到来するデータ主権に応えるLazarus AIが示すローカルLLMの可能性

How Lazarus AI Illustrates the Potential of Local LLMs in Response to the Rise of Data Sovereignty

青木岐夫

要　約　生成AIの普及に伴い，シャドーAIや自律的AIエージェントによる新たなセキュリティリスクが顕在化する中で，データを外部に送信せず企業の管理下で運用できるローカルLLMへの関心が高まっている．国内事例が示すように，ローカルLLMは実用段階にあり，継続的な運用改善を前提に，PoCから段階的に導入することが肝要である．ユニアデックスでは，抽出や精査に特化したLazarus AIを，機密性と検証可能性を担うハイブリッド運用の中核として位置づけている．

Abstract　As the adoption of generative AI brings emerging security risks like “shadow AI” and autonomous AI agents to the forefront, interest is growing in local LLMs that can operate under corporate control without transmitting data externally. Domestic use cases demonstrate that local LLMs have reached a practical stage of maturity. Therefore, a phased implementation starting with a Proof of Concept (PoC) and building upon a foundation of continuous operational improvement is essential. Uniadex positions Lazarus AI, a tool specialized in data extraction and scrutiny, as the core component of a hybrid operational model designed to ensure confidentiality and verifiability.

1. はじめに

2026年現在，企業への生成AI導入はすでに「始めるかどうか」の議論ではない．総務省の「令和7年版 情報通信白書」によれば，何らかの業務で生成AIを利用している日本企業は55.2%に達している[1]．文章作成，メール対応，業務フローの自動化にLLMが組み込まれる状況は，大企業だけでなく中堅・中小企業にも広がっている．

その一方で，普及の裏側では見えにくいひずみが蓄積している．CyeraとCybersecurity Insidersによる「2025 State of AI Data Security」調査[2]では，回答者の76%が自律的に行動するAIエージェントはセキュリティ確保が最も難しいと認識しており，70%が公開LLMへの外部プロンプトを高リスクと捉えている．さらに，同レポートは，調査対象企業の83%がAIを業務利用している一方，利用実態を可視化できている企業は13%に留まることも示している[2]．

これらは欧米企業を対象とした調査ではあるが，入力データが海外サーバーで処理されうるクラウドAIの構造は日本企業にも共通する．そのため，特に金融，医療，防衛などの規制産業では，機密データを組織の管理下に保持したまま利用できるAI基盤として，オンプレミス[*1]のデータセンターやプライベートクラウド[*2]への関心が高まっている．背景にあるのは，機密データをどこまで自社の管理下に置くかというデータ主権[*3]と，それに適したAI基盤を

どう整備するかという課題である．

本稿では，こうした背景を踏まえ，AIエージェント[*4]時代に顕在化したリスクとローカルLLMへの関心が高まる要因を整理し，実現アプローチの違いを比較する．その上で，ローカルLLMの中でも「生成」ではなく「抽出・精査」に特化したLazarus AIに着目し，同製品の技術的な特徴を紐解きながら，企業の機密データを安全に活用し，日々の意思決定を支える新たな基盤としてどのような価値をもたらすのかを検討していく．まず2章で生成AIの普及により顕在化したセキュリティリスクについて触れた後，3章でローカルLLMの必要性，4章でローカルLLMの実現アプローチを述べる．5章ではローカルLLMの段階的導入の重要性，6章ではコスト構造を説明する．7章でLazarus AIの技術的な特徴，8章でクラウドAIと併用するハイブリッド運用について述べる．

2.　AIエージェント時代が生み出した新たなセキュリティリスク

本章では，生成AIの普及に伴って顕在化したリスクを，シャドーAI，AIエージェントの自律性，そして禁止策の限界という三つの観点から整理する．

2.1 「シャドーAI」という構造的問題

シャドーAIとは，IT部門や経営層が関知・承認していないにもかかわらず，従業員が業務で独自に使用しているAIサービスを指す．かつての「シャドーIT」の現代版と言えるが，生成AIには入力データがモデルの学習に再利用され，外部へ流出するという特有のリスクが存在する．

独立行政法人情報処理推進機構（IPA）[3]や総務省[4]が警鐘を鳴らすように，従業員は利便性を優先し，組織の規定を逸脱して未承認ツールを使い続ける傾向にある．また，Gartner® も四半期ごとのEmerging Riskレポートで，シャドーAIがリスクマネジメント責任者にとって最重要関心事の一つであると述べている[5]．具体的には，企画書作成のために売上データをAIに貼り付けたり，契約書のレビューを外部サービスに依頼して機密情報を入力したりするケースが日常的に発生している．

Oktaが2026年2月に発表した調査[6]では，ITリーダーの86%がAIエージェントを「ミッションクリティカル」または「非常に重要」と位置付ける一方，データ漏洩・流出（83%）や過剰な権限付与（80%）を導入上の主要な懸念として挙げており[6]，AIの「必要性」と「危険性」のジレンマが浮き彫りになっている．「生成AIは危険だから全面禁止」という対応は，逆効果になる．NTTドコモビジネスが公開する資料[7]が端的に指摘するように，IT部門が承認するツールがなければ，従業員は個人的なアカウントで個人スマートフォン経由のAI利用を続けるだけである．禁止措置は見かけ上のリスク低減に過ぎず，実態としてはリスクの「不可視化」を招く．

2.2　AIエージェントの自律的行動がもたらすリスク

問題をさらに複雑にしているのが，AIエージェントの台頭である．単に質問に答えるチャットボットではなく，カレンダーを操作し，メールを送信し，外部APIを呼び出し，データベースにアクセスするような「行動する存在」としてのAIエージェントが，企業の業務フローに組み込まれ始めている．

トレンドマイクロが発表した「AIエージェントと脆弱性 PART 1[8]」は，この状況の危うさを具体的に示す．AIエージェントが外部のWebサイトを参照したり，外部ツールと連携したりするプロセスの中で，間接的なプロンプトインジェクション攻撃*5のリスクがある．悪意を持って設計されたWebページやドキュメントに隠された命令を読み込んだAIエージェントが，ユーザーの意図とは無関係に機密情報を外部へ送信してしまうシナリオは，もはや理論上のリスクではない．

より根本的な問題として，OWASPが公開する「2025年版LLMアプリケーションのトップ10リスク[9]」には，プロンプトインジェクション，機密情報の漏洩，サプライチェーンの脆弱性，データポイズニング*6が上位にランクされている．これらのリスクは，LLMが外部（クラウド）に存在することによって根本的に対処しにくくなる構造的な問題を抱えている．

