リコーのエンタープライズ特化型マルチモーダルAI：Qwen3.6/3.5-Ricohモデルの技術構造と市場における戦略的意義

Gemini 3.1 pro

1. エグゼクティブ・サマリー：エンタープライズAIにおけるパラダイムシフト

2026年6月5日、株式会社リコーは、中国のアリババクラウドが開発したオープンソースの大規模言語モデル（LLM）「Qwen」シリーズをベースモデルとして採用し、日本語ドキュメントにおける複雑な図表読解や論理的推論（リーズニング）能力を極限まで高めたマルチモーダル大規模言語モデル（LMM）、「Qwen3.6-Ricoh-27B-20260522」および軽量版「Qwen3.5-Ricoh-9B-20260522」の開発を発表した[1]。この画期的な開発は、経済産業省および国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）が主導する、国内における生成AIの基盤モデル開発力を強化するための国策プロジェクト「GENIAC」の第2期および第3期における成果として結実したものである[4]。

本発表は、単に「新たな国産AIモデルの登場」という事実を超え、企業におけるデータ利活用とセキュリティのジレンマを根本から解消するパラダイムシフトとして位置づけられる。これまで、多くの企業は「Gemini」や「GPT-4/5」クラスの強力な商用クラウドAIの推論能力を認識しつつも、請求書、経営資料、設計図といった極めて機密性の高い社内文書、いわば企業の「知の結晶」を外部のパブリッククラウドへと送信することに対し、データガバナンス上の強い懸念を抱き、AIの本格導入を躊躇してきた[3]。

リコーが開発した両モデルは、「オンプレミス環境（自社ネットワーク内のローカルサーバー）での完全な稼働」を大前提として設計されながら、独自の高度な強化学習およびカリキュラム学習によって、特定のドメイン（日本語の図表を含む複雑なドキュメントの読解と推論）においては、数千億から兆規模のパラメータを持つとされるクラウド型巨大商用AIモデルに肉薄、あるいはこれを凌駕する驚異的な性能を叩き出した[2]。本レポートでは、これら新モデルの根幹をなすベースアーキテクチャの優位性、リコー独自の学習手法がもたらした技術的ブレイクスルー、独自の図表読解ベンチマークが示す真の意義、そしてエコシステム全体としての産業界にもたらす波及効果について、多角的な視点から詳細な分析を提供する。

2. 開発の背景とマクロ環境：企業の「ダークデータ」とクラウドAIの限界

現代の企業活動において日々蓄積されるデータの大部分は、純粋なテキストデータではない。図、表、画像、グラフ、フローチャートなどが複雑に混在したPDF形式の報告書や、スライド資料といった視覚的なフォーマットで保存されている。従来のテキスト処理に特化した大規模言語モデルでは、これらの視覚的情報を伴うドキュメントを直接的に「読む」ことは不可能であり、結果として、企業の持つ膨大な知識の多くが検索不能かつ分析不可能な「ダークデータ」としてストレージの奥深くに埋も

れていた[3]。
この長年の課題を解決するブレイクスルーとして登場したのが、テキストと画像を同一のコンテキスト内で同時に処理できるマルチモーダル大規模言語モデル（LMM）である。LMMに請求書やマニュアルのスクリーンショットを入力すれば、AIは自ら視覚情報を解析し、必要な数値を抽出し、文脈に沿った要約やデータの比較を行うことが可能となる[4]。しかしながら、最先端のLMM機能を提供するGoogleのGeminiシリーズやOpenAIのGPTシリーズなどは、膨大な計算資源を必要とするため、原則としてクラウド上のAPI経由でサービスが提供されている。金融機関、製造業のR&D部門、医療機関、あるいは公共・自治体といった、厳格なデータガバナンスとプライバシー保護が求められる領域では、機密情報を含むドキュメントを外部ネットワーク上のパブリッククラウドに送信することは、セキュリティポリシーや法令遵守の観点から許容されないケースが依然として圧倒的に多い[2]。
さらに、2026年中盤のマクロ環境に目を向けると、OpenAIがAGI（汎用人工知能）社会への移行に向けた大規模な基金を設立し、Metaがグローバルなビジネスエージェントを展開するなど、ビッグテックによるクラウドAIの寡占化が進む一方で、サイバーセキュリティの脅威は新たな局面を迎えている。ロシア系のハッカー集団がAIエージェントを悪用してエンドポイント検知対応（EDR）を回避するマルウェアを自動生成する事案が発生するなど、AI自体が攻撃のベクトルとして利用されるリスクが顕在化している[7]。こうした背景から、日本の金融庁が地方銀行向けに独自のAI基盤モデルを無償提供する構想を進めるなど、国家や産業インフラのレベルにおいて、「安全に統制可能なローカル・オンプレミスAI」の重要性がかつてなく高まっている[7]。
したがって、市場には「外部ネットワークに一切接続せず、ローカル環境で自律的に動作し、なおかつ商用クラウドAIに匹敵する高度な図表読解能力を持つ軽量モデル」という、極めて技術的ハードルの高いソリューションに対する切実な要求が存在していた。リコーの「Qwen3.6-Ricoh-27B-20260522」および「Qwen3.5-Ricoh-9B-20260522」は、まさにこの産業界における決定的な空白地帯（ホワイトスペース）を埋めるための戦略的プロダクトとして設計されたのである[2]。

