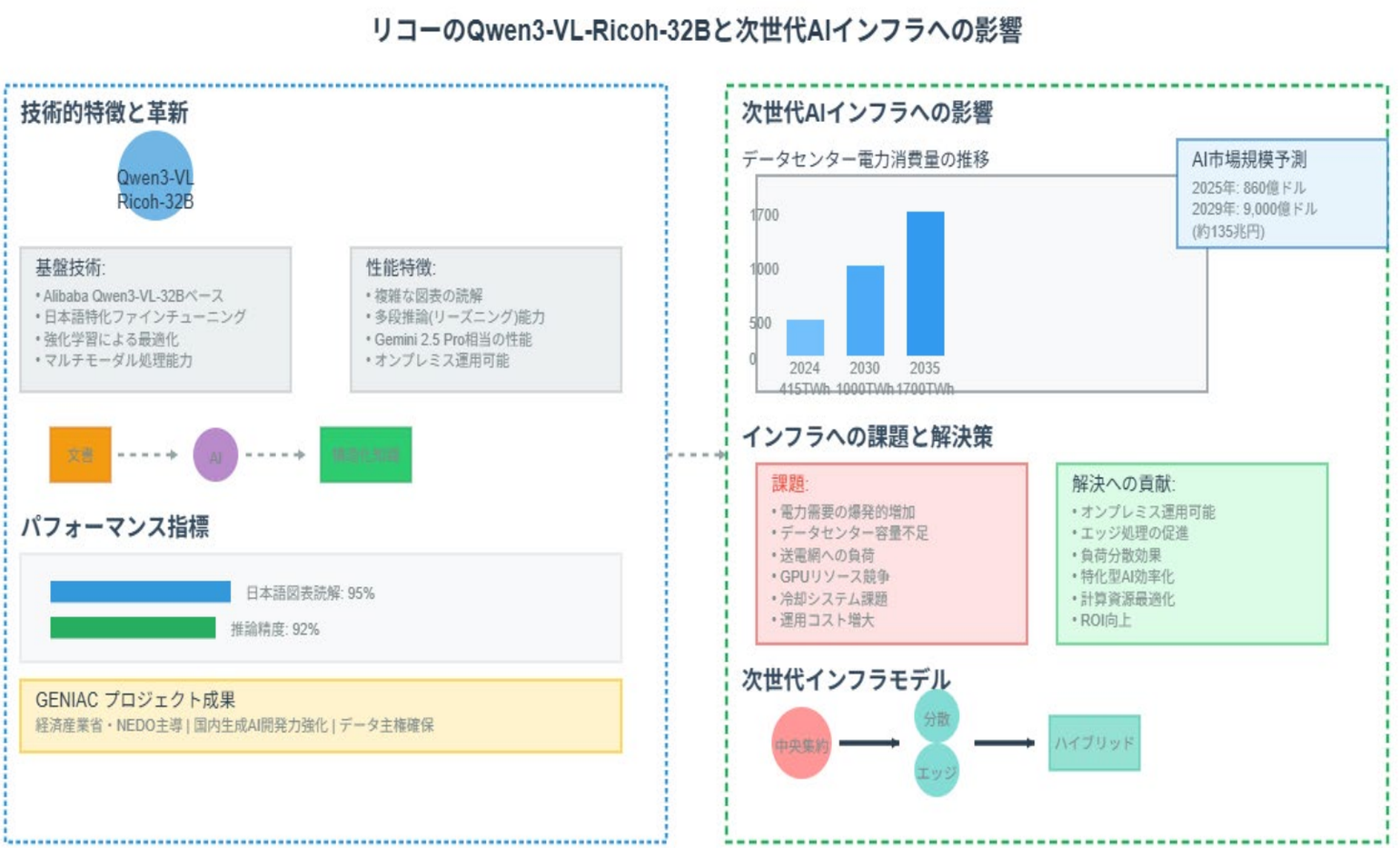
リコーの日本語特化型リーズニング LMM「Qwen3-VL-Ricoh-32B」の技術的評価と次世代 AI インフラへの影響

Felo AI



概要

株式会社リコーが開発した日本語特化型リーズニング LMM（大規模言語モデル）「Qwen3-VL-Ricoh-32B-20260227」は、経済産業省と NEDO が主導する国内生成 AI 開発強化プロジェクト「GENIAC」の成果として、AI 技術の新たな地平を切り拓いています 3 11 。このモデルは、Alibaba の「Qwen3-VL-32B-Instruct」をベースに、日本語の複雑な図表を含むドキュメントの読解能力を飛躍的に向上させたもので、その性能は Google の「Gemini 2.5 Pro」などの主要な商用モデルに匹敵します 9 13 37 。