2.3 「全面禁止」という解決策の限界

2.1節で述べた通り，「生成AIを禁止すればよい」という発想は逆効果になる．しかし，このAIの「必要性」と「危険性」のジレンマには別の解決策がある．それは，従業員が安全にAIを利用できる環境を企業自身が提供することである．すなわち，LLM基盤そのものを自社の管理下に置くというアプローチである（図1下「整備アプローチ」）．これにより，シャドーAIの動機を抑制し，従業員が承認済みのAIツールを利用しやすくなる．

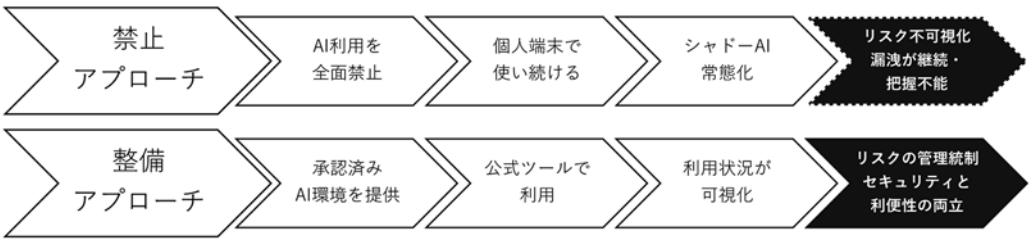

図1　生成AIの禁止措置と利用環境整備の構造的な相違
（参考文献[3][4][7]を踏まえて著者が作成）

3.　LLMの定義と普及加速の背景

本章では，ローカルLLM（Local Large Language Model）の定義と特徴，普及が加速した背景を述べる．

3.1　ローカルLLMとは何か

ローカルLLM（Local Large Language Model）とは，クラウド上の外部サーバーに常時依存せず，オンプレミス環境やプライベートクラウドなど企業の管理下で運用する大規模言語モデルを指す．NTTPCが公開する資料[10]では，クラウドLLMとの違いを「データを外部に送信することなく，マシン内で処理を完結できるため，機密情報の漏えいリスクを大幅に抑えられる」と整理している．重要なのは，適切な構成を取ることで，外部送信経路を限定または遮断しやすい点にある．

企業がChatGPT EnterpriseやGeminiなどのパブリックAPI*7を利用する場合，本質的には外部事業者が運用するAI基盤を利用していることになる．データプライバシーに関する契

約上の保証があったとしても，アーキテクチャ上，データが企業境界外の基盤を通過する構成は残る．共有インフラにおけるモデル反転攻撃（Model Inversion Attacks：攻撃者がモデルにクエリを投げて訓練データを抽出する手法）や，プロンプトインジェクションによるコンテキスト漏洩の理論的リスクは完全には消えない．これに対して，オンプレミス展開や閉域環境では，AIの知能そのものである「モデルの重み（学習データから得たパラメータ）」，利用者が入力した「推論ログ（対話履歴）」，そして社内資料などをAIが読み取れる形に変換した「ベクトル埋め込み（データの特徴量）」のすべてを，自組織の物理的な管理下に置くことができる（図2）．

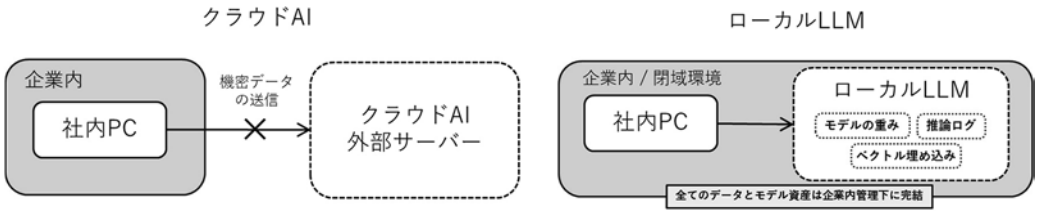

図2　クラウドAIとローカルLLMにおけるデータ処理境界の違い

3. 2　2025～2026年に普及が加速した三つの背景

ローカルLLMの普及が加速した背景は三つある．第一の背景は，モデルの軽量化・高性能化と，それを支えるハードウエア要件の低下である．2025年にはMetaのLlama系列，Alibabaの Qwen，MistralAIの各モデルが性能向上を遂げ，OpenAIもオープンソースモデル「gpt-oss」をリリースした．NTTが2025年に発表した国産軽量モデル「tsuzumi 2[11]」は，単一のGPUで動作しながら金融・医療・公共分野の知識を強化したモデルとして紹介されている．さらに，4ビットおよび8ビット量子化[*8]（GGUF，AWQフォーマット[*9]）の普及により，従来よりも少ないGPU枚数で運用できる構成が現実的になってきた．加えて，特定のGPUアーキテクチャに縛られない推論環境の多様化が進んだこと[12]により，ハードウエアの選択肢は大きく広がっている．

第二の背景は，導入ツールのGUI化・民主化である．かつてはコマンドラインに精通した一部のエンジニアに限られていたローカルLLMの運用も，2025年以降は「Ollama」や「LM Studio」といったツールの普及により，IT担当者でも試しやすい環境が整ってきた．これらのツールは，量子化モデルの変換や，Apple Siliconのユニファイドメモリ，AMD製アクセラレータなど，多様なハードウエア環境における差異の吸収にも寄与している[12][13][14]．

第三の背景は，規制環境の変化とデータ主権への関心の高まりである．欧州では「EU AI Act[*10]」の段階的な施行が進み，GDPR[*11]との整合性を確保する観点から，データがどの国・どのサーバーで処理されるかが経営レベルの論点になっている．日本でも個人情報保護や経済安全保障の観点から，海外ハイパースケーラーへの依存を見直す動きがあり，国内データセンターや顧客専有環境への関心を高める要因となっている．

4.　自社専用LLM環境（ローカルLLM）を実現する三つのアプローチの比較

企業がChatGPT等のパブリックAPIを避け，自社の統制下でLLMを運用する方法は一様

ではない．本稿では，他テナントから論理的または物理的に隔離された「自社専用のLLM環境」を広義のローカルLLMと定義し，大きく三つのアプローチに整理する．それぞれのアプローチには明確なトレードオフが存在し，どれが優れているかは組織の要件によって異なる．

