域的なパブクリッククラウドにデプロイするケースも含む



モデルの定義

- モデルの側面からは、①重み・ライセンス公開で自由に微調整できる**オープンなモデル**、または②重みは非公開でも有償ライセンスにより自社環境に導入して微調整・運用できる**商用のモデル**、のいずれかに該当する大規模言語モデルを対象とする。
- LLMよりも更に軽量なSLM*2は、性能が限られ、実ビジネスでの汎用的な応用に適さないこともあるため、対象外とする。



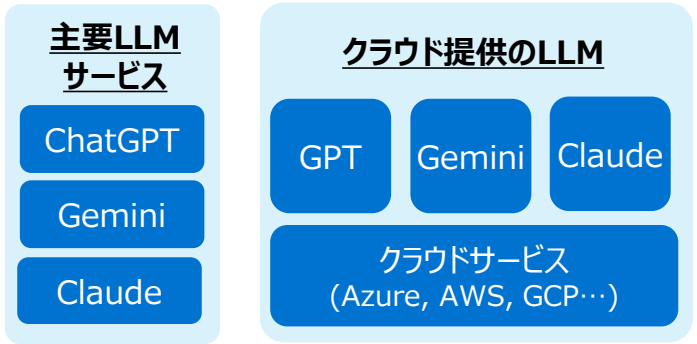
*2 Small Language Model
*3 SLMのパラメータサイズの定義はLLMと比較した相対的な値でしかなく、厳密な定義は存在しない。ただし、一般的に数B程度のモデルをSLMと呼んでいることが多い

ローカルLLMを取り巻くトレンド

- ChatGPT台頭以降、性能の観点からビジネスユースケースで用いられるLLMは一部のクローズドなLLMに集中していた。2024年4月以降、高性能かつオープンなLLMが次々と登場し、利用するLLMの**選択肢が増加**している。
- 最近は、クローズドなLLMに搭載された機能や技術が、**数か月後にオープンなLLMに取り入れられる**流れもある。

ChatGPT台頭以降 (2022/11~)

ビジネスユースケースにおけるLLMの活用では、性能の観点から、クローズドなクラウド提供のLLMサービス・APIの利用に集中[*1]

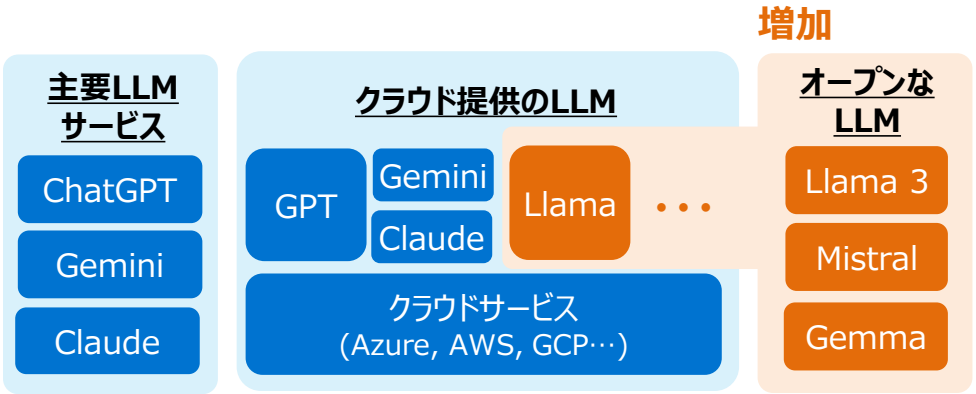


クラウドベースでの利用 (SaaS、API利用)

商用利用向けのモデルのベンダ、種類ともに**限定的**

高性能なオープンLLMの台頭 (2024/4[*2]~)

クラウド提供のLLMのみならず、オープンなLLMも選択肢として検討出来る環境に変化。



クラウドベースでの利用 (SaaS、API利用)

API利用に加え ローカルでの利用も可能

商用利用向けのモデルのベンダ、種類ともに**多様化**

*1 無論、オープンなモデルを開発して提供する組織はこの時点でもあったが、クラウド提供のLLMと性能面で大きな乖離があり、実用的とは言えなかった

*2 ここでは、クローズドなLLMの性能に引けを取らないとされる、Meta Llama 3が発表されたタイミングとしている。

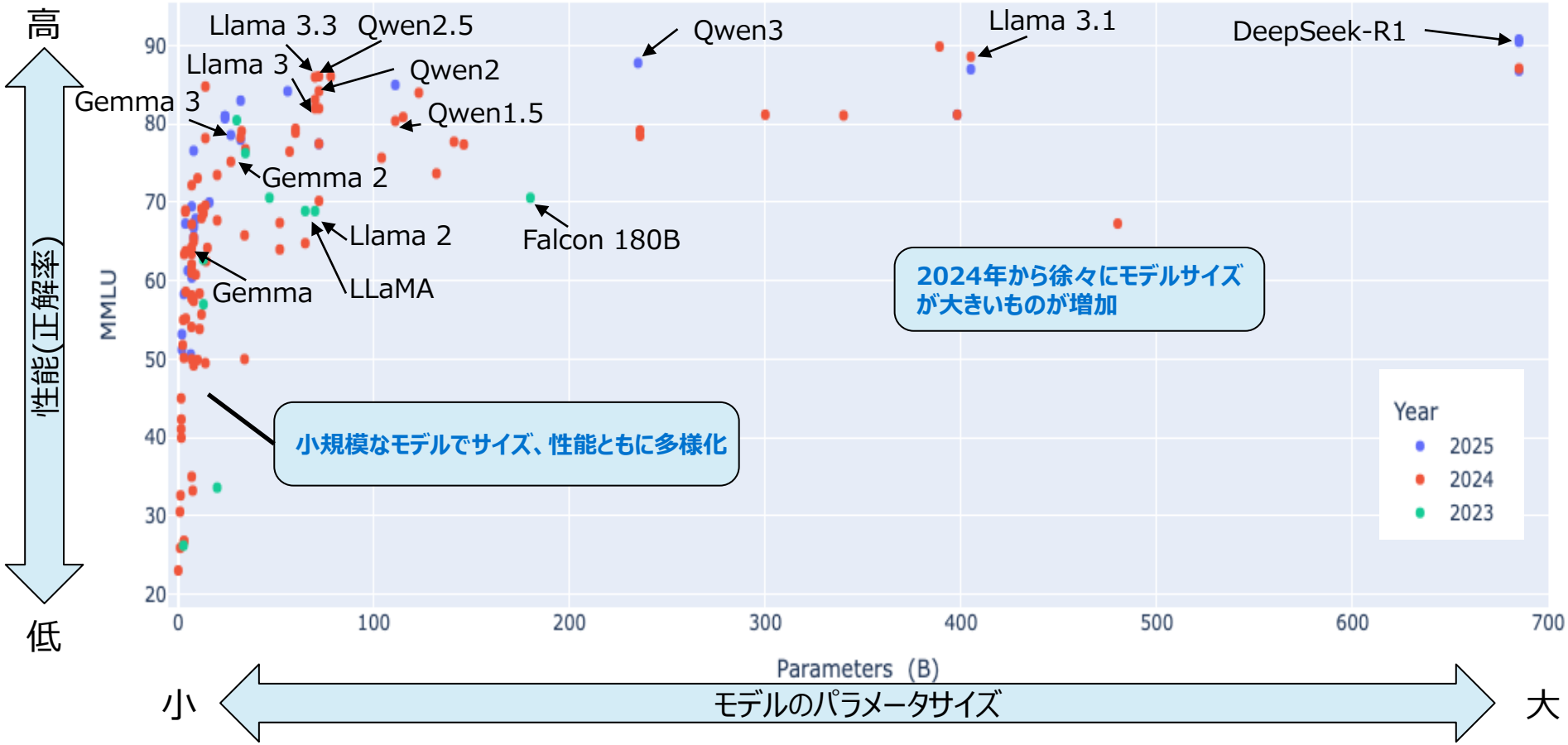
オープンなモデルの発展

- Meta社が2023年2月に「オープンサイエンス」の取り組みの一環として研究者向けにリリースした“LLaMA”を筆頭に、様々な組織や企業がローカルLLMとして利用できるオープンなモデルを発表。
- 2024年にオープンなモデルが多数公開されている。小型モデルが多いが、パラメータサイズも徐々に増加している。

リリースされているオープンなモデルの分布

MMLUベンチマーク*1の性能指標*2を縦軸に、横軸にモデルのパラメータサイズを取ったオープンなモデルの分布を示すグラフ

*1 初等数学やコンピュータサイエンス、法学など多分野のLLMの性能を評価するためのベンチマーク。
*2 MMLUベンチマークでは既にモデル間の微妙な性能差を測定できることができないのではないかと指摘されている点や、モデルの学習データにベンチマークデータが混入してしまう「ベンチマーク汚染」といった問題も指摘されていることに注意



[出所] “Models Table”, LifeArchitect.ai, https://lifearchitect.ai/models-table/ (参照：2025/5/8時点)を基に日本総研作成

モデルの選定基準

- ローカルLLMとして利用するモデルを選定する際にはいくつか基準がある。
- 利用したいユースケースに応じて、下表で示すような様々な観点を考慮して適切なモデルを選定する必要がある。