3. アーキテクチャの解剖：ベースモデル「Qwen」ファミリーの卓越性

リコーが自社のエンタープライズ向け独自モデルのベースとして選定したのが、アリババクラウドがオープンソースとして提供する「Qwen」ファミリーである。この選択は、モデルの基本性能の高さだけでなく、アーキテクチャの実用性とコミュニティの支持という観点から極めて合理的であった。ここでは、リコーの拡張の土台となったベースモデルが持つ生来の技術的強みを分析する。

3.1. Qwen3.6-27Bの高密度アーキテクチャと「Agentic」な特性

フラグシップモデルである27B（270億パラメータ）モデルのベースとなった「Qwen3.6-27B」は、2026年4月にリリースされたばかりの最新世代モデルである[2]。このモデルは、MoE（Mixture-of-Experts：専門家モデルの混合）アーキテクチャを採用した同ファミリーの35B-A3Bモデル（総パラメータ350億、アクティブパラメータ30億）等とは異なり、270億の全パラメータが常に推論に参加する「高密度（Dense）モデル」として設計されている[11]。Denseモデルは、MoE特有の複雑なルーティング処理を必要としないため、システムへの組み込みやデプロイメントが直感的であり、安定した挙動を示すことから、エンタープライズ向けの堅牢な基盤として理想的である[11]。開発者コミュニティや識者からも、「単なるベンチマーク上のスコアを追うだけでなく、実際のワークフローにおいて最も実用的で安定し

たオープンモデル」として絶賛されている[16]。
Qwen3.6-27Bが持つ技術的特性の中で、ドキュメント理解において特に強力に作用するのが以下の2点である。第一に、Agentic Coding（自律的コーディング・推論能力）である。このモデルは、単にコードのスニペットを生成するだけでなく、フロントエンドの開発ワークフローや、リポジトリ全体（プロジェクトを構成する複数のファイル群）にまたがる複雑な論理構造を高い精度で把握・推論する能力を備えている[9]。この「広範なコンテキストを横断して理解する」という特性は、プログラミングに限らず、企業内の複雑な階層構造を持つ仕様書や、複数の表が連動する決算資料などを横断して論理を組み立てる能力の土台となる。第二に、Thinking Preservation（思考プロセスの保持）機能である。これは、AIが推論を行う際の「思考の文脈（Thinking Context）」を対話の履歴全体にわたってメモリに保持するアーキテクチャ上の工夫である[9]。従来、AIはターンが変わるごとに思考プロセスをリセットしがちであったが、この機能により、多段階の推論（マルチステップ・リーズニング）において、過去のステップで導き出した論理や前提を忘れずに次のステップの計算や推論に活かすことができる。この反復的な問題解決能力の向上は、長文ドキュメントの解析において頻発するハルシネーション（もっともらしい嘘）の低減に直接的に寄与している[16]。

3.2. Qwen3.5-9Bの革新的なハイブリッドアーキテクチャ

一方、より少ないリソースでの運用を想定した軽量モデルのベースとなった「Qwen3.5-9B」は、90億パラメータという小規模なサイズの中に、現在のAI研究における最先端の構造パラダイムを組み込んでいる[19]。このモデルの最大の特徴は、「Gated Delta Networks」と従来の「Gated Attention」を緻密に組み合わせたハイブリッドアーキテクチャを採用している点にある[15]。具体的には、合計32の隠れ層が「8 × (3 × (Gated DeltaNet → FFN) → 1 × (Gated Attention → FFN))」という特殊なレイアウトで構成されている[15]。これは、長い系列データを極めて少ないメモリ消費と計算量で処理できる状態空間モデル（Mambaなどの系統に近いDeltaNet）の効率性と、TransformerベースのAttention機構が持つ圧倒的な文脈理解の正確性を融合させたものである[15]。
さらに、初期段階からのマルチモーダルトークンの融合（Early Fusion）を行う統合視覚言語基盤（Unified Vision-Language Foundation）を採用しており、テキストと画像を別々のモジュールで処理するのではなく、同一のコンテキスト内でシームレスに推論する能力を獲得している[19]。この結果、90億パラメータでありながら、ネイティブで262,144トークン（最大100万トークンまで拡張可能）という膨大なコンテキスト長をサポートしている[15]。このハイブリッド構造の威力は、ローカル環境での実証テストにおいても明白に示されている。例えば、一部の開発者によるESP32マイクロコントローラー用のコード生成とRAG（検索拡張生成）を組み合わせたテストでは、上位モデルである27Bが重いコンテキスト処理でタイムアウトを起こすようなローカル環境設定下であっても、Qwen3.5-9Bは10〜15秒という驚異的な速度でタスクを完了させ、ハードウェアの制約が厳しいエッジ環境における圧倒的なスピードと安定性を証明している[21]。

4. リコー独自のチューニング：強化学習とカリキュラム学習の精緻化

リコーは、上述した極めて優れたベースモデルに対し、日本企業のドキュメント処理に特化させるための独自の高度なチューニングプロセスを実施した。ここが、オープンソースモデルをそのまま利用するアプローチと、リコーの商用ソリューションを分かつ決定的な差異である。
リコーが用いた技術的ブレイクスルーの核心は、「強化学習（Reinforcement Learning）」および「カリ