4.1　アプローチA：OSS[*12]モデルの自社構築（DIY型）

LlamaやQwen，Mistralなどのオープンソースモデルを自社のGPUサーバー（オンプレミスまたはIaaS）にデプロイし，RAG[*13]（検索拡張生成）やファインチューニング[*14]で業務に最適化するアプローチである．米国のテクノロジー企業や金融機関では，このDIYアプローチが主流といわれており，モデルの重みへの直接的なアクセスを活用して差別化されたプロダクトを自社開発している．標準的な技術スタックは，Kubernetes上で稼働するLlama 3等のモデル，オーケストレーションを行うvLLMやTGI（Text Generation Inference），そしてアプリケーションロジックを担うLangChainやDify（ローコードプラットフォーム）などで構成される．

このアプローチの強みは，ライセンス費用が無料であること，自社のユースケースに合わせた深いカスタマイズができること，技術的な資産として社内に蓄積されることである．過去20年の判例ファイルで訓練された法務特化モデルや，社内のプライベートリポジトリへの完全なアクセス権を持つコーディングエージェントなど，パブリッククラウドでは実現困難なユースケースが開ける．

弱みは，高度なAI/MLエンジニア（AIや機械学習のエンジニア）による継続的な運用体制が不可欠であり，初期のモデル選定からRAGパイプラインの構築，ハルシネーション[*15]対策まで，品質を業務水準に引き上げるための工数と失敗コストが膨大になりやすいことである．近年はNVIDIA NIM（NVIDIA Inference Microservices）[15]のような推論最適化コンテナの登場でデプロイの障壁は下がりつつあるが，依然として運用難易度は高い．さらに，トレンドマイクロのレポート[8]が警告するように，公開リポジトリから取得したモデルファイル自体にマルウエア[*16]が仕込まれるリスクも現実に存在する．PoC[*17]に際してはまず7B～8Bパラメータの量子化モデルとRAGの組み合わせから始め，小さな業務課題で実証実験を行い，効果が出たら大規模モデルへ拡張する段階的アプローチが現実的である．

4.2　アプローチB：メガクラウドのプライベート環境（VPC型）

AWSやAzure，GCPなどのメガクラウドが提供する仮想プライベートクラウド（VPC）内にLLMを展開するアプローチである．「クラウド上にあるが，論理的に隔離されている」という位置づけである．近年は，NeoCloud（ネオクラウド）[16]と呼ばれる特化型GPUクラウド（CoreWeave, Nebius Group, Lambda Labs等）による「ベアメタル」インスタンス（単一の顧客に専有される物理サーバー）の提供も進んでいる．これは，オンプレミスに近い「物理的な主権（他テナントとの混在なし）」と，クラウド特有の「調達の速さと弾力性」を両立させる，第三の選択肢として注目されている．

このアプローチの強みは，自社でGPUハードウエアを保有・管理する負担がなく，スケーラビリティが高いことである．Azure OpenAI Serviceのプロビジョンドスループットや Amazon Bedrockなど，既存の企業向けクラウド契約の延長線上で導入できる場合が多い．冷却設備や電力供給の管理といった物理インフラの煩雑さを回避できる点も大きい．

弱みは，「厳密には自社の物理環境ではない」というガバナンス上の問題が残ることである．防衛関連企業や機密情報を扱う政府機関のように，高度なデータ主権要件を持つ組織にとっては，米国法（CLOUD Act[*18]）の適用範囲に含まれる可能性がある米国企業のインフラ上にデータを置くこと自体が許容できないリスクになりうる．また，継続的な API 利用料やインスタンス稼働費用がランニングコストとして積み上がる構造も無視できない．社内利用が想定以上に拡大した場合，短期間で月間予算を消化する恐れがある．

4. 3　アプローチ C：エンタープライズ特化のパッケージ型ローカル LLM

企業向けに最適化された形でハードウエアまたはソフトウエアがパッケージングされて提供されるローカル LLM ソリューションである．この領域では，設計思想の異なる複数の選択肢が存在する．例えば，Lazarus AI 社との戦略パートナーシップに基づきユニアデックス株式会社（以下，ユニアデックス）が展開するソリューション[17][18][19]は，ハルシネーションを抑制した高精度な文書処理・情報抽出と，セキュアな閉域環境での運用に軸足を置いた構成である．一方，日立ソリューションズが Allganize と提携して提供する「Alli LLM App Market[20]」は，SaaS 型からオンプレミス型まで選択できる生成 AI アプリ基盤として位置づけられている．また，NTT の「tsuzumi[11]」は，1GPU で動作可能な国産 LLM として，自社インフラや SIer 基盤へ組み込む前提で採用が進んでいる．

このアプローチの強みは，導入の確実性とコストの予測可能性である．LLM 単体ではなく，RAG に必要なベクトルデータベースや UI，権限管理，さらにはアプライアンスサーバー[*19]が一体の「プレパッケージ」として提供されるケースが多く，AI 専門エンジニアを持たない企業でも本番適用を実現しやすい．完全なオンプレミス・閉域網での稼働が可能なため，製造業の設計データや金融機関の顧客データといった機密性の高い要件に合致する．また，買い切り型やリース契約により初期コスト化しやすい点は，日本企業の稟議プロセスとも相性が良い．

弱みは，「モデルの陳腐化への対応」と「スケーラビリティの物理的限界」である．日進月歩で進化する LLM 業界において，閉域環境のパッケージ製品は最新モデルへのアップデートにタイムラグが生じやすい．また，処理リクエストが急増した際，クラウドのように即座にリソースを拡張することが難しく，独自の既存システムと深く API 連携させる際にもベンダーのアーキテクチャに依存する制約が生じやすい．

ただし，近年では知識基盤をモデル本体から分離するアーキテクチャも登場しており，業務知識の更新をモデルの再学習に依存させず，外部の知識基盤側で管理することで，モデルを頻繁に入れ替えることなく最新情報への追従を図る取り組みも進んでいる．Lazarus AI の Vector Knowledge Graph（VKG）はその一例であり，7. 4 節で詳述する．

本章で述べたローカル LLM を実現するための三つのアプローチの比較を表 1 にまとめた．

表1　ローカルLLM実現アプローチの比較

比較項目	A：OSS自社構築	B：VPC型	C：パッケージ型
初期コスト	中〜高 （GPU調達・構築費）	低	中〜高 （パッケージ・ハード費）
ランニングコスト	低 （電気代・保守費）	高 （Cloudサービス利用費）	低〜中 （保守ライセンス費）
AI/ML専門人材	高度な人材が必要	構成により必要	原則として限定的
基盤運用・保守	自社対応が中心	クラウド側が一部担当	ベンダー支援を利用可能
スケーラビリティ・弾力性	低（ハード追加が必要）	高（クラウドの弾力性）	低（ハード追加が必要）
カスタマイズ性	高	中	限定的
データ主権・機密性	極めて高い	ベンダー規約に依存	極めて高い
ハルシネーション制御	自社でパイプライン構築	ベンダー機能に依存	製品の設計・機能に依存
導入までの期間	長い	短い	短い