考慮すべき観点の例	詳細
ライセンス	モデルの利用にあたっては必ずそのモデルのライセンスを確認する必要がある。 ポイントとしては**①商用利用可能かどうか**、**②ライセンスの種類**(利用や改変、再配布に関する条件の確認)、**③その他利用制限事項がないか**(特定用途への使用禁止、再頒布・派生物公開義務、クレジット表示義務など)といった点。 例えば、商用利用が可能なLlamaのライセンスでは、月間のアクティブユーザが7億人を超える場合にはMetaに追加の承諾が必要となるなど、制限も存在する。
モデルのサイズ	モデルを運用するためのインフラコスト(ハードウェアや消費電力、ネットワーク設備などのコスト)に影響を及ぼすため、許容可能なサイズを事前に決定し、適切なサイズのモデルを選定すべき。
扱えるトークン数	ユースケースに応じて、モデルへ入力するテキストの長さ(⇒トークン数)が異なる。想定ユースケースがどの程度の長さが必要なのか見極めたうえで、選定するモデルのコンテキスト長が十分かどうか判断すべき。 目安としては、英語であれば4文字で1トークン程度、日本語であれば1文字で数トークン程度となる。
性能	扱いたいユースケースに応じて、どの程度の出力性能が出ていれば十分なのかを考え、公開ベンチマークなどを参考にモデルを選定。その際、できるだけ実際の想定ユースケースでの性能評価を自社で行い、性能を詳細に見極めたうえでモデルを導入していくのが望ましい。不適切な返答や有害な出力の有無など、LLMの毒性の見極めも重要。
学習データの透明性	モデルの学習に利用されているデータが公開されている場合もあり、どのようなデータを基に学習されているのか確認できる場合もある。使用データの内部統制が求められるようなユースケースでは、学習データが公開されているモデルが望ましい場合がある。また、使用しているデータのライセンスにも注意が必要である。
対応言語・マルチモーダル性	日本語に対応していないモデルもあるので、日本語での利用が想定されている場合には要注意。 また、画像入力なども必要なユースケースでは、テキストのみでなくマルチモーダルなモデルを利用する必要がある。

利用可能なモデルの例

ローカルLLMとして利用可能なオープンなモデルシリーズの代表例*1

モデルシリーズ名（開発元）	説明
Llama (Meta)	•オープンなLLMの先駆的存在。 •当初は、研究者向けに、大規模なインフラリソースが必要とせずに利用できるモデルとして提供。 •商用利用可能で、派生モデルも多数存在するシリーズ。
Phi (Microsoft)	•当初は、スマホなどでの利用が前提に開発されたSLM。 •徐々にサイズが拡大し、性能が伸びてきている。
Gemma (Google)	•Geminiと同じ技術を使って開発されたオープンなモデル。 •軽量だが、コンテキスト長の長さや高性能な点が特徴。
DeepSeek-R1 (DeepSeek)	•2025年1月にリリースされた初のオープンな推論モデル。OpenAI o1に匹敵する性能を示し、話題となった。 •高性能でMITライセンスで提供されたことから、既に多数の派生モデルが存在。
Qwen (Alibaba Cloud)	•多言語テキスト生成・長文推論・コード生成など、幅広い用途に利用できるモデル。 •一部のベンチマークスコアでOpenAIやDeepSeekの推論モデルを超える性能を示す。
Mistral/Mixtral (Mistral AI)	•フランスのスタートアップ企業が開発。 •MixtralはMoE(Mixture of Experts)と呼ばれるアーキテクチャの先駆的なモデル。
PLaMo (Preferred Networks) 国産LLM	•国からの支援により開発された純国産のフルスクラッチモデルで、日本語で世界最高レベルの性能を示す。 •オープンなモデルの提供だけでなく、API経由での提供も行っている。
Sarashina (SB Intutions) 国産LLM	•国内最大級のAI基盤を構築し、開発。 •中長期的に海外のLLMと同等の超大規模モデルの開発（1000B規模）を目指す
LLM-jp (NII) 国産LLM	•産学連携でオープンなLLM開発基盤の確立を目指して開発された。 •学習データも含めて全てオープンなLLMとしては、世界最大規模。

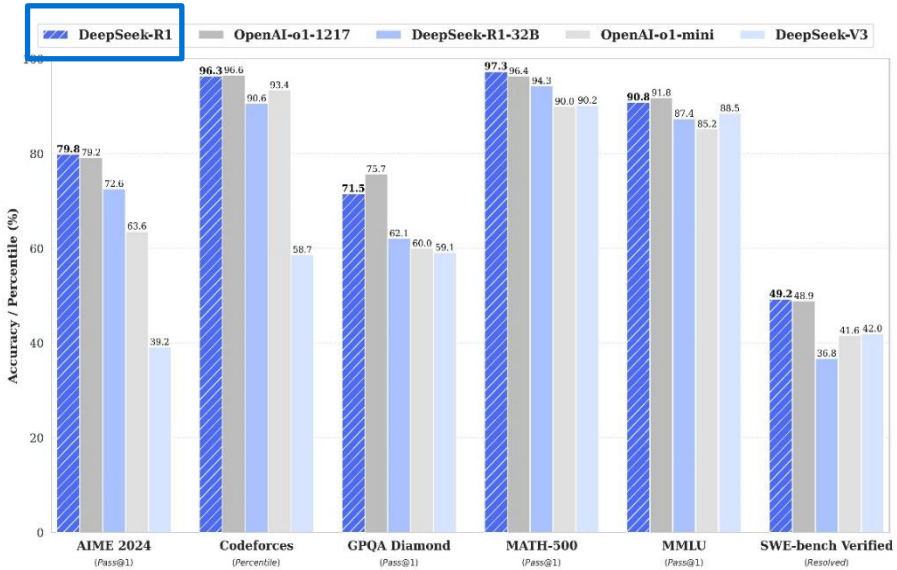
*1 表に抜粋して記載したモデルの情報はあくまでも執筆時点での目安であり、詳細は各モデルの最新情報を参照。

中国発のオープンなLLMの台頭とローカルLLMへの影響

- これまでLLM業界は米テック企業のクローズドなモデルが牽引してきたが、中国発のオープンなLLMであるDeepSeekが登場し、OpenAIが開発した最先端のモデルとほぼ同等の性能を見せたことで話題となった。
- 高性能なモデルがオープン化され、学習方法も公開されたことで、ローカルLLMの活用が加速する可能性がある。

DeepSeek（DeepSeek）

- 中国企業のDeepSeekは汎用的なV3モデル(2024年12月末)と推論に特化したR1モデル(2025年1月)をリリースし、それぞれGPT-4o、OpenAI o1モデルと同等の性能を示した。
- 米テック企業が大量のGPUリソースを使用して学習したのに対して、DeepSeekはより**低コストで学習**した（**真偽についてはさまざまな意見あり**）とされ、LLM開発に対する新たな可能性を示した。
- 商用利用も可能なMITライセンスの**オープンなモデルとして公開**され、派生モデルが多数開発されている。



[出所] "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning", https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf (参照：2025/3/27)

ローカルLLM活用加速への可能性

- 高性能なオープンなモデルを基にした派生モデル開発により、モデルの多様化が進み、ローカルLLM活用を後押しする可能性がある。
- これまで基本的にクローズドな戦略を取ってきたOpenAIのような組織でも、こうした潮流を受け、オープン化へ進む可能性もある。

主な著名人のモデルのオープン化への姿勢

著名人	説明
やや肯定派へ？ サム・アルトマン (OpenAI CEO)	Redditの人気コーナーで、オープン化についてOpenAIが「**歴史の間違った側にいる**」ことを認めた発言をした*1。 2025年4月には自身のXの投稿で、「**数か月後にオープンな推論モデルをリリース予定**」と明言*2した。
肯定派 イーロン・マスク (X/xAI CEO)	OpenAIがクローズドな戦略を取ってきたことに批判しており、xAIで開発した「Grok-1」をオープン化。 今後、後継モデルの**「Grok-2」もオープン化を表明**。
肯定派 エリック・シュミット (元Google CEO)	「西洋諸国は**オープンソースのモデル開発に注力**しなければ中国のオープンソースモデルに負ける」可能性を指摘*3。
やや反対派 ダリオ・アモデイ (Anthropic CEO)	AIの安全性を優先して開発していることもあり、現状は**モデル自体をオープンにする動きは見せていない**。
反対派 ジェフリー・ヒントン (AIの父)	2024年にノーベル物理学賞を受賞したヒントン氏は、「**比較的容易に悪用できてしまうため、オープンでLLMを公開することに対する懸念**」を示しており、現在でもその姿勢に変化はない*4。

*1 "Sam Altman: OpenAI has been on the 'wrong side of history' concerning open source", TechCrunch, 2025/1/31, https://techcrunch.com/2025/01/31/sam-altman-believes-openai-has-been-on-the-wrong-side-of-history-concerning-open-source/ (参照：2025/3/28)
*2 同氏は7/12に、モデルの安全性テストの観点からモデル公開時期を延期する旨の追加投稿を行っており、現時点で公開時期は未定。
*3 "Ex-Google chief warns west to focus on open-source AI in competition with China", https://www.ft.com/content/84cf0b2e-651d-4cb4-b426-ebc7afd634fa (参照：2025/3/28)
*4 "Open source LLMs could make artificial intelligence more dangerous, says 'godfather' of AI", https://www.techmonitor.ai/digital-economy/ai-and-automation/open-source-chatgpt-ai-llm-geoffrey-hinton?cf-view&cf-view (参照：2025/3/28)

ローカルLLMのメリット・デメリット一覧

- ローカルLLMを利用する際にはメリットとデメリットのトレードオフを考慮する必要がある。
- メリット・デメリットを基にして自社の状況や目的などを明確にしたうえでローカルLLMを利用するかどうか判断していくことが重要となる。

メリット

モデルのカスタマイズ性
特定業務に特化したモデルを開発する、自社独自のデータが利用しやすいなど、自社のニーズに合わせたチューニングや設定がしやすい。

モデルの制御可能性
- 自社でモデルを管理・運用するため、LLMベンダに依存するリスクを抑え、モデルの挙動を制御しやすくなる。
- Fine-tuningを行うような場合は、訓練データをある程度は自社で確認し選択したうえで利用できるため、**不適切な学習データが含まれるリスクを抑えられる**。

データセキュリティ・プライバシーの向上
機密情報を外部のクラウドに預けず、自社内で管理できる。

リアルタイム処理・オフライン環境への対応
エッジ側での高速な処理が必要な状況や外部ネットワークに接続しにくい状況でも対応できる。

デメリット

十分な性能が出ない可能性
クラウド提供のLLMと比較して、性能低下が生じる可能性がある。

初期導入費用の高さ
専用のハードウェアやネットワーク構築など、初期投資が必要。

モデル管理・運用コストの増大
セキュリティを担保する仕組みも含めて、モデルを管理・運用するための取り組みを自社で行う必要がある。

専門知識を持った人材確保の困難さ
ローカルLLMの開発・運用に関する専門知識を持った人材を確保する必要があるが、このような専門知識を持つ人材は貴重である。

ローカルLLMのメリット (1/2)