本モデルの最大の特徴は、高度なマルチモーダル性能と多段推論（リーズニング）能力を日本語環境に最適化した点にあります 3 9 。これにより、従来は AI による自動処理が困難だった企業内の「暗黙知」を形式知化し、業務効率化や意思決定支援に活用する道が拓かれました 9 。さらに、オンプレミス環境での運用が可能なサイズに設計されており、データセキュリティや主権を重視する企業にとって現実的な選択肢を提供します 9 。

この技術的ブレークスルーは、AI が単なるツールから社会基盤へと移行する現代において、次世代 AI インフラの在り方に大きな影響を与えます。AI の社会実装が加速するにつれて、データセンターの電力消費量は爆発的に増加しており、電力供給が AI 発展のボトルネックとなりつつあります 8 16 20 。リコーのモデルのような、特定の用途に特化し、効率化された AI は、計算資源の最適化を促し、インフラへの負荷を軽減する可能性を秘めています。これは、GPU の確保競争から、電力効率や物理的実装力へと競争の軸足が移りつつある AI インフラ市場の新たなトレンドと合致するものです 16 23 。

詳細レポート

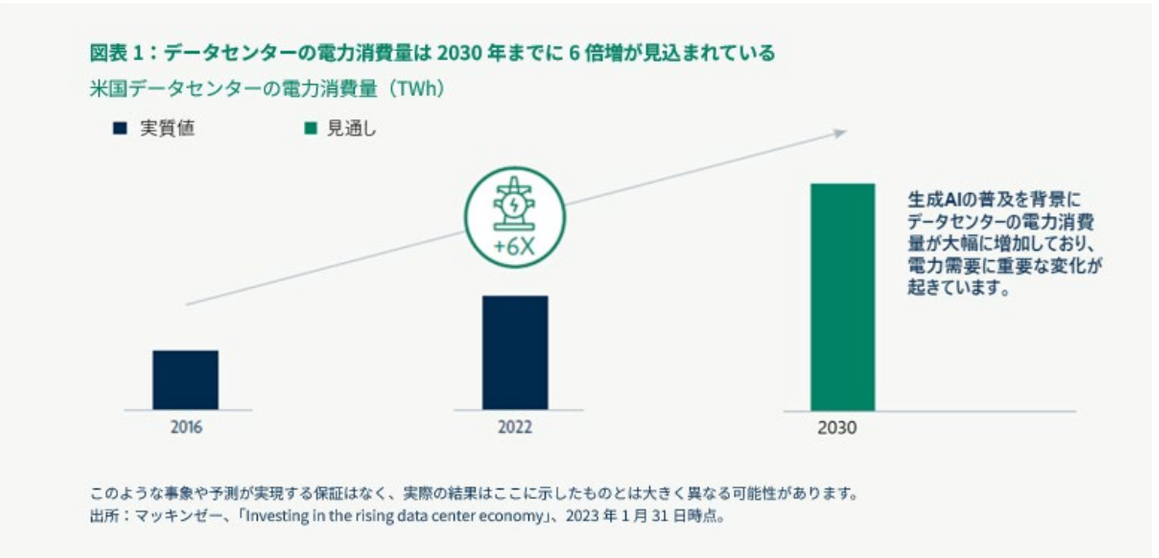
リコーのリーズニング LMM「Qwen3-VL-Ricoh-32B」の技術的詳細

開発背景と目的 リコーは、経済産業省と NEDO が実施する国内生成 AI 開発力強化プロジェクト「GENIAC」の第 3 期採択を受け、リーズニング性能を備えたマルチモーダル大規模言語モデル「Qwen3-VL-Ricoh-32B-20260227」の開発を完了しました 3 11 。この開発の背景には、企業内に存在する図や表組、画像を含む多様なドキュメント、いわゆる「暗黙知」を AI が活用できるようにするという課題がありました 9 。従来のテキスト検索だけでは意図した情報にたどり着けないケースが多く、マルチモーダルな処理能力が求められていました 9 。

技術的基盤と独自改良 本モデルは、Alibaba が開発したオープンソースのマルチモーダル LLM「Qwen3-VL-32B-Instruct」をベースモデルとして採用しています 9 13 31 。リコーはこれを基に、独自の改良を加えました。