5.　ローカルLLMの国内先行事例と段階的導入指針

本章では，国内における先行事例を踏まえながら，ローカルLLMを実運用へつなげる際の導入上の勘所を確認する．

5.1　国内の先行事例

ローカルLLMや閉域環境での生成AIの導入は，セキュリティ要件の厳しい業界を中心に実証・商用化がすでに進んでいる．

金融分野では，あおぞら銀行がneoAIと協業し，行内オンプレミス環境に金融業務に特化した独自のLLM基盤の構築を進めている[21]．また，同行は行内規定やマニュアルをRAG（検索拡張生成）技術と組み合わせて照会できる専用アプリを導入し，回答の根拠を提示させることでハルシネーションを抑えつつ検索業務を効率化している．常陽銀行も，クラウド上の閉域環境にChatGPTを構築し全行員向けに展開しているほか，RAGを活用した「営業ソリューション検索サービス」を追加し，行員の業務効率化を推進している[22]．

医療分野では，機密性の高い患者データの取り扱いが必須となるため，オンプレミス環境でのLLM活用が不可欠である．社会医療法人祐愛会織田病院は，オプティムが提供するオンプレミス生成AI「OPTiM AIホスピタル」を導入し，電子カルテと連携させた[23]．これにより，外部へデータを出すことなく，退院時看護サマリーの作成時間を約54%短縮するという実用的な成果を上げている．

これらのケースでは，顧客情報や患者の個人情報をパブリックな外部サーバーに送信しにくいという制約があるため，オンプレミス環境や専用の閉域環境で運用可能なLLM基盤が有力な選択肢となっている．これらの事例は，厳格なデータガバナンスが求められる領域においても，セキュアな環境構築とRAGの活用によって実用的な成果を目指せることを示している．

5.2　導入における留意点

「ローカルだから無条件に安全」という思い込みは禁物である．ローカルLLMは外部送信

リスクを排除するが，内部での権限管理，物理的なサーバーセキュリティ，マルウエア感染リスクは依然として存在する．トレンドマイクロが警告するモデルファイルへの悪意あるコードの埋め込みリスク[8]は，OSS モデルを外部リポジトリから取得する際に現実に発生しうる．

また，生成結果の正確性に対する過信も危険である．RAG のチャンク*20 設計が不適切であれば的外れな文書を参照し，文書の鮮度管理が不十分であれば古い情報に基づいた回答が生成される．単に検索精度を上げるだけでなく，「根拠に基づかない回答を徹底的に排除する」といった，事実性に特化したアーキテクチャを採用するか，自社で RAG の検索品質を継続的に改善する体制を持つかが問われる．

さらに，PoC はまず小さく始めるのが妥当である．文書検索，規定照会，情報抽出のように正解と根拠を確認しやすいユースケースから着手し，アクセス制御，ログ監査，ドキュメント更新責任を並行して設計しなければ，管理下で運用するという前提が形骸化する．「導入したら終わり」ではなく，継続的な運用改善を前提に据えることが肝要である．

これら先行事例と運用上の留意点は，ローカル LLM が実用フェーズにあることを示している．一方で，汎用モデルの延長線上では解決が難しい「事実の厳格な検証」という壁も明らかになった．

6. ローカル LLM 導入のコスト構造の考察

本章では，クラウド API とオンプレミス運用の費用構造を対比しながら，ローカル LLM 導入時に見るべきコストの全体像を整理する．

6. 1 「使うたびに払う」か「買って持つ」か

クラウド AI は「初期費用なしで今すぐ使える」という手軽さが魅力であった．しかし業務利用が本格化するにつれて，問題が顕在化してきた．クラウド API は処理した文字数（トークン数）に応じて課金される従量制のため，社員全員が日常的に使い始めたり，大量の文書を自動処理するシステムに組み込んだりすると，利用料が青天井で積み上がる構造にある．利用量が少ないうちはクラウドに軍配が上がる．1 日あたり 1000 件程度の問い合わせであれば，初期費用が不要なクラウドが明らかに安価である．しかし，保険金請求の自動処理や社内文書の一括解析など，大量のドキュメントを扱うような業務にクラウド API を使い続けると，その費用は現実的ではなくなる．

一方，自社サーバーで LLM を動かす場合の目安を示す．3 章で述べた OpenAI のオープンウェイトモデル「gpt-oss」を例にとると，gpt-oss:20b（20B パラメータ）は NVIDIA RTX 4090 を 1 ～ 2 枚搭載したワークステーションでの運用が現実的であり，日常的な文書検索・FAQ 対応用途の候補となる．より高い回答品質を求める場合は gpt-oss:120b（120B パラメータ）が選択肢となるが，こちらはより大きな GPU メモリを持つサーバー構成が不可欠になる．いずれの構成でも，機器調達の初期費用を除けば追加の処理コストは主として電力と運用保守費である．国内インフラ企業の試算でも，安定稼働時はオンプレミスが 3 年間で 30 ～ 50％のコスト削減になると報告されており[24]，利用頻度が高く，問い合わせ件数や処理文書量が大きい業務では，従量課金型のクラウド API よりオンプレミスの方が総コストで有利に転じる可能性がある（図 3）．

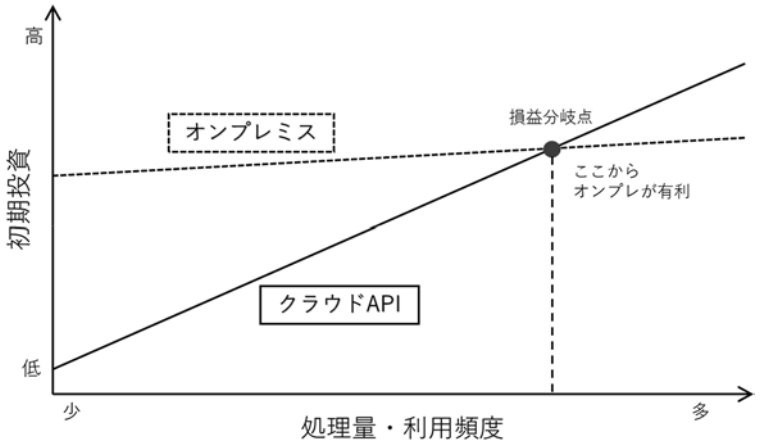

図3　クラウドAPIとオンプレミスLLMのコスト構造と損益分岐点

6.2　見落とされがちなコスト

ハードウエアの費用に加えて，自社運用では電力や冷却コストもかかる．高性能GPUを多数収容する高密度ラック（40kW超など）では液冷化の設備改修を要するケースもあり，電気代の高い日本では無視できない負担となる．また，システムの維持にはAI/MLエンジニアが不可欠であり，その人件費やSIerへの委託費も計上すべきである．