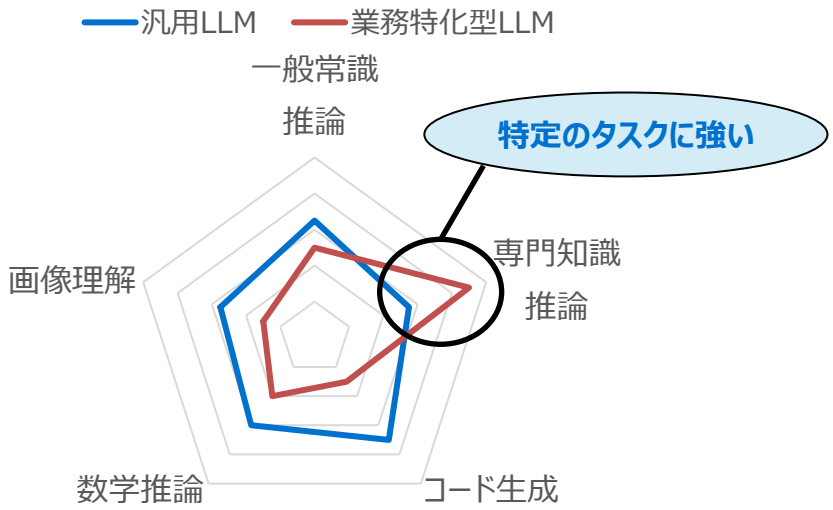
●ローカルLLMを利用することで、クラウド提供の汎用的なLLMと比較して、①モデルのカスタマイズ性、②モデルの制御可能性、③データセキュリティ・プライバシーの向上、④リアルタイム処理・オフライン環境への対応といったメリットがある。

メリット①：モデルのカスタマイズ性

- 自社の**特定業務に特化**させたモデルを構築でき、汎用的なLLMでは**不十分な性能を補える可能性**がある。
- **自社独自のデータや業界固有の知識**を利用したモデルを構築・運用でき、場合によってはユーザからの評価などのログを基にモデルを**継続的に改良**することもできる。
- ただし、特化させたモデルを開発しなくても、独自データの利用はRAGで十分に対応できる可能性もあることに注意。

モデルの性能イメージ

汎用的なモデルでは、あらゆる性能が満遍なく一定レベルに達している一方で、適切にカスタマイズされた業務特化型モデルでは**特定のタスクでより高い性能を出せる可能性**がある。



メリット②：モデルの制御可能性

- **自社でモデルを管理・運用できる**ため、バージョン変更による急な挙動の変更など、**LLMベンダに依存するリスクを抑えられる**。特に、AIエージェントについては挙動をプロンプトで制御するため、**モデルのバージョン変更に伴う影響が大きい**。
- 学習に使われたデータを公開しているオープンなLLMを使うことにより、**学習データの透明性**が確保できる。
- 推論モデルによって大量の推論が必要になる場合など、API利用での従量課金制による**コストの不透明さが解消**される。

LLMベンダに依存するリスク

リスク	説明
バージョン変更に伴う急な挙動変更	裏側でのモデルのバージョン変更によって、モデルの出力に急な挙動変更が生じ、LLMを利用する事業者側でプロンプトを再調整するなどの対応が必要となる。
料金改定に伴う事業コストへの影響	サービス利用料金をLLMベンダが引き上げることで、LLMを利用する事業のコストが大幅に増加するリスクがある。
サービス提供停止	• LLMベンダがサービス提供をやめてしまうことで、LLMに依存する事業の継続に影響が出る。 • 過去には、OpenAI社のコード生成用の「Codex」モデルがサポート停止を発表したことを受け、「学術論文の再現性が損なわれる」との指摘が実際にあった*。

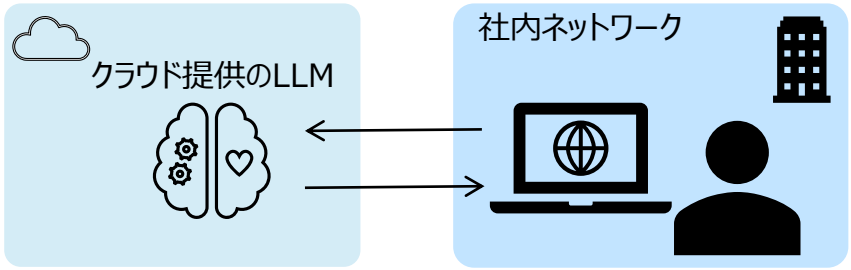
* [出所]"OpenAI's policies hinder reproducible research on language models", AI Snake Oil, Sayash Kapoor and Arvind Narayanan, https://www.aisnakeoil.com/p/openais-policies-hinder-reproducible (参照：2025/2/13)

ローカルLLMのメリット (2/2)

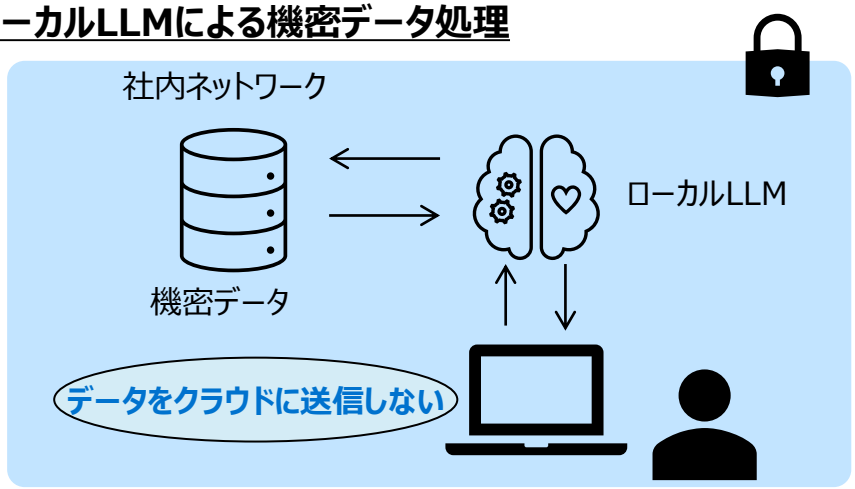
メリット③：データセキュリティ・プライバシーの向上

- ネットワーク的に閉じたシステム構成でデータを扱えるため、**情報漏洩のリスクを抑えられ**、**法規制や内部統制**への対応が容易。
- クラウド上で機密情報を扱うのは忌避されるが、ローカルLLMではデータをクラウドに送信する必要がないため、プライバシーや社外秘に関するデータなど**機密情報も扱える**。

クラウド提供のLLMによるデータ処理

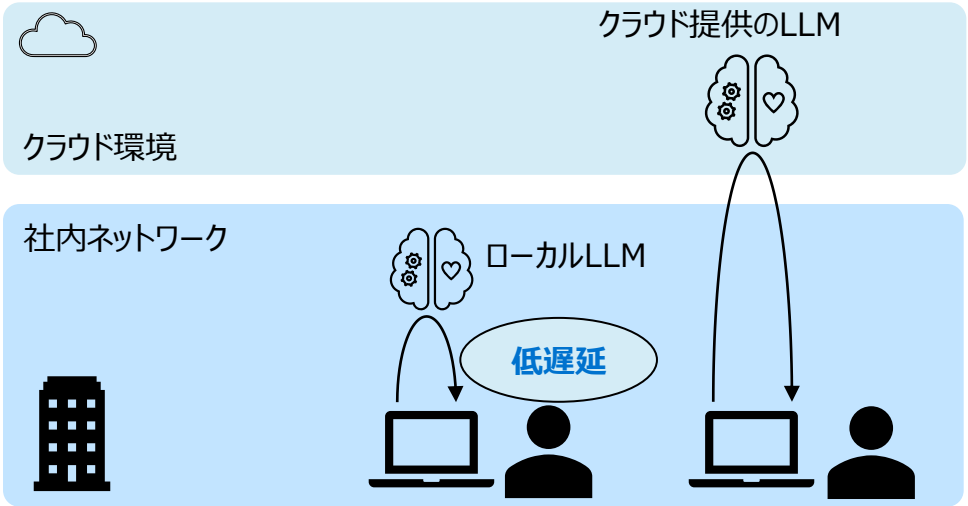


ローカルLLMによる機密データ処理



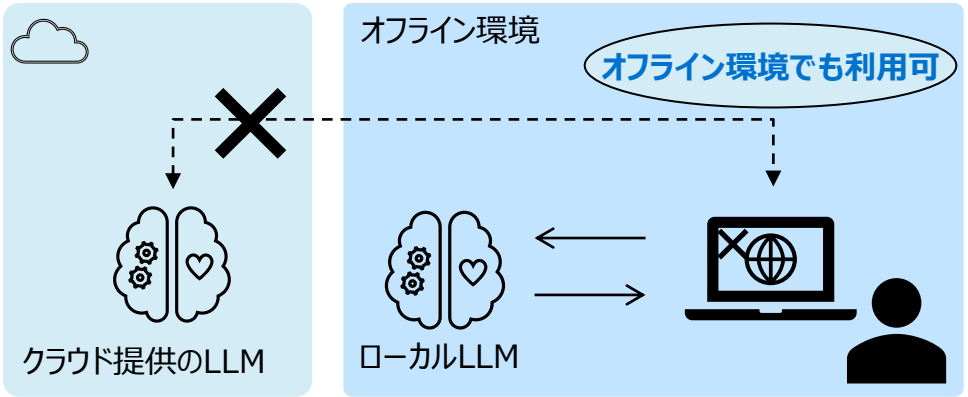
メリット④：リアルタイム処理・オフライン環境への対応

- クラウド側との通信が不要のため、クラウド提供のLLMと比較して**低遅延の処理**にも対応*できる。



*P.22で後述するように、アクセスが集中するような業務での活用は、スケーラビリティの観点からするとむしろ推論測度が遅くなり遅延が発生するようなケースもある点には注意が必要