キュラム学習（Curriculum Learning）」という二つの手法の精緻な適用にある[2]。通常の言語モデルの事後学習（ファインチューニング）では、大量の質問と正解のペアを与えてAIにパターンを暗記させるようなアプローチ（教師あり学習）が主流になりがちである。しかし、この手法では、訓練データに存在しなかった未知のレイアウトを持つ複雑な図表や、新たな論理ステップを要求されるタスクに直面した際、AIの応用が効かなくなる。これに対しリコーは、AIの「論理的推論（リーズニング）プロセス」そのものを根本から強化するため、強化学習における報酬関数（Reward Function）の設計を徹底的に見直し、精緻化した[4]。これにより、AIは単に答えを出力するだけでなく、「複数のステップからなる論理的な思考プロセスを経て結論を導き出す」能力を自己最適化するようになった[2]。

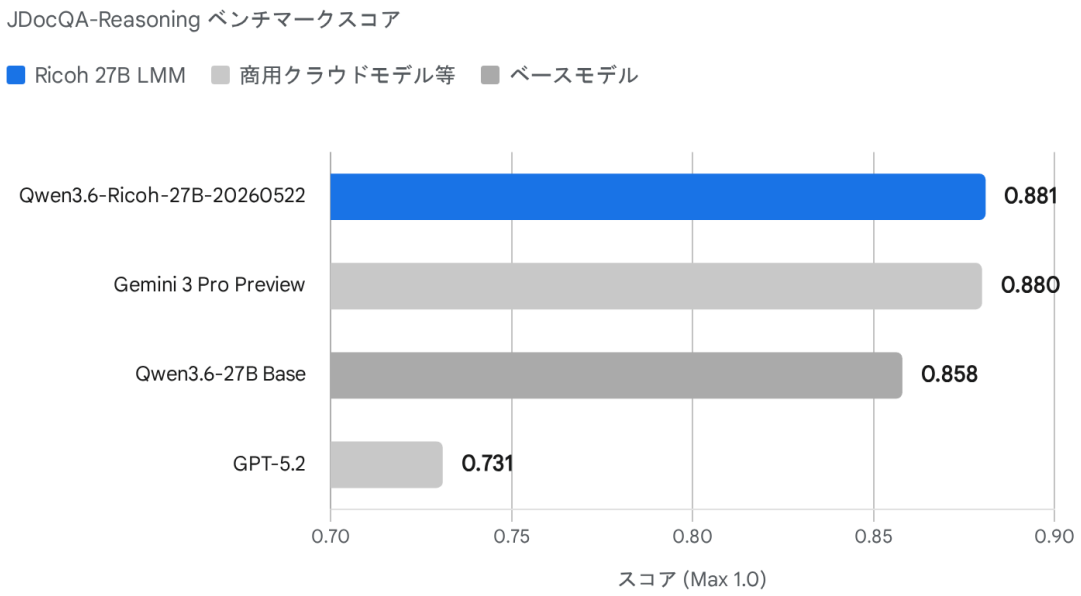
さらに、カリキュラム学習の概念を導入し、学習の難易度設計を高度化した[2]。これは、人間の学習プロセスと同様に、AIに対して最初は比較的単純なテキストや図表の抽出問題を提示し、段階的に難易度の高い複雑なドキュメントの読解（例えば、複数の表から数値を拾い出して四則演算を行い、最終的な傾向を推論するようなタスク）へと訓練を進める手法である。この漸進的なアプローチを用いることで、特定のフォーマットやデータセットにだけAIが過剰に適応してしまう「過学習（Overfitting）」を効果的に抑制しながら、汎用的かつ強靭な日本語ドキュメント読解能力を引き上げることに成功したのである[4]。

5. ベンチマーク評価が示すドメイン特化の威力：「逆転現象」の証明

リコーの独自の学習手法がもたらした成果は、厳格なベンチマークスコアに如実に表れている。性能評価にあたっては、リコー自身が開発し、2026年5月29日にオープンデータセットプラットフォームであるHugging Face上で無償公開した、日本語の図表読解や多段推論を評価するための独自ベンチマークツール「JDocQA-Reasoning」等が用いられた[4]。

このベンチマークの設計は極めて高度である。ベースとなったデータセットは、テキストと視覚情報を組み合わせた日本語QAデータセット「JDocQA」のテスト画像サブセットである。そこからさらに、棒グラフ、折れ線グラフ、企業の決算報告書、複雑な路線図など、ビジネスシーンで頻出する20種類以上の図表に特化し、独自のQAアノテーションを付与して構築された[5]。収録されている1,362問（あるいは関連資料による1,287問）の設問は、単なる情報の抽出にとどまらず、データを用いた計算、複数項目の比較、トレンド分析、さらには既存の要素から欠損しているデータを論理的に推論して補完するといった、極めて高度な多段推論（Multi-step reasoning）タスクによって構成されている[4]。

図表読解推論における性能比較：国産オンプレミスAIが商用クラウドAIに肉薄



独自ベンチマーク「JDocQA-Reasoning」（最高スコア1.0）における各モデルの性能。リコーの27Bモデルは、ベースモデルから大幅な性能向上を果たし、Gemini 3 Pro Previewと同等水準、GPT-5.2の参考値を上回る結果を記録した。