クラウドAPIはこれら運用の手間やコストが利用料に含まれるため比較しにくい．しかし，大量データの送受信には別途「データ転送料（Egress料金）」が発生し，従量制ゆえに予算が立てにくい点に注意すべきである．裏を返せば，固定費での予算管理を好む日本企業にとって，コストが読みやすいオンプレミスの方が社内稟議を通しやすい側面もある[25]．

6.3　「買ったGPUがすぐ古くなる」という懸念について

「高額なGPUを購入しても短期間で陳腐化する」という不安は，導入をためらわせる典型的な理由である．しかし，新規学習と推論用途では求められるハードウエアの更新頻度が異なる．推論に限定すれば，AWQ等の4bit/8bitへの「量子化技術」によりメモリ使用量を劇的に削減することができる．実際に海外事例では，量子化等により推論のリソース要件を最大70%削減しつつ品質を維持した報告もある[26]．こうした技術進化が旧世代ハードウエアのライフサイクルを延ばし，長期間の実運用に耐えられるようにしている．

7.　エンタープライズ特化型ローカルLLMの技術的特徴―Lazarus AIという選択肢

5章で確認した通り，企業が機密資産を知能化する上で，モデルの「表現力」以上に「事実の正確性」が成否を分ける．また前章では，クラウドAPIとオンプレミス運用のコスト構造と判断軸を整理した．本章では，これらを踏まえ，事実抽出の正確性と閉域運用を両立するソリューションとしてLazarus AIを取り上げる．

7.1　セキュリティ設計と日本展開―ユニアデックスとの戦略的提携

Lazarus AIは，文書処理と情報抽出に特化した企業向けAIである．保険・金融・政府機関などの高機密領域での利用を想定し，顧客専有のクローズド環境で処理を完結させる構成や，

参照元提示による検証可能性を特徴とする[19].

防衛産業や重要インフラにおいて求められる，インターネットから完全に遮断されたエアギャップ環境*21 や，強い閉域性を持つネットワーク環境への展開にも対応できる.

国内では，ユニアデックスが 2025 年 7 月に戦略パートナー契約を締結し，オンプレミス向け LLM ソリューションの提供を開始した[17]．これにより，RikAI2（7.3 節）や VKG（7.4 節）を中心とした構成を，日本企業の閉域環境へ展開する体制が整っている.

7.2　ATLS プラットフォーム――「AI を指揮するための AI」

Lazarus AI の技術的な中核を担うのが，ATLS と呼ばれるオーケストレーション・プラットフォームである[27]．ATLS は，タスクに応じて最適な AI モデルを選択・配分する「AI を指揮するための AI」として機能し，RikAI2 をはじめとする複数の専門モデルを統合的に制御する.

米国国防総省（DoD）などの情報分析事例では，膨大なマルチソース・データの相関分析において，人間が約 8 週間を要するタスクを数分へ短縮し，大幅なコスト削減を実現したとされる[19]．なお，2026 年 3 月時点では政府・防衛分野での実績が先行しており，民間向け展開のロードマップは今後順次公開される見込みである.

日本企業にとって ATLS が示唆するのは，「単一のモデルですべてを解決する」のではなく，「複数の専門モデルを協調動作させる」というエージェンティックなアプローチの有効性である．ここで重要なのは，2 章で提起した「AI エージェントが自律的に外部と通信し，機密情報を漏洩させるリスク」に対して，ATLS が明確な解決策を提示している点である．ATLS は「AI を指揮・監視する AI」として機能し，エージェントごとの権限を最小限に制限して，実行可能なアクションやデータアクセスの境界を厳密に統制する.

このように，暴走リスクを抑え込みながら複数の AI を安全に連携させ，実業務のプロセスに適合させるアプローチこそが，これからのエンタープライズ AI 運用における重要な鍵となる．この思想の核となるのが，次節で述べるフラッグシップモデル RikAI2 である.

7.3 「生成」ではなく「抽出・精査」という設計思想― RikAI2 の技術的革新

企業業務（法務判断，医療記録の参照，財務規定の照会など）における LLM 活用では，事実と異なる情報を生成するハルシネーションが致命的リスクとなる.

これに対して Lazarus AI は「We optimize for truthful reasoning」という思想を掲げ，一般的な生成型 AI よりも，文書からの正確な抽出と検証可能性を重視する[17]．その中核モデル「RikAI2」の技術的革新は，LLM と Vision Transformer（ViT）*22 を統合し，抽出型 AI*23（Extractive AI）としての動作原理を確立している点にある[28].

汎用的なマルチモーダルモデル*24 は視覚情報に対しても「生成（推測）」を行うためハルシネーションのリスクがある．一方で，RikAI2 は文書を視覚情報として直接解釈し，それを「厳格な事実抽出」にのみ用いる．そこに真の独自性がある（図 4）．RikAI2 では，手書き文字や複雑な表・グラフ・図面等のレイアウト構造を正確に把握でき，引用元の確実な明示（事後検証）を実現できる（図 5）．派生モデルとして構造化データ抽出に特化した「RikAI2-Extract」も提供されている.