- サーバ室内のセキュリティ対応など、インターネットに接続できないような**オフライン環境でLLMを使いたい場合**にも対応できる。



ローカルLLMのデメリット (1/2)

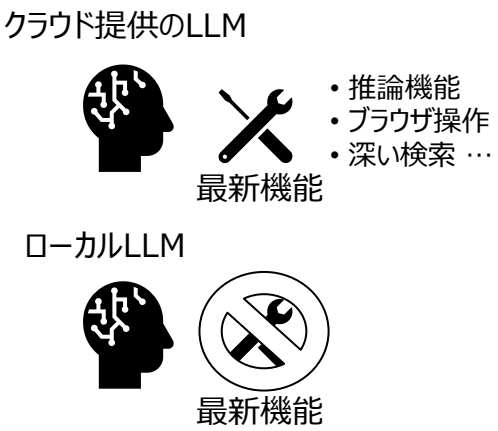
- ローカルLLMの開発・運用にあたっては考慮すべき事項もあり、デメリットとなり得る。
- 主なデメリットとして、①クラウド提供のLLMと比較して十分な性能が出ない可能性、②初期導入費用の高さ、③モデルの保守運用コストの高さ、④専門知識を持った人材確保の困難さなどが挙げられる。

デメリット①：十分な性能が出ない可能性

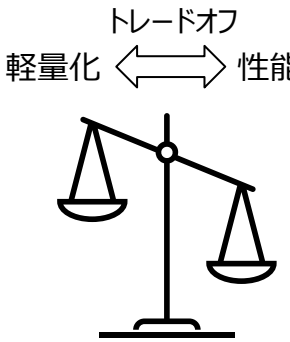
- 高性能なモデルや最先端の機能はクラウドで提供されることが多いため、**最先端の技術の恩恵を受けられにくく、性能が落ちる可能性**がある*。
- 一概に言えないが、クラウド提供のLLMと比較してローカルLLMは軽量となり、スケーリング則を考慮すると性能低下が生じやすい。導入の際は**軽量化と性能のトレードオフを考慮する必要**がある。

ローカルLLMで性能が出にくい要因

①最新技術が未導入*



②軽量化と性能のトレードオフ



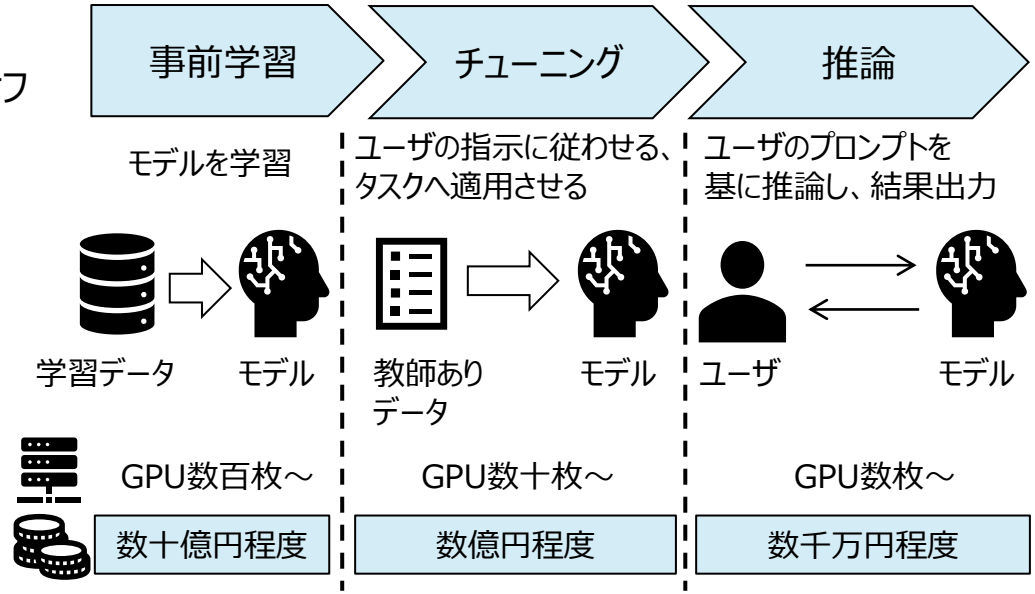
*先のページで記載したように、最近のトレンドとして、最新の技術もやや時間が経てばオープンなモデルにも取り入れられることが多く、ラグはあるもののモデルを更新すれば良くなってきている。

デメリット②：初期導入費用の高さ

- ローカルLLMにおいて必要な性能を担保するには、モデルの推論を行うのに十分なスペックのハードウェアを調達することになる。そのための専用ハードウェアやネットワーク構築に対する**高額な初期投資が必要**になる。

LLM開発・利用における初期投資金額規模の目安

ローカルLLMを開発・利用する際には、様々なフェーズ・導入方法があるが、各フェーズにおいて相応のインフラコストが必要となる。



※必要な計算リソースや金額は目安程度で、データの規模などによっても変化する。

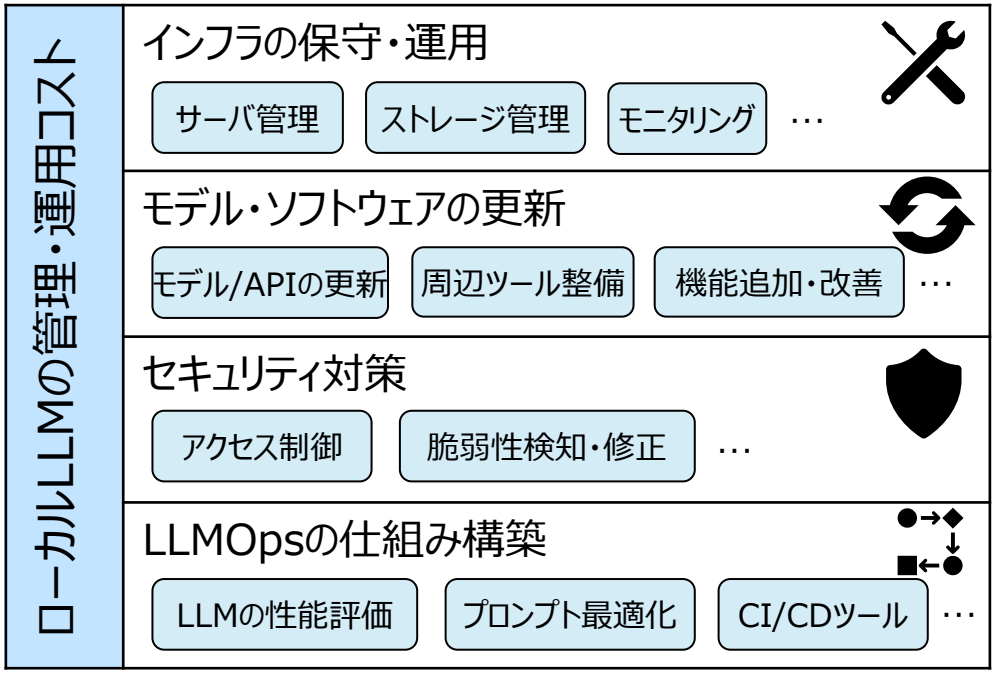
[出所]"【HPE AI Japan】HPEの生成AIオファリング", HPE AI Japan, https://community.hpe.com/t5/hpe-blog-japan/hpe-ai-japan-hpe%E3%81%AE%E7%94%9F%E6%88%90ai%E3%82%AA%E3%83%95%E3%82%A1%E3%83%AA%E3%83%B3%E3%82%B0/ba-p/7224113?profile.language=ja（参照：2025/3/17）を基に日本総研作成

ローカルLLMのデメリット (2/2)

デメリット③：モデル管理・運用コストの増大

- ローカルLLMを利用する場合には、クラウド提供のLLMを利用する場合と対比して、**モデルの管理・運用コストが発生**する。
- 具体的にはローカルLLMを利用するためのインフラの保守運用や、セキュリティ対策、モデルのアップデートといった対応が必要となる。
- また、LLMの性能をモニタリングして運用に繋げる**LLMOps*の仕組みも自社で構築する必要**がある。

＊ LLMを管理・運用する際の手法とツールを指す。



デメリット④：専門知識を持った人材確保の困難さ

- ローカルLLMの開発や運用にあたっては専門知識が必要となり、要件を満たす人材を確保する必要がある。
- LLMは新しい技術であるため、**LLMの開発・運用に関するノウハウを有する人材は現時点では極めて貴重**である。

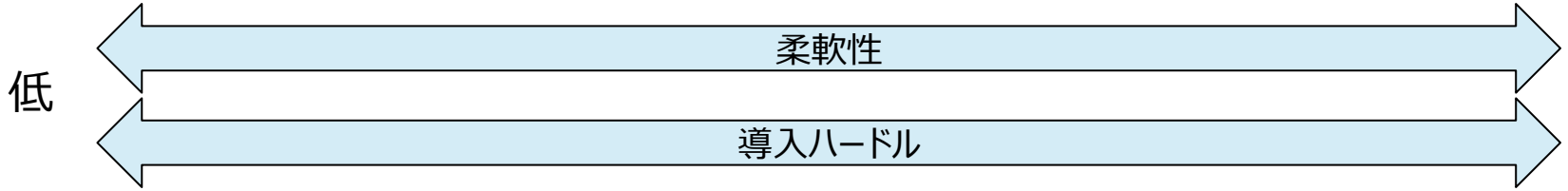
ローカルLLM開発・運用で求められる主なスキル

下表のスキルを持つ人材による開発・運用チームを編成する必要あり。

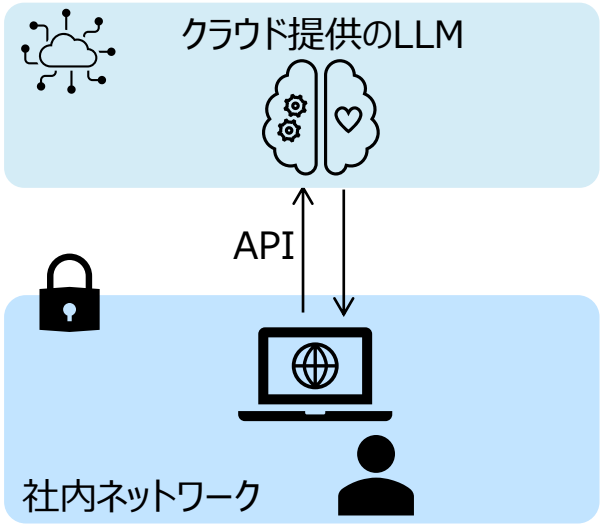
必要なスキル	具体的な説明
ディープラーニングや自然言語処理の専門知識	• Transformerなど、モデルのアーキテクチャの知識 • PyTorchなどの機械学習フレームワークの実装 • 事前学習、ファインチューニングの手法に関する知識 • 量子化、モデル圧縮技術などの知識 …
分散・並列処理に関する専門知識	• ハイパフォーマンスコンピューティングの知識 • GPUを利用した並列処理の実装 • 並列分散学習、推論コードの設計・実装 …
データエンジニアリングと前処理に関する専門知識	• 学習データセット作成に関する知識（大量なデータの収集、クレンジング、トークナイゼーションなど） • データ処理パイプラインの実装 • データの品質管理に関する知識 …
システム運用・インフラ管理に関する専門知識	• Linuxサーバやオンプレミスのクラスタ運用に関する知識 • ハードウェアのリソース最適化、トラブルシューティング • ネットワーク構築・管理に関する知識 • セキュリティ対策に関する知識 • 性能モニタリング、LLMOpsに関する知識 …

