リコー製マルチモーダル大規模言語モデル「Qwen3.6-Ricoh-27B-20260522」「Qwen3.5-Ricoh-9B-20260522」発表内容と評判・評価に関する調査報告

2026 年 6 月 11 日

Claude Fable 5

要旨

株式会社リコーは 2026 年 6 月 5 日、中国アリババクラウドのオープンモデル「Qwen3.6-27B」「Qwen3.5-9B」をベースに、日本語の図表読解・多段推論(リーズニング)性能を強化したオンプレミス対応のマルチモーダル大規模言語モデル(LMM)2 機種「Qwen3.6-Ricoh-27B-20260522」および「Qwen3.5-Ricoh-9B-20260522」を発表した [1]。自社開発の図表読解ベンチマーク「JDocQA-Reasoning」において、27B 版はスコア 0.881 を記録し、参考値として併記された Google「Gemini 3 Pro Preview」の 0.880 とほぼ同水準に達したとする [1]。ただし、この比較はリコー自作のベンチマーク・LLM-as-a-Judge 方式・自社測定という限定条件下の図表読解スコアであり、汎用性能全体での優位を意味するものではない点に留意が必要である [6]。本モデルは 2026 年 6 月下旬頃から「RICOH オンプレ LLM スターターキット」に搭載され、リコージャパンが提供する予定である [1]。

1. 発表の概要

発表主体は株式会社リコー(社長執行役員:大山晃)であり、発表日は 2026 年 6 月 5 日である [1]。発表されたのは、フラッグシップの「Qwen3.6-Ricoh-27B-20260522」(270 億パラメータ級)と軽量版の「Qwen3.5-Ricoh-9B-20260522」(90 億パラメータ)の 2 機種である [1]。いずれも経済産業省・NEDO による生成 AI 基盤モデル開発支援事業「GENIAC」の第 2 期・第 3 期の成果と位置づけられている [1,2]。

両モデルの特徴は、(1)テキストに加えて図表を含む文書画像を理解するマルチモーダル性、(2)複数の手がかりを統合して結論を導く多段推論(リーズニング)性能の日本語における強化、(3)軽量化によりオンプレミス環境で運用可能であること、の 3 点に集約される [1,4,5]。リコーは、機

密情報を社外に出せない企業・組織が自社環境内で商用クラウド AI に迫る日本語性能を利用できる点を訴求している[1,5]。

2. モデルの技術的詳細

2.1 ベースモデルとリコーによる追加開発

ベースモデルは、アリババクラウドが 2026 年 4 月 22 日に Hugging Face および ModelScope で公開した「Qwen3.6-27B」(27.8B パラメータの密(dense)マルチモーダルモデル、Apache 2.0 ライセンス)と「Qwen3.5-9B」である[1,7]。Qwen3.6-27B は、ネイティブ 262,144 トークンの文脈長(YaRN 拡張で約 101 万トークン)、ハイブリッドアテンション機構、思考/非思考モードの統合、201 言語対応などを特徴とし、コーディング性能では旧フラッグシップ級モデルを上回るベンチマーク結果が報告されている[7,8]。

リコーは、このベースモデルに対し独自の強化学習とカリキュラム学習を高度化して適用した。公式リリースによれば、強化学習では報酬関数設計を精緻化し、論理的推論の強化と過学習抑制を両立させ、カリキュラム学習では難易度設計を高度化し、複雑なドキュメントの読解能力を向上させたとする[1]。

2.2 軽量化とオンプレミス運用

FP16 版に加え、8bit 量子化版(AWQ-W8A16)および 4bit 量子化版(AWQ-W4A16)が用意され、GPU コストに応じた構成選択が可能である[1]。なお、リコー公式リリースには具体的な GPU・ハードウェア要件は明示されていないが、ベースモデルの Qwen3.6-27B は、FP8 量子化で単一の H200 または 2 枚の A100 で動作し、4bit 級量子化(Q4_K_M)では約 16.8GB の VRAM(16GB 級 GPU)で動作するとされており[9]、リコー版もこのコンパクトなハードウェア要件をおおむね継承するものと推定される(本段落のハードウェア要件はベースモデルの一般的要件からの推定を含む)。