図 4　汎用 LLM と RikAI2 における文書処理アーキテクチャの比較

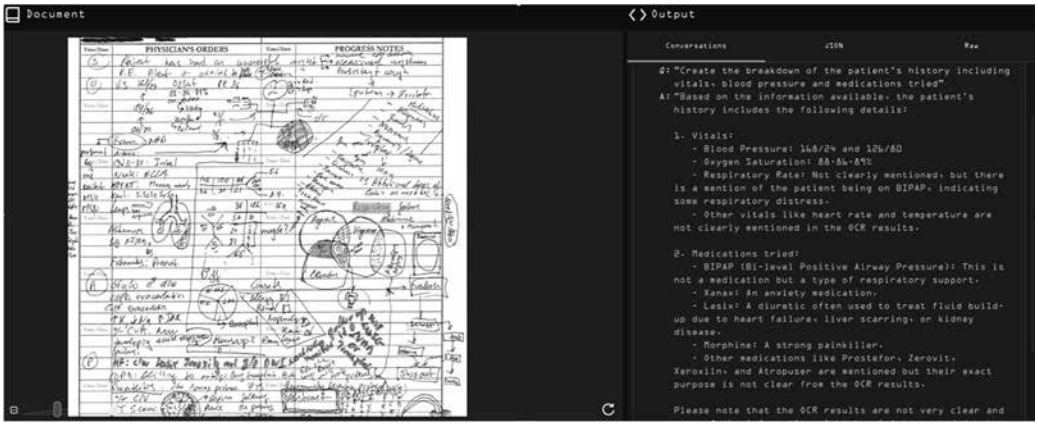

図 5　RikAI2 による手書き医療カルテの読み取り事例[19]

7. 4　Vector Knowledge Graph アーキテクチャ

RikAI2 の抽出精度を支える知識基盤が「Vector Knowledge Graph（VKG）」アーキテクチャである[19]．企業のドメイン知識*25 をモデルに組み込む際，ファインチューニングは再学習コストが高く，一般的な RAG（文書をチャンク分割するベクトル検索）では文書間の論理的関係や階層構造が失われやすい．加えて，OSS などを組み合わせた一般的な RAG 構成では，検索用の埋め込みモデルと生成用の LLM が異なるケースが多く，両者の意味や表現のズレが精度低下の要因となることがある．

VKG はベクトルデータベース*26 と知識グラフ*27 のハイブリッド構造により，検索の柔軟性を維持しつつ文書間の関係性を保持し，複雑な文脈理解を実現する．さらに，Lazarus AI のアーキテクチャはベクトル化から情報抽出に至るプロセスが統合的に設計されているとみられ，検索空間と生成モデル間の意味的なズレが生じにくく，より効率的で高精度な処理が期待できる．内部ドキュメントをインポートするだけで知識ベースがリアルタイムに更新されるため，再学習を伴わない運用が実現できる．これにより，新しいドキュメントの追加は知識グラフの拡張として処理され，業務マニュアルや規定集の更新にも即座に追従できる（図 6）．

以上を踏まえると，Lazarus AI は汎用的な生成性能よりも，規定集，契約書，設計書，報告書などの「根拠を示しながら読み解く・抽出する」業務に特化したローカル LLM である．閉域運用，参照元提示，文書構造理解の統合に同社の差別化要因がある．ATLS を含む全体アーキテクチャの国内展開にはまだ流動的な部分もあるが，出力の透明性と説明可能性を備えた RikAI2 と VKG を軸とする構成は，現時点でもセキュアな文書処理基盤として高い導入価値を持つ．

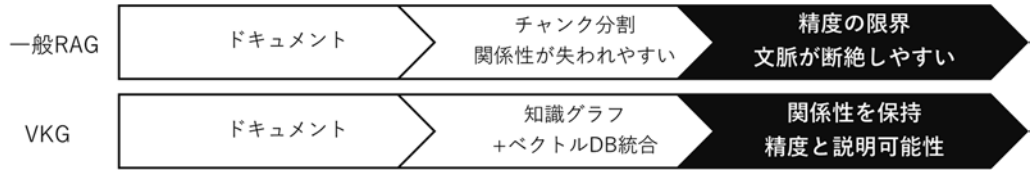

図 6　一般的な RAG と Vector Knowledge Graph の知識構造の違い

8.　Lazarus AI の位置づけとハイブリッド運用

本章では，Lazarus AI を全面置換のための製品としてではなく，クラウド AI と併用するハイブリッド構成の中での位置づけ，ガバナンス[*28]を維持しつつ安全に運用すべきかを考察する．

8. 1　ハイブリッド構成の中での位置づけとガバナンス

ここまで，データ主権を保護し，厳格な事実抽出を実現するローカル LLM としての Lazarus AI の特性を詳述してきた．しかし，実際のエンタープライズ環境において，すべての AI タスクをローカル LLM のみで完結させることは，インフラのコストや処理の汎用性を考慮すると必ずしも現実的ではない．実際の現場で求められるのは，機密性の高いデータの処理や厳格な事実抽出には Lazarus AI を配置し，一般的な文章の要約やアイデア出しにはクラウド LLM を利用するといった，適材適所の「ハイブリッド構成」である．これが企業における生成 AI 活用の現実解であり，Lazarus AI が担うべき中核的な位置づけとなる．

しかし，このデータ主権とリスク許容度に基づく運用を安全に行うためには，単に「重要なデータはローカルで処理する」というネットワークの境界防御だけでは不十分である．データが「いま，どこで，どのように処理されているか」をシステム内部で常に把握できる状態（可観測性[*29]：Observability）を作り，国際基準に沿った統合的なガバナンスを効かせることが不可欠となる．具体的には，ISO/IEC 42001[29]が求める AI ライフサイクル全体での透明性の確保や，NIST AI 600-1[30]が推奨する継続的な監視体制の構築である．同時に，OWASP Top 10 for LLM Applications[9]が警告するプロンプトインジェクションや，意図せぬデータの流出といった固有リスクに対しても，システム全体で一貫した防御ポリシーを適用しなければならない．

8. 2　AI 統合管理基盤とドメイン特化型導入の展望

こうしたガバナンス要件を満たしつつ，安全なハイブリッド運用を実現する現実的な手段が，AI ゲートウェイや統合監視ソリューションなどの可観測性基盤の導入である．これらを活用することで，ユーザーの入力に機密情報が含まれる場合には自動で Lazarus AI へ処理を振り分けるセマンティック・ルーティング[*30]といった動的なトラフィック制御ができるようになる．導入にあたっては，ゲートウェイが処理のボトルネックとならないようパフォーマンスを確保しつつ，自律型 AI[*31]（Agentic AI）に対する権限付与を最小限に留め，NIST が求める継続的な監視の原則をクラウド・ローカル双方に適用することが考えられる．

これらクラウドとローカルを横断する複雑なネットワーク境界の制御や，統合的な可観測性の確保は，AI モデル単体の知識だけでは実現できない．IT インフラ全体を俯瞰し，ネットワークやセキュリティに精通したパートナーが，企業ごとの要件に合わせてシステム全体を最適化することで，初めて安全なハイブリッド運用が成立する．これらのアーキテクチャ要件を踏ま