LLMのシステム構成パターン

- LLMを導入する際、①ローカルLLMを導入せずにクラウド提供のLLMのみを利用するパターン、②ローカルLLMのみを利用するパターン、③クラウド提供のLLMとローカルLLMをハイブリッドで利用するパターンの3種類のパターンがある。
- 前章で見たローカルLLMのメリット・デメリットを考慮し、目的やユースケースに応じた最適な導入パターンを検討すべき。



①クラウド提供のLLMのみ利用するパターン



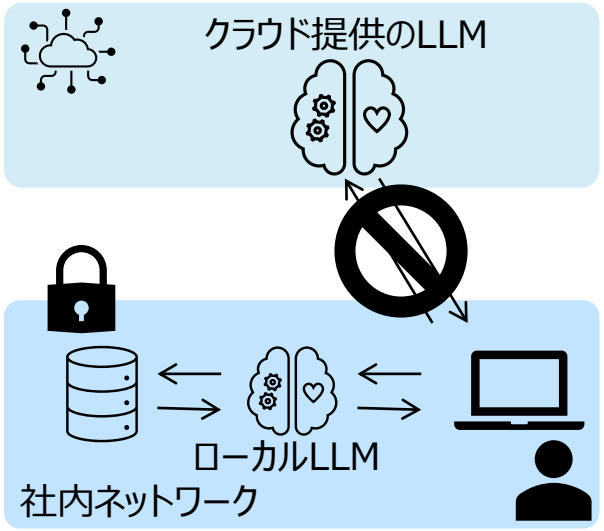
メリット

モデルの管理運用はLLMベンダに任せ、アプリケーションのみ社内で管理運用

デメリット

カスタマイズ性は低く、LLMベンダに依存

②ローカルLLMのみ利用するパターン



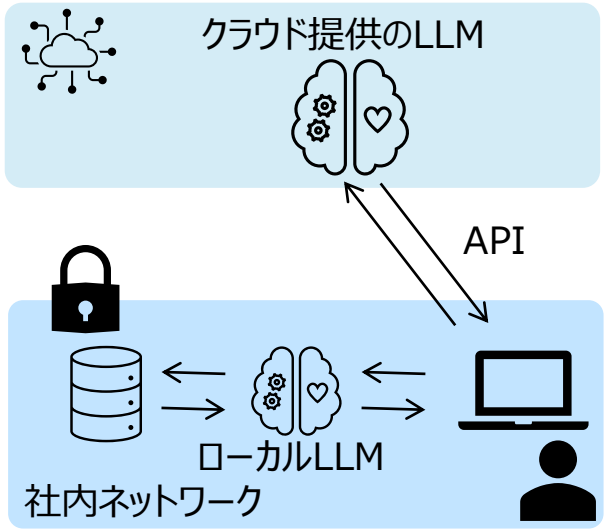
メリット

ローカルLLMのメリットを活かしたシステムを構築できる

デメリット

モデルの管理運用や、システム設計と構築を自社で実施する必要がある

③ハイブリッド式に利用するパターン



メリット

機密情報はローカルで処理し、それ以外の高度な処理をクラウド上で行うなど、それぞれの強みを活かしたシステムを構築できる

デメリット

ネットワークやシステム構成が複雑化し、システム全体の管理運用コストが増加する

業界ごとの主なローカルLLMの活用ユースケース

●顧客情報や高度な専門知識を活用することも多い金融業界、独自データやノウハウを有することが多い製造業、病歴情報などセンシティブな情報を扱うことも多い医療業界などではローカルLLMの活用ユースケースが想定される。

業界	ユースケース例	説明
金融	機密顧客データを活用した金融サービス実現	顧客の資産状況や取引履歴、投資嗜好などを基にした高度なパーソナライズによる金融サービス実現
	社内文書・レポートの要約や分析支援	膨大な内部報告書や議事録などを有効活用することにより、迅速かつ合理的な意思決定を支援
	金融教育・トレーニング支援	自社独自のノウハウを基にした金融教育やトレーニングを実施
	金融法規制対応支援	膨大な規則やガイドライン、コンプライアンス文書を確認し、最新の法規制への社内対応を支援
医療	電子カルテの要約	各患者の電子カルテの情報を要約し、医師が短時間で患者の症状経過や病歴、治療履歴などを漏れなく把握できるように支援
	遠隔診療における問診	患者が入力した問診情報を基に、症状の緊急度や受診すべき診療科についてアドバイス
	医師の診断支援	患者の問診情報から、考えられる疾患や治療方針の候補を提示し、医師の診断や治療計画作成を支援
	医療安全のための解析・改善提案	インシデントレポートの作成支援や、過去のレポートを解析して再発防止策や安全対策に向けた改善点の洗い出しを支援
製造業	設備の予知保全計画作成支援	自社設備のセンサデータや過去の故障データなどを基に設備の故障予知を行い、メンテナンスの計画作成を支援
	製造工程マニュアルや手順書作成支援	製造工程や機器操作に関する自社独自の情報を取り入れた現場向けのマニュアル・手順書作成により製造工程における品質を担保
	設計データ・技術文書の分析によるノウハウ抽出	自社が保有する設計データや技術文書などの内容を要約し、独自のノウハウを抽出し熟練者の技術を後継者に継承
	品質管理の自動化	製造品の検査画像を活用し、不良品の早期発見や品質レポート作成を支援

ローカルLLMの活用事例　～金融～

- 海外の大手金融企業では、LLMを独自開発してバックオフィス業務を中心に利用している例が見られる。
- 日本国内ではローカルLLMの事例がまだ多くはないものの、金融領域特化のLLMを開発する動きが見られる。

LLM Suite (JPMorgan Chase)

- JPMorgan Chaseが**データのセキュリティ確保**とLLMによる**ハルシネーション抑止**を目的として、**自社独自に開発**したLLM
- 社内版ChatGPTのような**汎用的な生成AI**で、5万人以上の従業員向けの「リサーチアナリスト」としての機能を提供
- 詳細は公開されていないが、機密の財務データを管理する「Connect Coach」や「SpectrumGPT」といった特定タスクに特化した**専用ツールとも連携**している。

[出所] "JPMorgan (JPM) Launches New AI Research Analyst Chatbot", Yahoo Finance News, https://finance.yahoo.com/news/jpmorgan-jpm-launches-ai-research-134500926.html?guccounter=1 (参照：2025/3/28)

みずほ特化型モデル (みずほFG×NTTデータ)

- NTTのLLM「tsuzumi」を基盤とした**「みずほ特化型モデル」**の開発に向け、NTTデータグループとの共同研究を発表(2024/12/18)
- 新人研修資料や社内固有のデータ、特定業界の企業データなどを学習させ、みずほFGや金融業界に対する**ドメイン知識を獲得**。
- 今後、開発したモデルをさまざまなユースケースに適用させ、業務特化のAIエージェントを開発していく予定とのこと。

[出所] "〈みずほ〉とNTTデータグループ、生成AI活用に向けた共同研究契約を締結 —NTT版LLM「tsuzumi®」のチューニングを通じた「〈みずほ〉特化型モデル」の構築—", みずほフィナンシャルグループニュースリリース, https://www.mizuho-fg.co.jp/release/pdf/20241218release_jp.pdf (参照：2025/3/28)

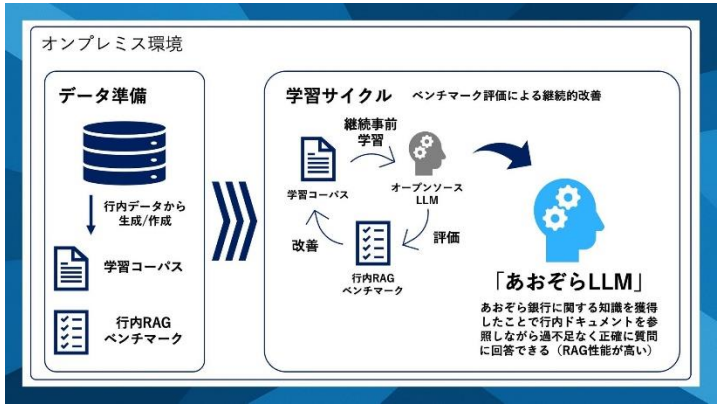
DeepSeek活用（中国工商銀行；ICBC　など)

- 中国最大手のICBCは、DeepSeekのオープンモデルのローカル環境へのデプロイと既存アプリとの接続完了を発表(3/8)。
- マルチモーダルで多様なタスクが協調できる多層フレームワークも実装済みであり、20以上の主要なビジネス領域で200を超えるアプリケーションをLLMにより強化している。
- 国策上、中国の金融機関はOpenAIなどが提供する米国のAIに頼ることは難しい。そのため、ICBCの他にも複数の金融機関で顧客サービス高度化や内部業務支援、与信審査などにDeepSeekを利用しており、中にはオンプレ環境で運用しているものもあると見られる。