データソース: リコー プレスリリース, Innovatopia

ベンチマークの分析から明らかになる最も重要な洞察は、特定領域における巨大モデルの完全なる凌駕、すなわち「逆転現象」の証明である。評価結果において、270億パラメータ規模のリコーモデル「Qwen3.6-Ricoh-27B-20260522」は、最高スコア1.0のうち「0.881」という極めて高い数値を記録した[4]。これは、ベースモデルであるQwen3.6-27B（0.858）や、前世代のリコーモデルであるQwen3-VL-Ricoh-32B（0.826）を明確に上回るだけでなく、汎用型AIの最高峰として君臨し、パラメータ規模が数千億から兆に達すると推測されるGoogleの「Gemini 3 Pro Preview」の参考値（0.880）と事実上同等に並び、さらにOpenAIの「GPT-5.2」の参考値（0.731）を大きく引き離す結果となった[4]。

この事実は、あらゆる知識を網羅しようとする汎用的な巨大モデル（AGI志向のモデル）に対し、ターゲットドメインを「日本語の図表読解と多段推論」に限定し、そこにリソースを集中して徹底的に鍛え上げた中型特化モデルが、実務的なタスクにおいて十分に対抗、あるいは凌駕し得ることをデータで実証したものである[4]。この成果は、ハードウェアリソースや消費電力に厳しい制約のあるオンプレミス環境において、いかにして商用トップレベルのAIを実運用に乗せるかという命題に対する、極めて有力な解答となる。

以下に、リコーが実施した各種ベンチマークの総合結果を整理する。特筆すべきは、LMMとしての

図表読解能力（VLM性能）のみならず、テキストベースの「ELYZA-tasks-100」や「Japanese MT-Bench」においても、ベースモデルを凌駕する高い水準の自然言語処理能力（LLM性能）を同時に達成している点である[4]。

モデル名 / 評価軸	**JDocQA-Reasoning(図表推論・最高1.0)**	**JDocQA(図表読解・5点満点)**	**ELYZA-tasks-100(日本語テキスト・5点満点)**	**Japanese MT-Bench(日本語テキスト・10点満点)**
Qwen3.6-Ricoh-27B-20260522	**0.881**	**4.22**	**4.64**	**9.48**
Qwen3.6-Ricoh-27B (AWQ-W4A16 4bit量子化)	0.868	4.20	4.62	9.35
Qwen3.6-27B（ベースモデル）	0.858	4.15	4.58	9.35
（参考）*Gemini 3 Pro Preview*	*0.880*	*4.24*	-	-
（参考）*GPT-5.2*	*0.731*	*3.93*	-	-
Qwen3.5-Ricoh-9B-20260522	**0.782**	**4.00**	**3.95**	**7.93**
Qwen3.5-9B（ベースモデル）	0.762	3.89	3.76	7.65
（参考）*Qwen-3-VL-Ricoh-8B (前作)*	*0.718*	*4.00*	-	-

(データ出典：[4] ※推論および評価は5回実施された平均値。ELYZA-tasks-100およびJapanese MT-Bench等の評価には、Azure OpenAI Serviceを用いたLLM-as-a-Judge方式を適用)
上記のデータが示す通り、9B（90億パラメータ）の軽量モデル「Qwen3.5-Ricoh-9B-20260522」に関

しても、前作の8Bモデル（0.718）やベースモデル（0.762）を明確に上回る推論スコア（0.782）を達成している[2]。これは、推論能力の向上というリコーの学習手法の有効性が、モデルサイズに依存せず普遍的に機能していることを証明しており、スマートフォンクラスのエッジデバイスや小規模な部門サーバーでの高度な運用に向けた高いポテンシャルを示している。

6. オンプレミス実装のボトルネックを破壊する「量子化技術」

リコーのエンタープライズAI戦略の真髄は、単にベンチマークで高得点を叩き出す高性能なモデルを開発したことにとどまらず、それを日本の一般的な企業のインフラにおいて「現実的なコスト」でデプロイ（導入・展開）するための技術的アプローチを同時に提供している点にある。その中核をなすのが「量子化（Quantization）」技術の適用である。

通常、27Bクラスの大規模言語モデルをフル精度（FP32）あるいは半精度浮動小数点（FP16）で動作させる場合、膨大なVRAM（ビデオメモリ）帯域幅を要求され、NVIDIAのH100やA100といった、1基あたり数百万円から一千万円規模に達するハイエンドのデータセンター向けAIアクセラレータ（GPU）が不可欠となる[4]。これは、巨大なIT予算を持つ一部の大企業を除き、中堅・中小企業や、特定の部門単位でのオンプレミス導入において、事実上乗り越えられない致命的なコスト障壁となっていた。

このハードウェアの壁を打ち破るため、リコーはFP16版のリリースと並行して、モデルのパラメータを8bit（AWQ-W8A16）および4bit（AWQ-W4A16）の整数表現へと圧縮した「量子化モデル」を最初からラインナップとして用意している[2]。量子化とは、AIの推論を司る神経網（ウェイト）の計算精度をあえて意図的に粗く丸めることで、メモリ消費量と推論時の計算負荷を劇的に削減する高度な技術手法である[4]。一般的に、量子化を行うと計算の解像度が下がるため、推論能力の大幅な劣化が懸念される。しかし、驚くべきことに、表に示した通りリコーの最も圧縮率の高い4bit量子化モデル（AWQ-W4A16）であっても、JDocQA-Reasoningにおけるスコアは「0.868」を維持している[4]。この数値は、圧縮前のベースモデルであるQwen3.6-27Bの「0.858」を依然として上回っており、推論能力の劣化を実務上影響のない最小限のレベルに封じ込めつつ、モデルサイズを物理的に劇的に圧縮することに成功したことを意味する[4]。