えた上で，現実のハイブリッド構成において重要なのは，6章で述べたLazarus AIが万能基盤ではなく，機密性と検証可能性を要する領域を担う中核として位置づけられる点である．Gartner® は，2027年までに企業で利用される生成AIモデルの過半数が業務・ドメイン特化型モデル（特定の業界や業務機能に特化したもの）になると予測している．これは2024年時点の1%から増加となる[31]．この潮流において，文書理解と検証可能性を重視したLazarus AIは，既存のクラウドAIを全面置換するのではなく，限定領域から導入するローカルLLMの有力な選択肢となる．

9. おわりに

本稿では，生成AIの急速な普及を背景にセキュリティとデータ主権の観点からローカルLLMが求められる背景を整理し，Lazarus AIの技術的特徴と企業導入上の位置づけを考察した．

同製品の意義は，単に閉域環境で運用できる点にとどまらない．RikAI2による文書理解・情報抽出とVKGによる知識基盤，そしてATLSによる複数モデル協調の構想を通じて，汎用LLMが持つ「表現の多様性」よりも「推論の正確性と検証可能性」を優先した点にある．とりわけ厳格なガバナンスが求められる規制領域において，この設計思想は実運用上の重要な意味を持つ．汎用的な生成AIが持つ利便性を享受しつつも，自社のコアコンピタンス[*32]となる機密データは，確実な制御下で安全に処理・抽出できなければならない．Lazarus AIの抽出・検証機能に，自社環境に適したインフラ統合を掛け合わせることは，企業がデータ主権を保護しながらAIを業務適用するための有力なアプローチとなる．

さらに将来を見据えれば，この堅牢な推論基盤は，自律的に業務を遂行するAgentic AI時代における「マルチエージェント協調」の重要な布石となる．例えば，クラウド上の汎用エージェントが収集した外部市場データを，オンプレミス環境のLazarus AIが社内の機密ドキュメントと安全に照合・精査するといった，セキュリティ境界を越えた連携が実現できる．事実に基づく正確性を担保した環境があってこそ，AIは単なる「回答者」から信頼できる「エージェント」へと飛躍できるだろう．

＊1　クラウドを使わず，自社の建物内に設置したサーバーでシステムを運用する形態．
＊2　特定の企業・組織が専用に利用するクラウド環境．パブリッククラウドとは異なり，他社とインフラを共有しない．
＊3　自社のデータをどの国・どの組織が管理・処理するかを自ら決定できる権利・状態．
＊4　メール送信・ファイル操作・外部サービスの呼び出しなど，人に代わって自律的に作業を実行するAI．
＊5　AIへの入力（プロンプト）に悪意ある命令を紛れ込ませ，意図しない動作をさせる攻撃手法．
＊6　AIの学習データに意図的に誤情報を混入させ，モデルの動作を狂わせる攻撃．
＊7　不特定多数の企業・開発者が利用できる公開されたAIサービスの接続口．ChatGPTやGeminiなどが該当し，インターネット経由でデータを送受信する．
＊8　AIモデルのパラメータのビット数を削減し，精度をほぼ保ちながらメモリ消費・計算コストを削減する技術．
＊9　LLMの量子化モデルの代表的なファイル形式．メモリを節約しながら，LLMをローカル環境で動かすために使われる．
＊10　EUが定めたAIに関する包括的な規制法．AIのリスクレベルに応じて義務や禁止事項を定めており，段階的に施行されている．
＊11　EUの個人情報保護法（General Data Protection Regulation）．個人データの収集・処理・移転に関するルールを定めており，違反には高額の制裁金が科される．AI活用においても

データの取り扱い方法に影響する．

＊12 オープンソースソフトウエアのこと．ソースコードが公開されており，誰でも無償で利用・改変・再配布できるソフトウエアの総称．本稿では Llama や Qwen，Mistral などの公開された AI モデルを指す．

＊13 社内文書などを検索して取得した情報をもとに AI が回答を生成する手法．事実性の向上に有効．

＊14 汎用 LLM モデルに特定業務のデータを追加学習させ，専門領域に最適化すること．

＊15 AI が事実と異なる情報を，もっともらしく生成してしまう現象．

＊16 悪意を持って作られた不正なソフトウエアの総称．ウイルス・ランサムウエアなどが含まれる．文中では OSS の LLM モデルのファイル自体に仕込まれるリスクとして言及されている．

＊17 新しい技術やシステムが実際に機能するかを小規模で検証する「概念実証」のこと．本格導入前に効果や課題を確認するための試験的な取り組み．

＊18 米国企業に対して，海外サーバー上のデータも米政府の要求に応じて開示を義務付ける米国法．外資クラウド利用のリスク要因．

＊19 特定の用途に特化して，ハードウエアとソフトウエアがあらかじめ一体化された状態で提供されるサーバー機器．

＊20 RAG で長い文書を扱う際に，小さなブロック（チャンク）に切り分けること，およびその分割方法の設計．分割が不適切だと文脈が失われ，的外れな回答が生成される原因になる．

＊21 インターネットや外部ネットワークから物理的に完全に隔離された環境．防衛産業や重要インフラ，高度な機密情報を扱う組織で求められる，最も強固な閉域運用形態．

＊22 画像を細かなパッチに分割し，文書のように解析する視覚 AI モデル．RikAI2 の基盤技術

＊23 文書の内容をもとに新たな文章を「生成」するのではなく，文書中に書かれている事実をそのまま「抜き出す」ことに特化した AI．根拠が明示できるため，ハルシネーションが起きにくい．

＊24 テキスト・画像・音声など複数の種類のデータを統合的に扱える，特定用途に限定しない幅広い目的向けの AI モデル．GPT-4o や Gemini などが該当する．

＊25 特定の分野，業界，または業務に関する専門的な知識のこと．

＊26 テキストや画像などを数値化（ベクトル）して保存し，単語の完全一致ではなく「意味の類似性」で高速検索を可能にする AI 向けのデータベース．

＊27 データや概念を「点」とし，その関係性を「線」で結んでネットワーク状に構造化したもの．AI が単なる単語の検索だけでなく，情報同士のつながりや複雑な文脈を理解するための基盤技術．

＊28 AI や機密データを安全かつ適正に管理・運用するための社内ルールや統制（アクセス制御，監査，情報漏洩対策など）の仕組み．

＊29 システムの外部出力から内部の状態や動作を把握する能力．AI 運用においては，モデルの回答精度，処理遅延，トークン消費量，エラーなどをリアルタイムで監視・追跡し，問題を迅速に特定する仕組み．