2.3 提供形態と想定ユースケース

両モデルは 2026 年 6 月下旬頃から「RICOH オンプレ LLM スターターキット」に搭載され、リコージャパンが提供する予定である[1]。同キットは 2025 年 4 月 7 日に提供開始されたもので、GPU サーバ 1 台でリコー製 LLM とノーコード AI 開発基盤「Dify」等をプリインストールして提供する形態をとり、基本構成は一部門(30 人程度)を想定、価格は「1500 万円から」と報じら

れている(本発表モデル搭載構成の価格は確認できなかった)[11]。業種・業務別のファインチューニングにも対応する [1]。想定ユースケースとして、製造業における設計図と要求仕様の適合確認、金融・保険における約款・報告書の読解と要点抽出、公共・自治体における申請書類・行政文書処理、会議・提案資料からの情報抽出・分析などが挙げられている [1,5]。

3. ベンチマーク結果と評価方法

3.1 図表読解性能

リコー公式リリースに掲載された図表読解ベンチマークの結果は以下のとおりである [1]。「JDocQA-Reasoning」は GENIAC 第 3 期においてリコーが独自開発したベンチマーク(テストデータ 1,362 問、図表または図表とテキストから多段推論で回答を導く問題で構成)であり、2026 年 5 月 29 日から Hugging Face で無償公開されている [1,2]。

モデル	JDocQA-Reasoning(1.0 満点)	JDocQA(5 点満点)
Qwen3.6-Ricoh-27B-20260522	**0.881**	**4.22**
同 8bit 量子化版(AWQ-W8A16)	0.873	4.21
同 4bit 量子化版(AWQ-W4A16)	0.868	4.20
ベースモデル Qwen3.6-27B	0.858	4.15
(参考)Gemini 3 Pro Preview	0.880	4.24
(参考)Gemini 2.5 Pro	0.838	4.08
(参考)GPT-5.2	0.731	3.93
(参考)前作 Qwen3-VL-Ricoh-32B-20260227	0.826	4.08
Qwen3.5-Ricoh-9B-20260522	0.782	4.00
ベースモデル Qwen3.5-9B	0.762	3.89

3.2 日本語テキスト性能

日本語テキストタスクのベンチマーク結果は以下のとおりであり、いずれもベースモデルからの性能向上が示されている [1]。

モデル	ELYZA-tasks-100(5 点満点)	Japanese MT-Bench(10 点満点)
Qwen3.6-Ricoh-27B-20260522	4.64	9.48

ベースモデル Qwen3.6-27B	4.58	9.35
Qwen3.5-Ricoh-9B-20260522	3.95	7.93
ベースモデル Qwen3.5-9B	3.76	7.65

3.3 評価条件と留意点

評価は Azure OpenAI Service の GPT-4.1(JDocQA のみ GPT-4o)を採点役とする LLM-as-a-Judge 方式で行われ、各ベンチマーク 5 回実施の平均値が採用されている [1]。また、Gemini 3 Pro Preview および GPT-5.2 のスコアは前作リリース(2026 年 3 月 30 日)時点の参考値であり、評価時期・条件が異なる [1,2]。すなわち「Gemini 3 Pro Preview 並み」との訴求は、リコー自作ベンチマーク・自社測定・参考値比較という限定条件下の図表読解スコアに基づくものであり、第三者による独立検証結果は本報告作成時点で確認できなかった [6]。

4. リコーの LLM/LMM 開発の系譜と本発表の位置づけ

リコーは 2022 年に LLM 研究に着手し、2023 年 3 月の独自 LLM(60 億パラメータ)発表以降、2024 年 1 月に 130 億パラメータ版、2024 年 8 月に日英中 3 言語対応の 700 億パラメータ LLM(Llama-3-Swallow-70B ベース)、2025 年 4 月に GPT-4o 相当を主張するモデルへと段階的に発展させてきた [2,3]。マルチモーダル化については、2025 年 6 月に GENIAC 第 2 期で LMM 基本モデルの開発を完了し、2026 年 1 月に Qwen2.5-VL-32B ベースの 32B LMM[3]、2026 年 3 月 30 日にリーズニング対応の「Qwen3-VL-Ricoh-32B-20260227」(Gemini 2.5 Pro 匹敵を主張)を発表している [2]。

本発表(6 月 5 日)は、ベースモデルを Qwen3-VL 系 32B から最新の Qwen3.6-27B へ刷新し、5 月に公開した自社ベンチマーク JDocQA-Reasoning で前作(0.826)を上回るスコア(0.881)を示した連続的な進化であり、比較対象も前作時の Gemini 2.5 Pro から最新の Gemini 3 Pro Preview へ引き上げられた [1,2]。