この技術的ブレイクスルーにより、企業は超高価なサーバー専用GPUを複数台導入せずとも、市販のハイエンドデスクトップPCに搭載されるようなコンシューマー向け最上位GPU（例えば最大24GBのVRAMを持つモデルなど）や、ユニファイドメモリ環境を持つワークステーション等を用いて、商用クラウドに匹敵する高度な図表推論AIを、自社の閉域ネットワーク内で完全にオフラインで走らせることが可能となる[4]。これは、エンタープライズAIの「コストパフォーマンス革命」と呼ぶにふさわしい成果である。

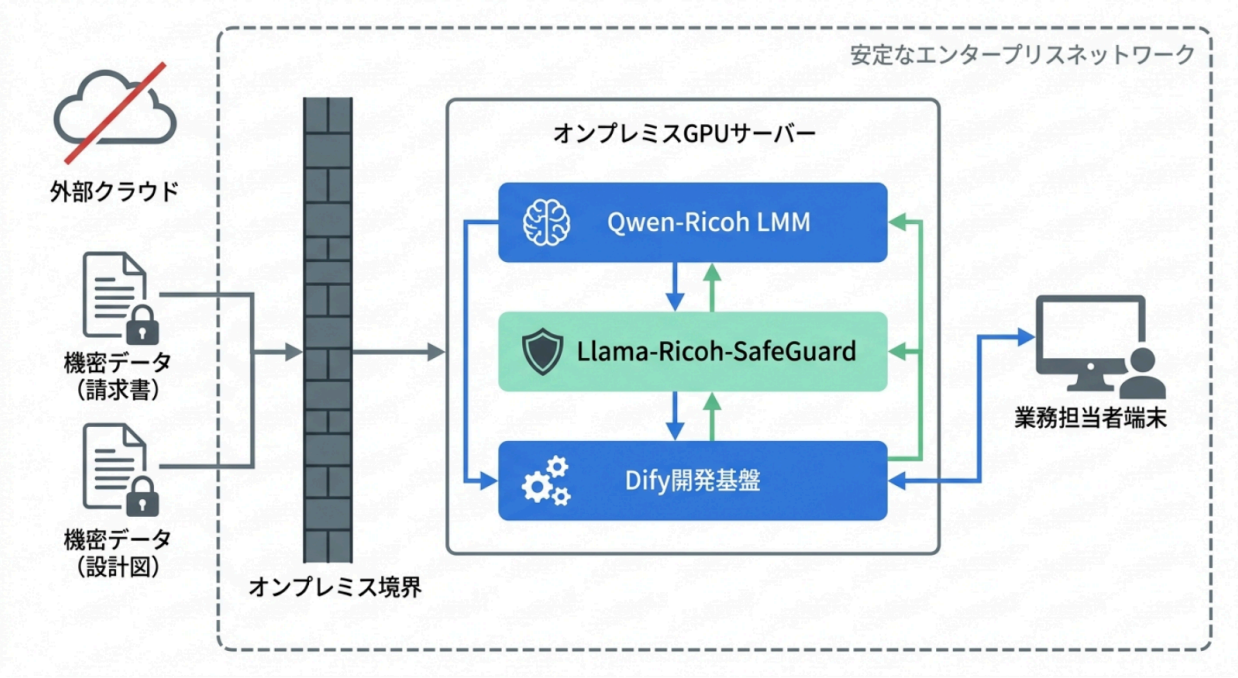
7. 安全なAIエコシステムの構築と「RICOH オンプレLLMスターターキット」による民主化

いかにAIモデルの推論性能が優れており、量子化によって軽量化されていたとしても、GitHubやHugging Faceにアップロードされたオープンソースの「重みデータ（Weights）」のファイル群だけを渡されて、セキュアな推論サーバーを自前で立ち上げ、業務アプリケーションを構築できる企業は、IT業界のごく一部に限られている。リコーは、この「導入と適用の壁」を越えるため、2026年6月下旬頃より、リコージャパン株式会社を通じてこれらの新モデル群を「RICOH オンプレLLMスターターキット」という包括的なソリューションに組み込んで提供する体制を整えた[2]。