＊30 入力されたテキストの意味を理解して，最適な処理先やモデルに振り分ける仕組み．

＊31 目標に基づき，自律的に計画・実行・ツール利用を行う AI エージェント群の総称．

＊32 他社には真似できない，顧客に独自の価値を提供する企業の中核的な能力のこと．

参考文献

［1］ 総務省，令和 7 年版 情報通信白書，2025 年 7 月，
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/html/nd112220.html

［2］ Cyera，Cybersecurity Insiders，2025 State of AI Data Security Report，2025，
https://www.cyera.com/research-labs/2025-state-of-ai-data-security-report

［3］ 独立行政法人情報処理推進機構，「テキスト生成 AI の導入・運用ガイドライン」，
独立行政法人情報処理推進機構，
https://www.ipa.go.jp/jinzai/ics/core_human_resource/final_project/2024/f55m8k0000003spo-att/f55m8k0000003svn.pdf

［4］ 総務省，AI 利活用ガイドライン，
https://www.soumu.go.jp/main_content/000624438.pdf

［5］ Gartner®，Emerging Risk Deep Dive: Shadow AI，Ben Fisher et al.，18 August 2025
https://www.gartner.com/en/documents/6714034（Gartner Subscription Required）
GARTNER is a trademark of Gartner, Inc. and its affiliates.

［6］ Okta，Enterprise Buyer Survey: AI Agent Security，2026 年 2 月，
https://www.okta.com/newsroom/articles/enterprise-buyer-survey-ai-agent-security/

［7］ NTT ドコモビジネス，「生成 AI が情報漏えいを引き起こす？「シャドー AI」のリスクと対策を解説！」，https://www.ntt.com/bizon/shadow-ai.html

［8］ トレンドマイクロ，「AI エージェントと脆弱性 PART 1」，
https://www.trendmicro.com/ja_jp/research/25/f/unveiling-ai-agent-vulnerabilities-

part-i-introduction-to-ai-agent-vulnerabilities.html
[9] OWASP, Top 10 for LLM Applications 2025, https://genai.owasp.org/llm-top-10/
[10] NTTPCコミュニケーションズ，セキュア×高速！ローカルLLMが変える企業AI活用とデータガバナンス最前線，2025年12月，
https://www.nttpc.co.jp/gpu/article/knowledge22_local_llm.html
[11] NTT，NTT版大規模言語モデル「tsuzumi 2」，2025年10月
https://www.rd.ntt/research/LLM_tsuzumi.html
[12] ITmedia PC USER，M4 Mac miniで「gpt-oss」は動く？ 動作が確認できたローカルLLMは……，2025年9月，
https://www.itmedia.co.jp/pcuser/articles/2509/30/news035.html
[13] Ollama，AMD support，2024年3月，https://ollama.com/blog/amd-preview
[14] PC Watch，Ryzen AI Max+ 395を「ビデオメモリ128GB積んだ鬼GPU」にしてしまう方法，2025年10月，
https://pc.watch.impress.co.jp/docs/column/nishikawa/2054963.html
[15] NVIDIA，NVIDIA NIMマイクロサービス，
https://www.nvidia.com/ja-jp/ai-data-science/products/nim-microservices/
[16] NTTドコモビジネス，「ネオクラウドとは？」，
https://www.ntt.com/business/services/xmanaged/lp/column/neocloud.html
[17] ユニアデックス株式会社，Lazarus AIとの協業によるオンプレミス向けLLMソリューション提供，2025年7月，
https://www.uniadex.co.jp/news/2025/20250717_lazarus-ai.html；
[18] ユニアデックス株式会社，Lazarus AI製品情報，
https://solution.uniadex.co.jp/service/product/lazarus-ai.html
[19] Lazarus AI，Technology，https://www.lazarusai.com/technology
[20] 日立ソリューションズ，Alli LLM App Market，
https://www.hitachi-solutions.co.jp/allganize/
[21] PR TIMES，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，2025年4月17日，
https://prtimes.jp/main/html/rd/p/000000026.000109048.html
[22] PR TIMES，【常陽銀行】生成AI「ChatGPT」のバージョンアップおよびRAGを活用した「営業ソリューション検索サービス」の取り扱い開始について，2025年9月18日，https://prtimes.jp/main/html/rd/p/000000033.000081984.html
[23] オプティム，オンプレミス生成AIサービス「OPTiM AIホスピタル」によって退院時看護サマリー作成にかかる時間を54.2%削減―社会医療法人祐愛会織田病院様，2025年10月6日，https://www.optim.co.jp/media/cat-case/aih_251006-01
[24] Ragate，オンプレミスLLM構築の実践：セキュアな閉域環境での大規模言語モデル運用術，https://www.ragate.co.jp/media/developer_blog/xooise_gx7xb
[25] マクニカ，クラウドでは難しい社外秘データの活用に向けた生成AI ～NVIDIAを例にローカルLLMについて考える～，
https://www.macnica.co.jp/business/semiconductor/articles/nvidia/148236/
[26] Reinforz BizDev，「生成AI原価の真実と収益最大化戦略：GPUコスト時代の単価設計と粗利管理」，https://bizdev.reinforz.co.jp/4557
[27] Lazarus AI，ATLS for Department of Defense Information Analysis，
https://www.lazarusai.com/government/atls-for-department-of-defense-information-analysis
[28] アルゴスサービスジャパン株式会社，「ハルシネーション（幻覚）」を最低限に抑制できるLazarus RikAI2，https://note.com/argossj/n/n5b33813b0b75
[29] ISO, ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system, https://www.iso.org/standard/81230.html
[30] NIST, NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,
https://www.nist.gov/itl/ai-risk-management-framework
[31] Gartner®，Press Release, Gartner Forecasts Worldwide End-User Spending on GenAI Models to Total $14.2 Billion in 2025，2025年7月10日，
https://www.gartner.com/en/newsroom/press-releases/2025-07-10-gartner-forecasts-worldwide-end-user-spending-on-generative-ai-models-to-total-us-dollars-14-billion-in-2025

※　上記の参考文献に示したURLのリンク先は，2026年5月14日時点での存在を確認.

執筆者紹介　青　木　岐　夫（Michio Aoki）

2021年ユニアデックス(株)中途入社．AI・DX分野における新規サービスの開発・導入支援業務に携わる．現在はLazarus AIの開発・導入支援を主業務とし，生成AI関連の最新技術動向を継続的に調査・研究している．JDLA認定 E資格保有.