5. オンプレミス LLM 需要の背景

製造業の設計図面、金融機関の約款・稟議書、官公庁の行政文書など、機密性の高い文書を社外クラウドに送信できない組織において、自社環境内で完結する生成 AI への需要が拡大している [5,11]。量子化により必要 GPU 数とコストを抑制でき、クラウドの従量課金も発生しないため、

中堅企業でも導入しやすい。GENIAC 第 3 期では「軽量・業界特化」型の AI 開発が主要な傾向となっており(24 件採択)、リコーの取り組みはその代表例と位置づけられる [10]。

6. 世の中の評判・評価

6.1 メディア報道

報道は概ねプレスリリースの要約にとどまる。AI Watch(インプレス)[4]、ZDNET Japan[5] 等が報じ、ZDNET Japan は、社内ベンチマーク「JDocQA-Reasoning」での評価において Google の「Gemini 3 Pro Preview」に匹敵する水準に達し、従来のオープン系モデルを上回る性能を示したと要約している [5]。日経クロステック、ASCII、CNET Japan、PC Watch 等による本件専用記事は本報告作成時点で確認できなかった。

6.2 批評的論評

最もまとまった批評は Web メディア innovaTopia(2026 年 6 月 9 日)によるものである [6]。同記事は、(1)ベンチマークがリコー自身の開発・公開によるものであり、評価方式も AI が採点役を担う LLM-as-a-Judge である点、(2)比較対象の Gemini や GPT のスコアが第三者の公式評価ではなくリコー側測定の参考値である点を「割り引いて見るべき」と指摘する一方、データセットを Hugging Face で無償公開し第三者検証を可能にした姿勢を「誠実なアプローチ」と評価した [6]。また、「日本語文書の図表読解という限定された土俵での結果であり、汎用的な性能全体でリコーのモデルが Gemini 3 Pro Preview を上回ったわけではない」こと、土台が中国アリババ製の Qwen であり、LLM の国産化を語る際に基盤が海外モデルに依存している現実は技術主権の議論で避けて通れないことも論じている [6]。

6.3 技術コミュニティ・専門家の反応

X(旧 Twitter)、はてなブックマーク、Qiita、Zenn、note 上で、本発表(6 月 5 日)を名指しした技術コミュニティの個別反応や、固有名の AI 専門家・アナリストによる論評は本報告作成時点で確認できなかった。なお、Qwen ベースであることへの技術主権・セキュリティ上の懸念や、LLM-as-a-Judge 評価方式への批判は技術コミュニティに一般論として存在するが [8]、本件への直接の論評ではない。

7. 競合状況と市場における位置づけ

最も直接的な競合は、本発表の約 2 週間前(2026 年 5 月 19 日)に NTT が発表した「tsuzumi 2

Vision」である。同モデルは純国産(フルスクラッチ)開発で、300 億パラメータのテキスト基盤に図表理解アダプタを追加し、1GPU(NVIDIA A100 相当・約 40GB メモリ)で動作して図表入り日本語ビジネス文書に対応する点で、「図表×オンプレミス×日本語」という土俵がリコーと重なる。このほか、国産・日本語特化 LLM 市場には NEC(cotomi)、Preferred Networks(PLaMo)、富士通(Takane)、ELYZA、ストックマーク、サイバーエージェント、rinna 等が存在する。

開発アプローチの観点では、NTT・PFN がデータ主権を重視したフルスクラッチ路線をとるのに対し、リコーは ELYZA・rinna・ABEJA 等と同様に海外オープンモデル(Llama/Qwen/Gemma)への日本語追加学習路線をとる。その中でリコーは「Qwen ベース+図表読解+オンプレミス+業種特化+ベンチマーク無償公開」という組み合わせに独自色を出している [1,2,6]。

8. 考察:導入検討・知財実務上の留意点

第一に、性能訴求の割り引き評価である。「Gemini 3 Pro Preview 並み」は限定条件下の図表読解スコアであり、導入検討にあたっては自社の実文書(設計図・約款・帳票等)での PoC 検証が必須である。ベンチマークデータは Hugging Face で公開され第三者検証が可能であるため、調達前に独自データでの再評価が推奨される [1,6]。なお、4bit 量子化版でもスコア 0.868 と FP16 版(0.881)からの劣化が小さい点は、GPU 予算が限られる組織にとって導入判断上ポジティブな材料である [1]。