このスターターキットは、単にAIモデルがインストールされた箱を販売するものではない。企業のDX

推進や業務効率化を加速させる生成AI導入の最初の一歩として、お客様の社内ネットワーク環境下へのGPUサーバーの物理的な設置構築から、LLMの安定的な動作環境の整備、導入時の教育、そして運用開始後の継続的な伴走支援に至るまでをワンストップで提供するパッケージソリューションである[27]。このソリューションの優秀性は広く認められており、「2025年日経優秀製品・サービス賞 最優秀賞」を受賞するなど、すでに高い市場評価を確立している[27]。
このキットのインフラストラクチャにおける重要なピースが、オープンソースのLLMアプリケーション開発プラットフォームである「Dify」の標準搭載である[27]。Difyを活用することで、高度なプログラミング知識を持たない現場の業務担当者（シチズンデベロッパー）であっても、視覚的なインターフェースを通じて、自社のドキュメントデータベースとAIを連携させた独自のチャットボットや、RAG（検索拡張生成）システムを構築することができる[29]。リコーはさらに、金融や医療といった業種向けに最適化されたDifyのテンプレートアプリケーションも提供しており、例えば那須赤十字病院における「退院サマリー作成支援」など、実際の医療現場のワークフローにオンプレミスAIを組み込む実証と支援をすでに行っている[29]。加えて、企業独自の専門用語、社内スラング、特有の言い回しなどをモデルの根幹に学習させる個別ファインチューニング（追加学習）の有償対応も用意されており、各企業の持つ「暗黙知」をAIの思考プロセスに直接組み込む経路を確保している[27]。

高セキュアなオンプレミスAI環境の統合アーキテクチャ



外部ネットワークから完全に切り離された自社環境内で、リコーのLMMとガードレールAIが連携し、企業の機密ドキュメントを安全に処理・解析するデータフロー。

さらに、リコーのエンタープライズAI戦略における「安全性」を全方位的に担保する上で欠かせないのが、本モデル群の発表の約2週間前、2026年5月20日に無償公開された自社開発のガードレール

機能組み込みLLM「Llama-Ricoh-SafeGuard-20260520」の存在である[31]。生成AIを企業業務に導入する際、データ漏洩と同等に懸念されるのが、AIの「暴走」や不適切な出力によるブランドリスクである。この「セーフガードモデル」は、ユーザーがAIに入力するプロンプトと、基盤AIが生成して返す出力回答の双方のトラフィックを常時監視し、暴力、差別的表現、プライバシー侵害など、14種類に分類される不適切な内容（有害情報）を自動で判別・検知し、必要に応じて遮断する機能を持つ[32]。圧倒的な図表解析能力を持つ「Qwen-Ricoh LMM」をエンジンとし、厳格な入出力監視機構である「SafeGuard」をブレーキとして両輪でオンプレミスのGPUサーバー内に配備することで、企業は情報漏洩のリスクをゼロにするだけでなく、倫理的リスクやコンプライアンス違反のリスクをも完全に自社のコントロール下に置くことができる。この二段構えのアーキテクチャこそが、AIのコア業務適用に対する経営陣の心理的ハードルを劇的に引き下げる決定的な要因となる。

8. 開発者コミュニティおよび産業界からの評価と圧倒的な透明性

リコーが発表したこれら一連の取り組みと成果は、国内外のAI研究者・開発者コミュニティ、およびAI導入を検討する産業界の双方から、極めて高い評価と関心を集めている。その根底にあるのは、同社の技術的誠実さと透明性に対する信頼である。

本プロジェクトに対する最大の評価ポイントの一つは、リコーが自社モデルの驚異的な性能証明の根拠として用いたベンチマークツール「JDocQA-Reasoning」のデータセットそのもの、および評価を行うためのプログラムコードを、Hugging Face上で「Apache License 2.0」および「CC BY-SA 4.0」という寛容なオープンソースライセンスのもと、完全無償で一般公開したことである[2]。AI業界において、企業が自社開発モデルの優位性をアピールする際、ブラックボックス化された独自の非公開データを用いて「自称スコア」を主張するケースが散見される。しかし、リコーは問題セットを公開し、さらに評価プロセスにおいてAzure OpenAI Serviceを用いた「LLM-as-a-Judge（AIによる自動採点）」方式を採用することで、第三者がいつでも自身の環境で客観的に検証し、再現性を確認できる環境を世界に向けて提供した[4]。この「徹底した透明性と検証可能性の担保」という姿勢は、AIの性能競争において非常に誠実なアプローチであると技術者層から絶賛されており、ひいては日本の生成AI基盤モデル全体の開発力底上げ（GENIACプロジェクトの真の目的）に直接的に貢献するエコシステム形成の動きとして高く評価されている[4]。

9. 産業別ユースケースと未来への展望：知の利活用の新時代

商用クラウドAIに肉薄する図表リーズニング性能が、量子化技術によって現実的なコストでオンプレミス環境に実装可能となったことにより、これまで「紙」や「PDFの図表」という物理的・フォーマット的制約に縛られていた様々な業界のコア業務における革新が、いよいよ現実味を帯びている[2]。

1. 製造業・エンジニアリング： 設計・開発部門において、図面（CADデータを出力した画像）と、それに対応する要求仕様書（テキスト）をQwen-Ricohモデルに同時に入力することで、寸法、材質、公差などの適合確認作業を大幅に自動化できる。また、製造ラインでの予期せぬトラブル発生時に、過去数十年分の膨大なトラブルシューティングマニュアルや報告書から、テキストの症状と機器の図解を横断的に解析し、瞬時に原因箇所の特定と対処法を提示する高度なエージェントとして機能する[2]。
2. 金融・保険業におけるコンプライアンスと審査： 保険商品の約款のような、極小の文字と複雑な条件分岐の表で構成されるドキュメントや、融資先の顧客から提出されるフォーマットが統一されていない多種多様な決算報告書を、人間の審査担当者に代わって高精度に読解す