- **ファインチューニングと強化学習**: ベースモデルの弱点を補強する学習データを用いてファインチューニングを実施 13 。さらに、独自の報酬関数を設定した強化学習を適用し、過学習を抑制しながら学習効率を高めました 9 。この報酬関数は、出力の正確さだけでなく、「日本語による推論プロセス」の出力にも高い報酬を与えるよう設計されており、単に正しい答えを出すだけでなく、「考え方」そのものを日本語環境に最適化しています 13 19 。
- **日本語への特化**: 思考プロセス自体を日本語化することで、日本語文書の読み取り精度を向上させました 9 。これ

により、ユーザーは回答の判断根拠や前提条件を日本語で確認でき、実務利用における信頼性が大幅に高まっています 9 13 。



卓越した性能と評価 「Qwen3-VL-Ricoh-32B」は、複数ページにまたがる図表を関連付けて理解し、読解難易度の高い質問に対しても高精度な回答を生成する能力を持ちます 9 。例えば、科学材料の複雑なフロー図を正確にトレースしたり、複数ページのグラフから傾向を読み取ったりすることが可能です 9 。
性能評価においては、リコーが独自に開発したベンチマークツールが使用されました 3 。このツールは、既存のデータセットに加え、日本企業特有の複雑な図表をテストデータの中核に据えており、より実用的な環境でのリーズニング性能を測定します 3 。この評価において、本モデルは Google の「Gemini 2.5 Pro」といった最先端の大型商用モデルと同等のスコアを記録しました（2026 年 2 月 17 日時点）9 13 37 。
展開とアクセシビリティ リコーは、320 億パラメータの基本モデルに加え、その技術を応用した 80 億パラメータの軽量モデル「Qwen3-VL-Ricoh-8B-20260227」を無償で公開しています 3 11 。また、基本モデルはパラメータサイズを社内サーバーに搭載可能なレベルに抑えており、機密性の高い社内文書を扱うためのオンプレミスでの活用を可能にしています 9 。

AI の進化とインフラへの要求増大

AI の「インフラ化」と市場の爆発的成長 2026 年現在、AI は個別のツール群から、検索、生産性向上、メディア制作などを統合した「エージェント主導のスーパーアプリ」という社会基盤（インフラ）へと変貌を遂げています 4 。この構造的変化は、グローバルな AI インフラ投資の急増を促しています。世界の AI インフラ市場は、2025 年第 3 四半期に過去

最高の 860 億ドルに達し、2029 年には 9,000 億ドル（約 135 兆円）規模に達すると予測されています 16 。ゴールドマン・サックスも、世界の AI 関連支出が年率 30%超で成長し、5000 億ドルを超えると見ています 4 。
この投資の焦点は、単なる GPU の調達競争から、それを支える電力、冷却システム、データセンター建屋といった物理的な実装力へとシフトしています 16 23 。

データセンターの巨大化と電力危機 AI の高度化は、膨大な計算資源を必要とし、データセンターの電力需要を前例のないレベルにまで押し上げています 8 14 。

項目	従来型データセンター	AI データセンター（現在）	AI データセンター（次世代）
ラックあたり消費電力	2〜4 kW	最大 140 kW	600 kW〜1 MW
データセンター規模	10 MW 程度	500 MW 未満	最大 2 GW（計画中）

*出典: McKinsey & Company, Deloitte2 10 *

この電力需要の急増は、既存の電力網に深刻な負荷をかけています 8 。国際エネルギー機関（IEA）は、世界のデータセンターの電力消費量が 2024 年の 415TWh から、2035 年には最大で 1,700TWh に達する可能性があると予測しています 20 。特に、データセンターが集中する地域では、局所的な電力需給の逼迫が深刻な問題となっています 20 。米国では、データセンター建設への支出が過去 3 年で 3 倍に増加しており、電力と送電網の容量がインフラ構築の最大の課題として認識されています 6 10 。

「Qwen3-VL-Ricoh-32B」が次世代 AI インフラに与える影響

オンプレミス利用によるインフラ負荷の分散 リコーのモデルは、高性能でありながらオンプレミス環境で動作可能な点が大きな特徴です 9 。これにより、全てのデータをクラウド上の巨大データセンターに集約する必要がなくなり、インフラ負荷の分散に貢献します。

- **データ主権とセキュリティ**: 機密情報を外部に出すことなく AI を活用できるため、金融、医療、製造業など、セキュリティ要件が厳しい業界での AI 導入が加速します 9 16 。
- **エッジコンピューティングの促進**: データ発生源の近くで処理を行うエッジコンピューティングへの移行を後押しします。これは、電力制約が厳しくなる中で、省電力な推論専用チップの活用と並行して進む重要なトレンドです 16 。

特化型 AI による計算資源の最適化 あらゆるタスクを一つの超巨大モデルで処理するのではなく、「Qwen3-VL-Ricoh-32B」のように特定のドメイン（日本語の図表読解）に特化した高効率モデルを活用することで、計算資源の最適化が可能になります 12 。