第二に、知財・契約デューデリジェンスである。オンプレミス運用によりクラウド経由のデータ送信リスクは解消されるが、(1)Qwen のライセンス(現状 Apache 2.0 だが、将来のライセンス方針変更の可能性が技術コミュニティで議論されている [8])、(2)中国製ベースモデルに由来する出力傾向・バイアスの残存可能性、(3)学習データの権利クリアランス、を確認項目とすべきである。提供契約時には、モデルのライセンス条件、権利継承、再学習・ファインチューニングの可否と成果物の帰属を明示的に確認することが推奨される。

第三に、判断を変えうる今後の注目点として、(1)第三者の独立評価(Nejumi Leaderboard 等)における図表読解優位の再現性、(2)NTT tsuzumi 2 Vision 等の純国産競合との自社文書での実測比較、(3)スターターキット価格とクラウド API 従量課金の TCO 分岐点、(4)Qwen のライセンス動向、が挙げられる。

9. 結論

本発表は、「日本語の図表文書読解」という具体的ユースケースに焦点を絞り、海外オープンモデルの改良によって商用クラウド AI に迫る性能をオンプレミスで実現するという、GENIAC 第 3 期の「軽量・業界特化」路線を体現した動きである [1,10]。性能主張には自社測定・自作ベンチマークという留保が付くものの、評価データの無償公開により第三者検証の道を開いている点は評価でき [6]、機密文書を扱う製造業・金融・公共分野におけるオンプレミス LLM 選択肢として、純国産路線の NTT tsuzumi 2 Vision 等と並ぶ有力な比較検討対象になると考えられる。

参考文献

1. 株式会社リコー ニュースリリース「リコー、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」を開発」2026 年 6 月 5 日 https://jp.ricoh.com/release/2026/0605_1
2. 株式会社リコー ニュースリリース「リコー、「GENIAC」第 3 期において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を開発」2026 年 3 月 30 日 https://jp.ricoh.com/release/2026/0330_1
3. 株式会社リコー ニュースリリース「リコー、「Qwen2.5-VL-32B-Instruct」ベースのマルチモーダル LLM を開発」2026 年 1 月 8 日 https://jp.ricoh.com/release/2026/0108_2
4. AI Watch(インプレス)「リコー、日本語リーズニング性能を強化した LMM「Qwen3.6-Ricoh-27B-20260522」を開発」2026 年 6 月 https://ai.watch.impress.co.jp/docs/news/2114771.html
5. ZDNET Japan(Yahoo!ニュース配信)「リコー、オンプレ対応マルチモーダル LLM を開発--軽量モデルでクラウド AI 級の日本語推論性能を実現」2026 年 6 月 https://news.yahoo.co.jp/articles/a4243456cbffff434d49994c5483e27ffa7773ac
6. innovaTopia「リコー「Qwen3.6-Ricoh-27B」開発、オンプレ対応 LMM が独自日本語推論ベンチマークで Gemini 3 Pro Preview 参考値と同水準」2026 年 6 月 9 日 https://innovatopia.jp/ai/ai-news/107897/
7. LLM Stats "Qwen3.6-27B Benchmarks, Pricing & Context Window" https://llm-stats.com/models/qwen3.6-27b
8. note(zephel01)「Qwen3.6-27B リリース!長所・短所、モデル比較、X/Reddit での評判まとめ」2026 年 4 月 https://note.com/zephel01/n/n5e76d565696b
9. Will It Run AI "Qwen 3.6 27B VRAM & Hardware Requirements — Dense 27B GPU Guide (2026)" https://willitrunai.com/blog/qwen-3-6-27b-vram-requirements
10. 日経クロステック「鍵は「軽量・業界特化」の AI 開発 経産省の GENIAC 第 3 期に 24 件採択」 https://xtech.nikkei.com/atcl/nxt/mag/nc/18/020800017/080801302/
11. 日経クロステック「リコージャパンがオンプレでの生成 AI 導入サービス、LLM 設定済み GPU サーバー提供」2025 年 4 月 https://xtech.nikkei.com/atcl/nxt/column/18/00001/10491/

(注)各 URL の最終閲覧日は 2026 年 6 月 11 日。本報告の数値・日付・モデル名は、リコー公式ニュースリリース(文献 1)を一次情報として記載した。確認できなかった事項はその旨を本文中に明記している。