る。財務諸表の数値の矛盾を検知したり、必要な要点を抽出してサマリーを作成する業務を、外部クラウドに顧客情報を一切出すことなく瞬時に実行する[2]。

3. 公共・自治体を通じた行政**DX**：住民から提出される手書きに近い申請書類や、独自のフォーマットを持つ各種届出、さらには複雑な統計図表が含まれる行政文書の処理支援を行う。職員のデータ入力や確認作業の負荷を劇的に低減し、行政サービスのスピードと質の向上を通じたデジタルトランスフォーメーション（DX）の基盤となる[2]。
4. 全産業を横断する次世代ナレッジマネジメント：企業内のファイルサーバーに眠る過去の膨大な会議資料や、PowerPoint等で作成された提案スライドからの自律的な論点抽出や傾向分析。従来のキーワードベースのテキスト検索技術では絶対に引っかからなかった「グラフが示す売上のトレンド」や「競合比較表における自社の強み」などをAIが視覚的に解釈し、メタデータとしてインデックス化することで、企業全体の集合知の活用レベルを飛躍的に押し上げる[2]。

10. 結論：日本のエンタープライズAI戦略の新たな羅針盤

リコーによるマルチモーダル大規模言語モデル「Qwen3.6-Ricoh-27B-20260522」および「Qwen3.5-Ricoh-9B-20260522」の開発と市場投入は、世界の生成AI進化の焦点が、単なる「パラメータ規模の拡大をひたすら追求する汎用的な巨大AGIの開発」から、「特定領域の深い理解と、運用コスト、そしてエンタープライズセキュリティのバランスを極限まで最適化した『実務特化型AI』の社会実装」へと明確にシフトしていることを鮮明に示している[7]。

本件から産業界全体が導き出すべき最重要のインサイトは、「優れたベースアーキテクチャの選定、目的に合致した良質なカリキュラム学習と強化学習の精緻化、そして適切なベンチマーク設定という三位一体のアプローチがあれば、ハードウェア制約の厳しいオンプレミス環境下（かつ量子化モデルを用いた運用）であっても、商用クラウドのフラグシップモデルと同等の、あるいはそれを上回る事業価値を創出することが十分に可能である」という証明である[4]。

「RICOH オンプレLLMスターターキット」や「Llama-Ricoh-SafeGuard」といった周辺ソリューション群との強固な統合により、リコーは単にベンチマークで勝つための優れた「AIの脳」を作っただけでなく、それを日本の一般的な企業が明日から安全かつ確実に実業務に組み込んで利用できる「AIのインフラと手足」までをパッケージ化して提供することに成功した[4]。今後、リコーが本プロジェクトで培った要素技術を、自社の提供する企業向けAIプラットフォーム「Hi.DEEN（ヒデン）」の高度化や、さらなる業種・業界特化型モデル群の継続的な開発・統合へと展開していくロードマップ[2]を踏まえれば、同社が日本のエンタープライズAI市場における強力なリーダーシップを発揮し、デファクトスタンダード（事実上の標準）の一つを形成していく可能性は極めて高い。

データガバナンスの確保と、AIによる業務革新の推進という、一見相反するジレンマに直面しているすべての企業にとって、リコーのQwen-Ricohモデル群の登場は、次世代の知の利活用（ナレッジマネジメント）に向けた最も現実的かつ強力な推進力となるであろう。

引用文献

1. リコー、マルチモーダル大規模言語モデル「Qwen3.6-Ricoh-27B ..., 6月 11, 2026にアクセス、https://jp.ricoh.com/release/2026/0605_1
2. リコー マルチモーダル大規模言語モデルを開発 軽量・オンプレミス ..., 6月 11, 2026にアクセス、https://www.oalife.co.jp/new_p/8779/
3. リコー, 6月 11, 2026にアクセス、