- **投資対効果（ROI）の向上**: 用途に応じて最適なモデルを選択することで、不要な計算コストを削減し、AI インフラへの投資対効果を高めることができます 16 22 。2026 年以降、AI 投資は初期の衝動的な導入フェーズから、ROI を重視した最適化のフェーズへと移行すると見られています 16 。

国内 AI 開発力とインフラ自律性への貢献 GENIAC プロジェクトから生まれた本モデルは、日本の AI 開発力を強化し、海外の巨大プラットフォームへの依存を低減させる上で重要な役割を果たします 3 11 。

- **ハイブリッド戦略の実現**: 米国勢のプラットフォームを活用しつつも、データ主権や経済安全保障の観点から、自律的な計算基盤を国内に保持するという「ハイブリッド戦略」の具体的な選択肢となります 16 26 。

新たなインフラビジネスモデルの創出 AI の普及は、電力供給のあり方そのものにも変革を迫っています。データセンターと発電所を併設するコロケーションや、電力網の状況に応じて計算タスクを柔軟に移動させる「計算タスクのモビリティ」といった革新的なアプローチが検討されています 10 。リコーのモデルのように、企業の現場で実用的に使える AI が増えることは、こうした新しいインフラソリューションへの需要を喚起し、AI とエネルギーの協調を前提とした新たなビジネスモデルの創出を加速させるでしょう 20 。

1. arXiv:2505.09388v1 【cs.CL】 14 May 2025

2. The AI infrastructure of the future
3. “はたらく”を支えるリコーの大規模言語モデル（LLM）
4. AI はツールから「インフラ」へ、今後 10 年の競争優位を ...
5. https://scholar.google.com/citations?view_op=view_...
6. How AI Is Transforming Data Centers and Ramping Up ...
7. GENIAC」第 3 期においてリーズニング性能を備えたマルチ ...
8. インフラストラクチャー：AI の普及を背景とした電力需要の増加
9. リコー、企業の暗黙知を AI 対応にするマルチモーダル新モデル
10. Can US infrastructure keep up with the AI economy?
11. リコー、「GENIAC」第 3 期においてリーズニング性能を備えた ...
12. AI の社会実装と加速するインフラ投資
13. 日本語で推論”できるマルチモーダル LLM を開発 「Gemini 2.5 ...
14. Infrastructure of the Future: The Impact of AI and the Cloud
15. リコー、「GENIAC」第 3 期においてリーズニング性能を備えた ...
16. AI インフラ市場、2029 年に 135 兆円規模へ爆発的拡大
17. リコー、図表を読み取れるリーズニング性能を備えたマルチ ...
18. AI Infrastructure: New Opportunity, but Old Principles Apply
19. リコー、強化学習で多段推論を獲得したビジネス文書向け LLM ...
20. 2035 年に向けた AI・デジタル技術とエネルギーの協調と課題
21. リコー、複雑な図表も読み解くマルチモーダル AI を開発
22. What is AI Infrastructure?
23. AI インフラ投資 2028 年予測｜96 兆円市場の物理制約を突破 ...
24. Executive summary – Energy and AI – Analysis – IEA
25. ソフトバンクの次世代 AI インフラを自動運転やロボットでどう使う ...
26. デジタル社会を支えるデータセンター（1）米国・EU・日本の政策
27. 岐路に立つ AI：AI がどのようにインフラストラクチャー投資を再 ...
28. GENIAC」第 3 期においてリーズニング性能を備えたマルチ ...
29. “はたらく”を支えるリコーの大規模言語モデル（LLM）
30. リコー、図表を読み取れるリーズニング性能を備えたマルチ ...
31. リコー、企業の暗黙知を AI 対応にするマルチモーダル新モデル
32. リコー、「GENIAC」第 3 期においてリーズニング性能を備えた ...
33. 楽天 AI『DeepSeek 採用』の真相と日本発 LLM の現在

34. リコー、複雑な図表も読み解くマルチモーダル AI を開発
35. リコー、「文書×AI」を現場仕様に──Qwen2.5－VL－32B 基盤
36. GENIAC」第 3 期においてリーズニング ...
37. リコー、日本語推論特化のマルチモーダル AI モデルを発表
38. 日本語で推論”できるマルチモーダル LLM を開発 「Gemini 2.5 ...
39. リコー、「GENIAC」第 3 期においてリーズニング性能を備えた ...
40. GENIAC」第 3 期においてリーズニング性能を備えたマルチ ...
41. リコー、日本語推論特化のマルチモーダル AI モデルを発表