[出所] "Multiple banks deploy DeepSeek AI models for customer service, credit approval", Global Times, 2025/3/13, https://www.globaltimes.cn/page/202503/1330054.shtml (参照：2025/4/25)

あおぞらLLM(※仮称) (あおぞら銀行×neoAI)

- あおぞら銀行は、スタートアップ企業neoAIと共同で、金融分野かつ行内情報に特化した**「あおぞらLLM」**をオンプレミス環境で開発。
- 法人・リテール業務の事務規定管理業務を想定し、行員の実際のタスクや運用フローを踏まえたチューニングによる精度向上を確認。



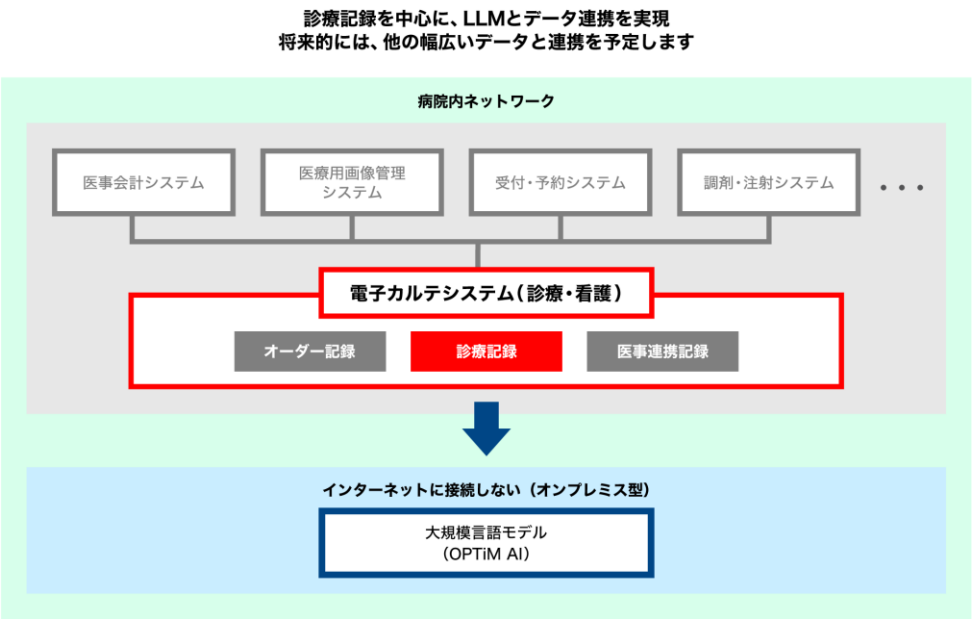
[出所]あおぞら銀行 x neoAI オンプレミス型次世代AI基盤構築に向けて、金融・行内特化LLMを開発", PR TIMES, https://prtimes.jp/main/html/rd/p/000000026.000109048.html (参照：2025/4/25)

ローカルLLMの活用事例　～医療～

● 医療業界において、患者のプライバシーを保護しつつ、医療事務を支援するためにローカルLLMを導入する事例が見られる。

織田病院における臨床現場への導入

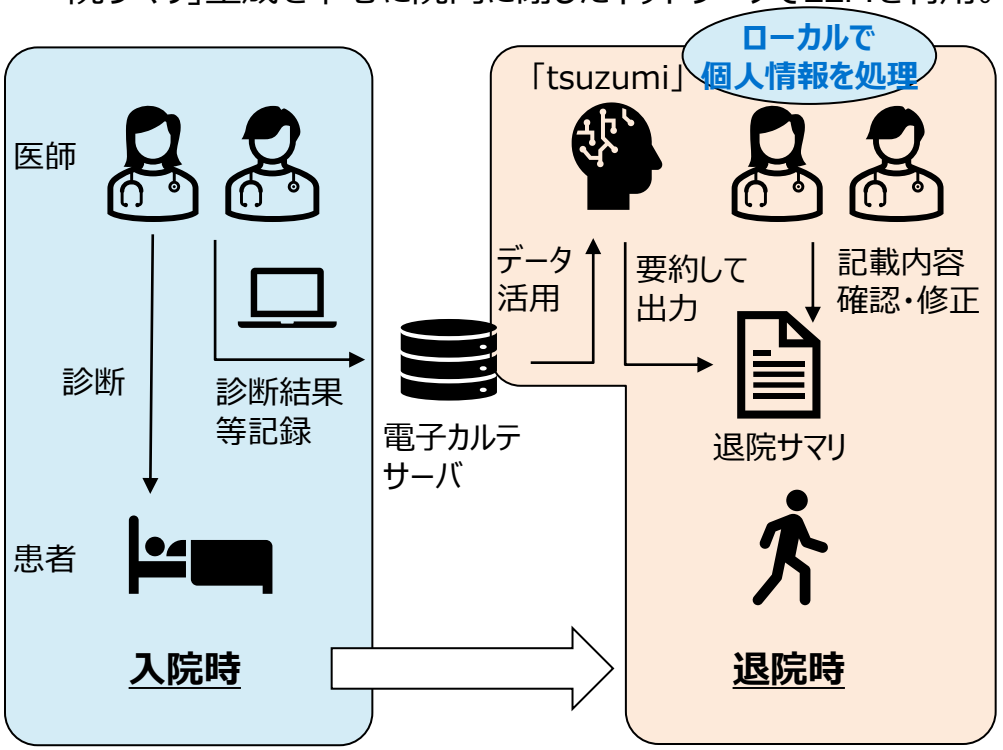
- 佐賀県の祐愛会織田病院、株式会社オプティム、株式会社シーエスアイの3社による臨床現場でのローカルLLM導入事例
- 外部ネットワークに接続しないLLMを構築し、電子カルテシステムと連携して患者の個人情報を保護しつつ医療従事者を支援
- 「入退院時看護サマリー」の自動生成を中心に取り組み
- 将来的には幅広いデータとの連携を予定



[出所]"【国内初※1】外部ネットワークへの接続を必要としないセキュアな大規模言語モデル（LLM）「OPTiM AI」を発表電子カルテと連携し、臨床現場にオンプレミス導入開始～医師・看護師・病院関係者の働き方改革を支援～", OPTiM プレスリリース, https://www.optim.co.jp/newsdetail/20240329-pressrelease-01 (参照：2025/3/28)

三重大学医学部附属病院における実証実験

- NTT西日本と協力して、NTTのLLM「tsuzumi」を活用した電子カルテの要約に関する実証実験を開始(2024/11/1)。
- 電子カルテのデータを利用し、入院中の治療経過をまとめた「退院サマリ」生成を中心に院内に閉じたネットワークでLLMを利用。



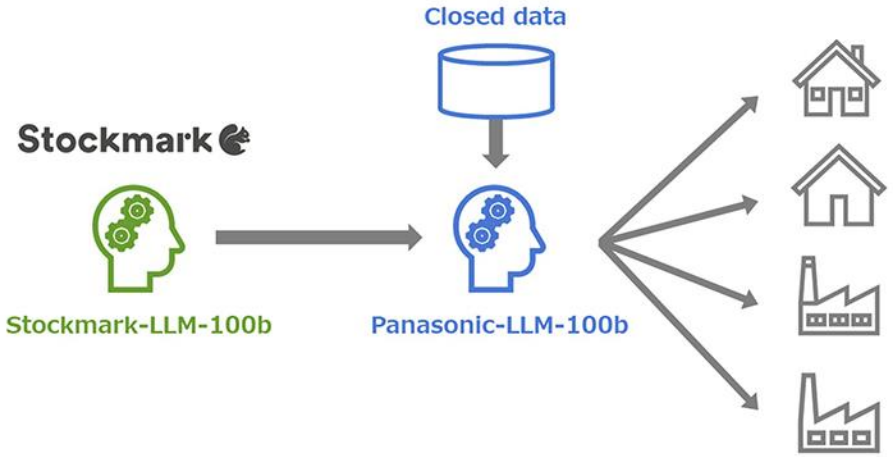
[出所]"三重大学とNTT西日本が医療DX推進に向けた包括連携協定を締結NTT版LLM「tsuzumi」による電子カルテ要約の実証実験を開始～", NTT西日本ニュースリリース, https://www.ntt-west.co.jp/news/2411/241101a.html (参照：2025/3/28)
を基に日本総研作成

ローカルLLMの活用事例　～製造業・官公庁～

- 製造業では、設計データなど独自データを有することが多く、その知識を秘匿するためにローカルLLM活用が有効となる場合がある。
- 官公庁では組織ごとに独自のデータや業務フローを有することも多く、個人情報を取り扱うことも多いためローカルLLM活用が有効となる事例が多く存在する。

パナソニックHD×ストックマーク

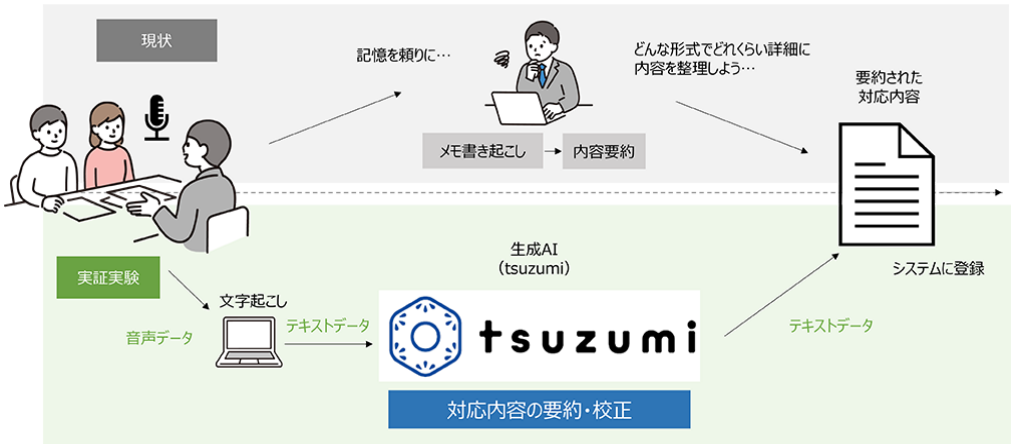
- ストックマーク社のモデルを基に、パナソニックグループ**独自のデータで追加学習**させた**自社専用のLLMを構築**すると発表。企業で利用する独自のLLMとしては**国内最大レベル**とされる。
- 今後は、工場などの**オフライン環境**での利用も見据えた**小型化も図る**とのこと。