https://jp.ricoh.com/-/media/Ricoh/Sites/jp_ricoh/release/2026/pdf/0605_1.pdf?rev=e9af3fdd2d7f4150b0c6bb3f50c219b4&sc_lang=ja-JP
4. リコー「Qwen3.6-Ricoh-27B」開発、オンプレ対応LMMが独自日本 ..., 6月 11, 2026にアクセス、https://innovatopia.jp/ai/ai-news/107897/
5. Ricoh unveils open benchmark for AI reasoning on Japanese business documents, 6月 11, 2026にアクセス、https://jp.ibtimes.com/ricoh-unveils-open-benchmark-ai-reasoning-japanese-business-documents-101269
6. Ricoh Develops AI Models for Japanese Document Reasoning, Matching Gemini 3 pro Benchmark Score | IBTimes JP, 6月 11, 2026にアクセス、https://jp.ibtimes.com/ricoh-develops-ai-models-japanese-document-reasoning-matching-gemini-3-pro-benchmark-score-101546
7. Notable AI-Related News (May 31 - June 6)｜HIROE - note, 6月 11, 2026にアクセス、https://note.com/hiroe28/n/n611d35ffeb95?hl=en
8. リコー マルチモーダル大規模言語モデル「Qwen3.6-Ricoh-27B-20260522」開発 日本語リーズニング性能を強化し6月下旬より提供 |【印刷業界ニュース】ニュープリネット, 6月 11, 2026にアクセス、https://www.newprinet.co.jp/%E3%83%AA%E3%82%B3%E3%83%BC%E3%80%80%E3%83%9E%E3%83%AB%E3%83%81%E3%83%A2%E3%83%BC%E3%83%80%E3%83%AB%E5%A4%A7%E8%A6%8F%E6%A8%A1%E8%A8%80%E8%AA%9E%E3%83%A2%E3%83%87%E3%83%AB%E3%80%8Cqwen3-6-ricoh-27b-20
9. Qwen3.6 27B - API Pricing & Benchmarks | OpenRouter, 6月 11, 2026にアクセス、https://openrouter.ai/qwen/qwen3.6-27b
10. Qwen/Qwen3.6-27B - Hugging Face, 6月 11, 2026にアクセス、https://huggingface.co/Qwen/Qwen3.6-27B
11. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model, 6月 11, 2026にアクセス、https://qwen.ai/blog?id=qwen3.6-27b
12. Qwen3.5 & Qwen3.6 Usage Guide - vLLM Recipes, 6月 11, 2026にアクセス、https://docs.vllm.ai/projects/recipes/en/latest/Qwen/Qwen3.5.html
13. Qwen/Qwen3.6-35B-A3B - Hugging Face, 6月 11, 2026にアクセス、https://huggingface.co/Qwen/Qwen3.6-35B-A3B
14. qwen/qwen3.6-27b - LM Studio, 6月 11, 2026にアクセス、https://lmstudio.ai/models/qwen/qwen3.6-27b
15. Qwen3.5-9B: Specifications and GPU VRAM Requirements - ApX Machine Learning, 6月 11, 2026にアクセス、https://apxml.com/models/qwen35-9b
16. Qwen3.6-27B is here, 6月 11, 2026にアクセス、https://medium.com/data-science-in-your-pocket/qwen3-27b-is-here-197fb2256b97
17. Qwen3.6:27b is the first local model that actually holds up against Claude Code for me, 6月 11, 2026にアクセス、https://www.reddit.com/r/LocalLLM/comments/1t3pjkn/qwen3627b_is_the_first_local_model_that_actually/
18. Qwen3.6 is the large language model series developed by Qwen team, Alibaba Group. - GitHub, 6月 11, 2026にアクセス、https://github.com/QwenLM/Qwen3.6

19. Qwen3.5-9B - API Pricing & Benchmarks | OpenRouter, 6月 11, 2026にアクセス、https://openrouter.ai/qwen/qwen3.5-9b
20. Qwen/Qwen3.5-9B - Hugging Face, 6月 11, 2026にアクセス、https://huggingface.co/Qwen/Qwen3.5-9B
21. Qwen3.5-9B Surprised Me - Faster and More Reliable Than Larger Models for My Setup, 6月 11, 2026にアクセス、https://www.reddit.com/r/LocalLLM/comments/1rjm2kf/qwen359b_surprised_me_faster_and_more_reliable/
22. qwen3.5:9b - Ollama, 6月 11, 2026にアクセス、https://ollama.com/library/qwen3.5:9b
23. リコー、生成AIの推論性能を測る独自ベンチマークを無償公開 | リコーグループ 企業・IR, 6月 11, 2026にアクセス、https://jp.ricoh.com/release/2026/0529_1
24. ricoh-ai/JDocQA-Reasoning · Datasets at Hugging Face, 6月 11, 2026にアクセス、https://huggingface.co/datasets/ricoh-ai/JDocQA-Reasoning
25. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models - arXiv, 6月 11, 2026にアクセス、https://arxiv.org/html/2403.19454v1
26. Qwen3.5 - How to Run Locally | Unsloth Documentation, 6月 11, 2026にアクセス、https://unsloth.ai/docs/models/qwen3.5
27. 「RICOH オンプレLLMスターターキット」が「2025年日経優秀製品・サービス賞 最優秀賞」を受賞, 6月 11, 2026にアクセス、https://prtimes.jp/main/html/rd/p/000000172.000043114.html
28. RICOH オンプレLLMスターターキットとは？企業 ... - DENZAI-ZeuS, 6月 11, 2026にアクセス、https://denzai-zeus.com/ogawa-news/ricoh-on-prem-llm-starter-kit
29. リコージャパン、「オンプレLLMスターターキット」でDify活用を伴走支援するオプションを提供, 6月 11, 2026にアクセス、https://cloud.watch.impress.co.jp/docs/news/2053109.html
30. ricoh-ai/Qwen-3-VL-Ricoh-8B-20260227 - Hugging Face, 6月 11, 2026にアクセス、https://huggingface.co/ricoh-ai/Qwen-3-VL-Ricoh-8B-20260227
31. リコー、自社開発のセーフガードモデルを無償公開 | リコーグループ 企業・IR, 6月 11, 2026にアクセス、https://jp.ricoh.com/release/2026/0520_1
32. リコー、生成AIの有害な入出力を遮断するAIモデルを無償公開 - IT Leaders, 6月 11, 2026にアクセス、https://it.impress.co.jp/articles/-/29366
33. リコー、自社開発のセーフガードモデルを無償公開 - AFPBB News, 6月 11, 2026にアクセス、https://www.afpbb.com/articles/-/3635889
34. Datasets - Hugging Face, 6月 11, 2026にアクセス、https://huggingface.co/datasets?other=benchmark