[出所]"パナソニックHDとストックマーク、国内最大規模（1,000億パラメータ）の独自日本語LLM「Panasonic-LLM-100b」開発で協業", パナソニックニュースリリース, https://news.panasonic.com/jp/press/jn240702-3 (参照：2025/3/28)

山口県×NTT西日本

- NTTの「tsuzumi」を利用して、山口県庁の庁内業務への生成AI適用、および機微データを扱う業務への実証実験を2024年10月から開始。
- オンプレミス環境の小型GPUサーバ上で、機微なデータを扱う業務の対応記録の要約・校正や、各種業務マニュアルの検索・要約等の実証を行う。
- 業務に特化したチューニングまでを行い、評価する予定とのこと。



[出所]"機微データを扱う業務への大規模言語モデルtsuzumi活用に関する実証実験を開始", NTT西日本 ニュースリリース, https://www.ntt-west.co.jp/news/2409/240917a.html (参照：2025/3/28)

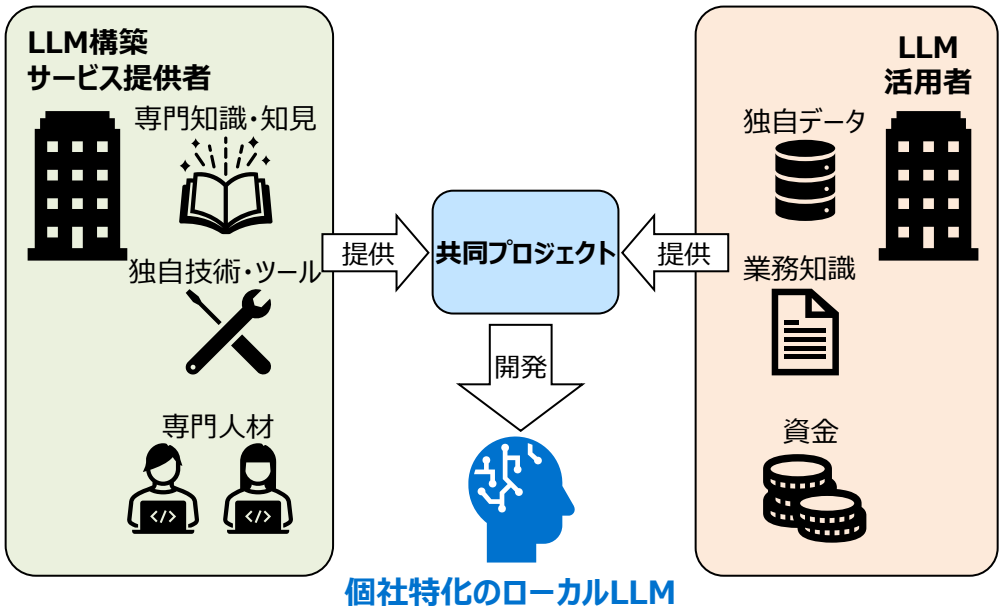
ローカルLLM構築サービス

- 様々な企業・業界においてローカルLLMを活用するニーズがある一方で、**導入に際してはハードルが高い**。
- これを受けて、ローカルLLMを構築するためのサービスを提供し始めている企業が存在する。こうしたサービス提供者と組むことで、独力ではローカルLLMの開発・運用が難しい企業においても導入が進む可能性がある。

ローカルLLM構築に向けた共同での取り組み

- ローカルLLM構築に関する知見を有するLLM構築サービス提供者と、自社独自データを有するLLM活用者が**共同でプロジェクト**を進めることで**個社特化のLLMを開発**する取り組みが増加。
- LLM活用者単体ではローカルLLMの開発・運用が難しい場合でも、知見を持った事業者と組むことで**導入障壁が下がり、ローカルLLM活用が進む可能性**がある。

個社特化のローカルLLM開発の共同プロジェクトのイメージ図



ローカルLLM構築サービスの例

- 大手ITベンダやスタートアップ企業など、様々な企業でローカルLLM構築サービスを提供する事業者が出て来ている。
- ローカルLLM活用には、モデルの開発だけでなく**運用も重要**となってくるため、**モデルの保守・運用を担うサービス**も増加する可能性がある。

分類	サービス名	サービス提供者	説明
モデル開発に特化	独自LLM開発支援プログラム	ELYZA	・2023年から提供する、独自のLLMを開発するサービス ・事後学習基盤構築から支援
モデル運用に特化	美琴 powered by cotomi	NEC・大塚商会	・NEC製「cotomi」がプレインストールされたサーバを提供 ・保守運用支援も行う
	GBase On-premises	Sparticle・SB C&S	・オンプレミス専用のRAGを提供するサービス ・2024年12月にSB C&Sがディストリビューター契約締結
モデルの開発・運用をサポート	RICOH オンプレLLM スターターキット	リコージャパン	・リコー製のLLMを利用 ・環境構築から運用支援まで対応
	"ローカルLLM"を用いた生成AI基盤構築・導入支援	neoAI・マクニカ	・2025年3月に発表された両社の協業による支援サービス

ローカルLLMの技術課題(1/2)

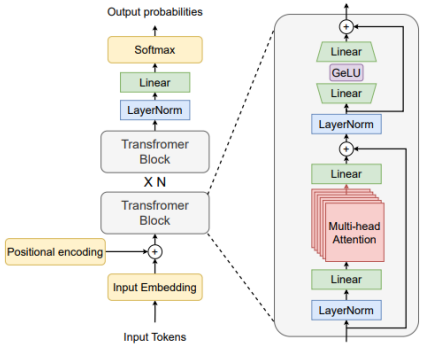
モデルの小型化と性能の両立

- 計算リソースに要するコストを抑えられるため、ローカルLLMには小型のモデルの方が良いが、**十分な性能が出ないことが多かった**。
- 最近は推論モデルの台頭などもあり、小型モデルでも実用的になりつつあるが、今後は、**新たなアーキテクチャの模索**などによりモデルの小型化と性能を両立させる取り組みが一層進むと考えられる。

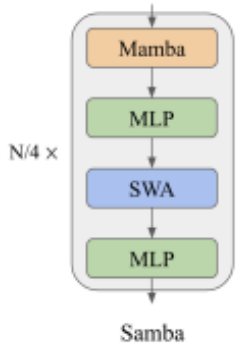
状態空間モデルとの融合による小型の高性能モデル開発

Preferred Networksは、状態空間モデルを取り入れたアーキテクチャと高品質なデータセットによりPLaMo2を開発。特に、PLaMo2 8Bは前バージョンのPLaMo 100Bと同等の性能を示し、**高性能なまま大幅なサイズ縮小**を実現。

PLaMo 100B
(Decoder-only transformerベース)



PLaMo2 8B
(Sambaベース)



※両者ともにアーキテクチャ図は別論文から引用したイメージ図である点に注意

[出所] G. Chalvatzaki et al., "Learning to Reason over Scene Graphs: A Case Study of Finetuning GPT-2 into a Robot Language Model for Grounded Task Planning", https://arxiv.org/abs/2305.07716 (参照：2025/4/16)のFig.2より抜粋

[出所] L. Ren et.al, "Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling", https://arxiv.org/abs/2406.07522 (参照：2025/4/16)のFig.1より抜粋

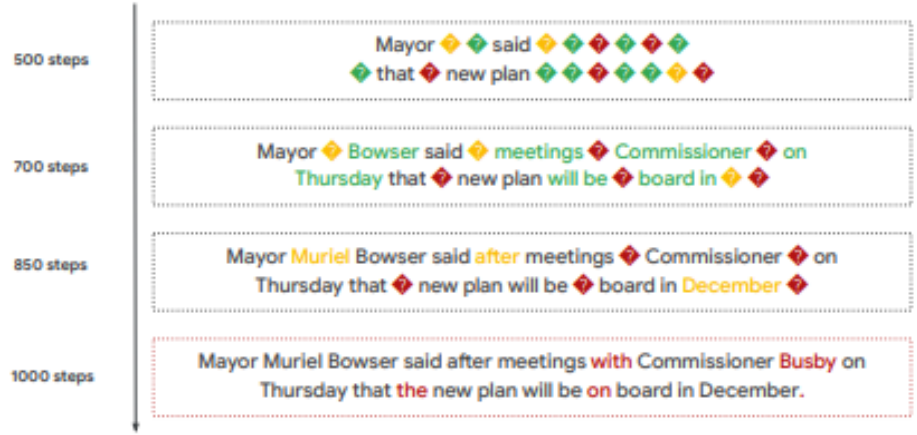
推論速度の向上

- オープンなLLMでは、クラウドで提供されるLLMと比較し推論速度が低下することも多いが、推論速度向上により、リアルタイム処理が重要なケースにより活用されていく。
- vLLM*1やNvidia Dynamo*2など、LLMで高速な推論を行うためのライブラリやミドルウェアが出て来ており、こうしたライブラリが今後拡充されていくものと考えられる。
- また、拡散モデルベースの高速な言語出力ができる実用的なLLM*3が出てくるなど、新たな推論の仕組みにも期待がかかる。

*1 vLLM; https://docs.vllm.ai/en/latest/ (参照：2025/4/18)
*2 Nvidia Dynamo; https://www.nvidia.com/ja-jp/ai/dynamo/ (参照：2025/4/18)
*3 Mercury; https://www.inceptionlabs.ai/news (参照：2025/4/18),
Gemini Diffusion; https://deepmind.google/models/gemini-diffusion/ (参照：2025/5/22)

拡散モデルベースのLLM

画像生成で多用されている拡散モデルをベースとしたLLMも出現。従来の自己回帰モデルのLLMのように順次単語を生成するのではなく、文章全体を並列して生成する仕組みのため非常に高速な文章生成ができる。



[出所] J. Shi et.al, "Simplified and Generalized Masked Diffusion for Discrete Data", https://arxiv.org/abs/2406.04329 (参照：2025/4/16)

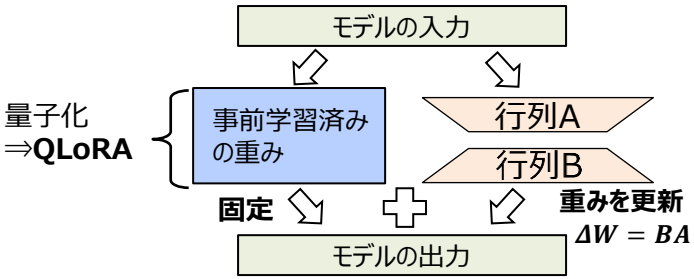
ローカルLLMの技術課題(2/2)

効率的なファインチューニング・学習手法の確立

- 現在はLLMに業界固有、もしくは企業固有の知識を学習させるファインチューニングのコストが非常に高い。そのため、モデル自体を更新せずに済むRAGなどの方法が積極的に取り組まれているが、モデル自体をチューニングした方が良い場合もある。
- 今後、ファインチューニングや学習コストを下げる手法が確立されれば、独自データによるLLMのカスタマイズが進む可能性がある。

LoRA/QLoRA

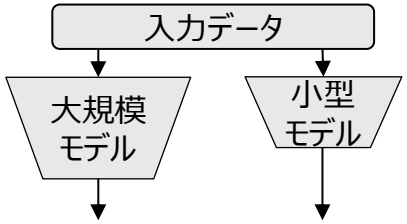
LoRA(Low-Rank Adaptation)とは、事前学習済みの重みを固定し、一部の層に低ランクの行列を追加して、この行列のみを学習させる手法。
更に、量子化手法を組み合わせたQLoRAにより、計算に必要なメモリを削減し効率的なファインチューニングを実現



[出所] E. Hu et.al, "LoRA: Low-Rank Adaptation of Large Language Models", https://arxiv.org/abs/2106.09685
(参照：2025/5/1)のFig.1を基に日本総研作成

知識蒸留(Knowledge Distillation)

事前に学習済みの大規模なモデルを使って教師データを作成し、より小型のモデルを学習させる手法。元となるモデルよりも少ないリソースで学習できる。

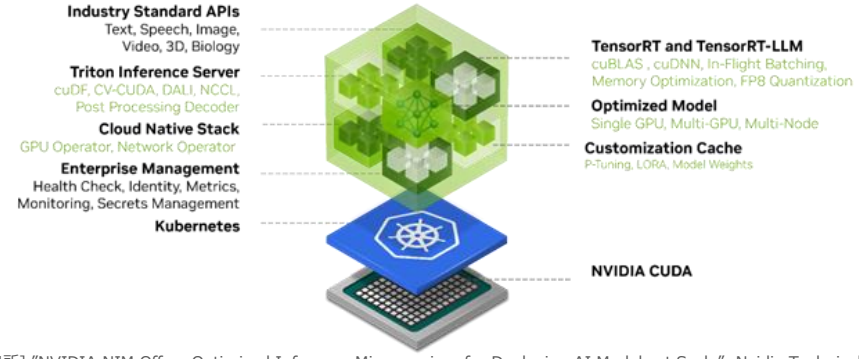


モデル管理手法・ツールの確立

- ローカルLLMを活用する際には、モデルの開発だけでなく、自前でのモデルの運用が必要。そこで、LLMを効率的に管理・運用する(LLMOps)技術・支援ツールの開発が進んでいる。
- 効率的なLLMOps実現のための手法やツールが確立されてくれば、ローカルLLMの導入が一層進むものと考えられる。

NIM(Nvidia Inference Microservices)

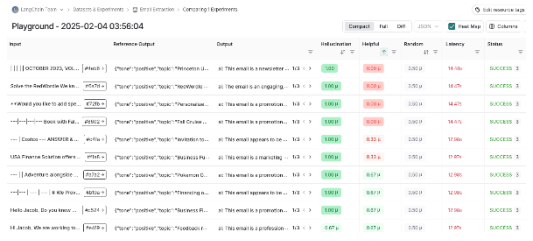
Nvidiaから発表された、LLMの推論を実行するためのマイクロサービス群。
GPU処理に最適化されたコンテナ形式で利用でき、モデルの多様な環境へのデプロイやGPU処理性能に優れている。
Kubernetesとの互換性もあり、スケーラビリティも担保できる。



[出所] "NVIDIA NIM Offers Optimized Inference Microservices for Deploying AI Models at Scale", Nvidia Technical Blog, https://developer.nvidia.com/blog/nvidia-nim-offers-optimized-inference-microservices-for-deploying-ai-models-at-scale/
(参照：2025/5/2)

LangSmith

LLMの性能をモニタリングするためのプラットフォーム。プロンプトやモデル出力・実行ログのトレース、カスタムダッシュボードの作成などの機能が存在する。



※イメージ図

[出所] LangSmith Documents, https://docs.smith.langchain.com/evaluation/concepts
(参照：2025/5/2)

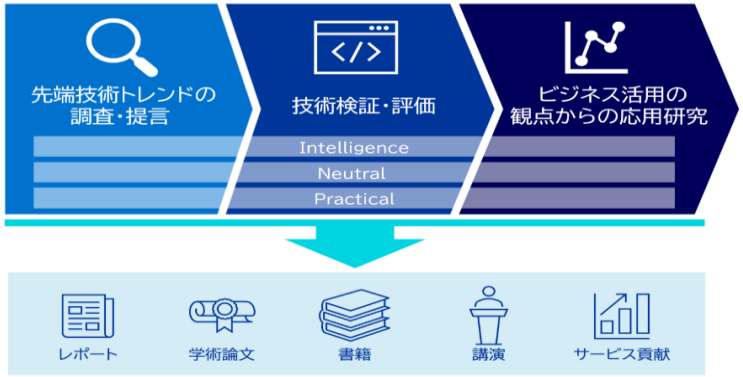
今後の展望

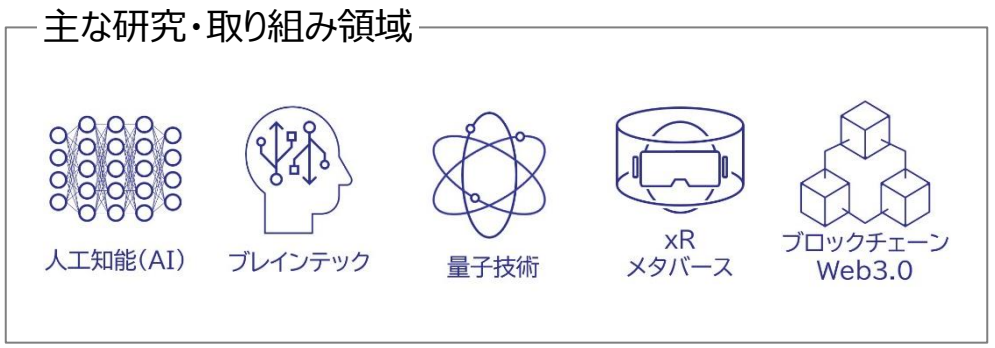
- オープンなLLMの軽量化や高性能化が進み、ローカルLLMとして実用的なモデルが増加する。結果として、ローカルLLMを利用するモチベーションや意義が高まっていく。
- クラウド提供のLLMとローカルLLM、エッジデバイスでのSLMのハイブリッド活用など、**多様なシステム構成が検討・実現**されていく。

ローカルLLMを取り巻く状況	今後数年で起きる変化	
独自データ・モデル活用による競争優位性確保 モデルの小型化と精度の両立 機密情報のLLMへの活用促進 多様な環境でのLLM活用促進 推論速度の向上 効率的な学習・ファインチューニング手法の確立 モデル管理手法・ツールの確立	汎用的なモデルと特化型モデルの開発の二極化	• クラウド提供のLLMのように、**巨大で汎用的なモデル**を開発していくアプローチと、**小型で個別の業務や自社業務に特化したモデル**を開発していくアプローチの二極化が進む。 • ほとんどの企業・組織では、後者のアプローチを模索していくと考えられる。**高性能なオープンなモデルによって、独自のモデル開発を後押しするものと期待**される。
	ローカルLLM導入の段階的拡大	• ローカルでの自社独自データ活用や、業務特化型モデルの活用により自社の強みを見出したい企業が、ローカルLLM導入を進めていくものと考えられる。 • まだ限定的なユースケース・事例しかないが、**今後様々なユースケースでの活用が期待**され、段階的に拡大していく。
	ハイブリッドなシステム構成でのLLM活用	• システムアーキテクチャ同様、**LLMのシステム構成においても万能な構成は存在しない**。そのため、クラウド上のLLMとローカルLLM、エッジデバイスでのSLMのそれぞれの強みを活かした**多様なシステム構成が検討**され、**ユースケースに応じた使い分け**がされていく。 • 今後流行するAIエージェントにおいても、ローカルLLMを活用したものが増加していくものと予想される。

先端技術ラボ

先端技術を活用したITサービスの創出に向けた技術の目利き役として、「先端技術トレンドの調査・提言」、「技術検証・評価」、「ビジネス活用の観点からの応用研究」に取り組んでいます。





当社ホームページの **特集サイト** では、ＩＴ分野における先端技術の調査レポート、及び所属する部員のプロフィール詳細がご覧いただけますので、ぜひご参照ください。

本レポート執筆者へのメディア取材や講演などに関するご相談につきましては、当社ホームページの **問い合わせフォーム** よりご連絡ください。

株式会社日本総合研究所

日本総研は、シンクタンク・コンサルティング・ITソリューションの3つの機能を有するSMBCグループの総合情報サービス企業です。
東京本社　〒141-0022　東京都品川区東五反田2丁目18番1号　大崎フォレストビルディング
大阪本社　〒550-0001　大阪市西区土佐堀2丁目2番4号


日本総研
The Japan Research Institute, Limited