ChatGPT

リコー製マルチモーダルLMM調査報告書

エグゼクティブサマリ

株式会社リコーは、2026年6月5日にマルチモーダル大規模言語モデル「Qwen3.6-Ricoh-27B-20260522」と「Qwen3.5-Ricoh-9B-20260522」を発表した。前者は Alibaba Cloud の Qwen3.6-27B を、後者は Qwen3.5-9B をベースにし、日本語でのリーズニング性能を強化した企業向け LMM と位置づけられている。リコーは両モデルを、図表を含む企業ドキュメントの読解・推論に強い「オンプレミス導入可能な LMM」として位置づけ、2026年6月下旬から「RICOH オンプレLLMスターターキット」での提供を予定している。 1

現時点で最も強い根拠は、**リコー公式の評価結果**である。そこでは 27B モデルが JDocQA-Reasoning で 0.881、JDocQA で 4.22、ELYZA-tasks-100 で 4.64、Japanese MT-Bench で 9.48 を記録し、同社は「Gemini 3 Pro Preview に近い性能水準」と説明している。一方で、この比較は Ricoh 自身が実施した **LLM-as-a-Judge 評価**であり、Azure OpenAI Service の gpt-4.1 と gpt-4o を審判モデルに用い、Gemini 2.5 Pro / Gemini 3 Pro Preview の値は 2026年3月30日時点の参考値を再掲したものである。したがって、**27B/9B 両モデルの価値は高いが、公開されている証拠の性質は「自社評価中心」であり、第三者の同一条件再評価は未成熟**である、というのが最も厳密な読み方になる。 2

導入判断としては、**日本語の図表・文書読解を社内閉域で回したい企業**には有力候補である。特に製造、金融・保険、公共・自治体、ドキュメント中心業務での適合性が高い。他方で、**価格、正確なハードウェア BOM、現行2モデルの公開ライセンス、個別安全性評価の詳細、第三者ベンチマーク**は未公開または未成熟であり、PoC なしに本番採用を即断すべき段階ではない。結論として、27B は「高精度オンプレ文書推論」、9B は「より広いハードウェア環境での実務導入」に向くが、企業導入ではクラウド最先端モデルとの**併用検証**が合理的である。 3

本報告書で明示的に扱う調査項目は、次の七点である。

- 公式情報
- モデル仕様
- 性能評価
- 導入・運用面
- 世間の評判
- 法的・倫理的観点
- 推奨・結論

本報告書では、**「未公開」**は公式資料に開示が見当たらない項目、**「推定」**は公開された仕様値から本報告書が算出した補助的見積りを意味する。

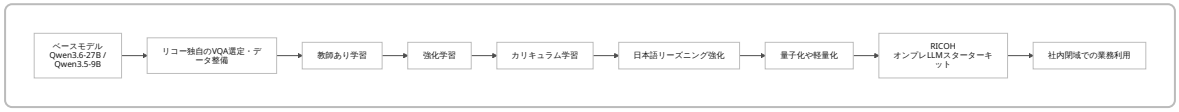
公式情報と開発位置づけ

最上位の一次情報は、2026年6月5日のリコーのニュースリリースである。ここでリコーは、27B モデルを「Qwen3.6-27B ベース」、9B モデルを「Qwen3.5-9B ベース」と明示し、いずれも日本語リーズニング性能を強化したマルチモーダル大規模言語モデルとして発表している。さらに、両モデルは「図表を含む多様なドキュメントを高精度に読み取り、推論すること」が主眼であり、提供形態はクラウド API ではなく、**オンプレミス導入可能なスターターキット**であることが明示されている。 2

この発表は、リコーが GENIAC 第2期・第3期で進めてきた LMM 開発の延長線上にある。技術ページでは、LMM_27B を「GENIAC 基盤モデル開発第3期で開発した LMM_32B の後継」、LMM_9B を「第3期で開発・公開した LMM_8B の後継」と整理し、前者を高精度かつ低運用コスト、後者を「汎用 GPU サーバー1台構成でも活用できる」コンパクトモデルと説明している。つまり、今回の 27B / 9B は、**第3期で確立した文書推論系 LMM を、より実装寄りに再編した商用近接モデル群**とみるのが妥当である。 4

リコーの技術ページは、これらのモデル群を「日本企業特有の複雑な図表を含む文書」に対応する企業独自生成 AI と位置づける。そこでは、公開データに加え、研究機関連携や独自開発による学習データを用い、小型化と高性能の両立を志向し、ラックマウント型から小型 PC サーバーまでを視野に入れたラインアップを構築していると説明されている。これは、単なるベースモデルの日本語微調整ではなく、**文書 AI とオンプレ配備を一体化した製品戦略**である。 5

以下の図は、公式資料に基づいて整理した開発・提供の流れである。



上図のうち、強化学習、カリキュラム学習、VQA 選定、画像トークン圧縮、モデルマージはリコーが GENIAC 第3期の説明で明示した技術要素であり、現行 27B / 9B のニュースリリースでも「従来から取り組んできた強化学習およびカリキュラム学習を高度化し適用した」と説明している。 6

モデル仕様と技術構成

公開情報から確認できる事実を厳密に整理すると、**ベースモデルの詳細は比較的明確だが、Ricoh 版に固有の変更点は限定的にしか公開されていない**。27B ベースの Qwen3.6-27B は「Vision Encoder 付きの causal language model」で、27B パラメータ、64 層、hidden dimension 5120、Gated DeltaNet と Gated Attention を組み合わせたハイブリッド構成を持つ。9B ベースの Qwen3.5-9B も同じく Vision Encoder 付きで、9B パラメータ、32 層、hidden dimension 4096 である。いずれもベースモデルとしては 262,144 トークンのネイティブ文脈長を持ち、1,010,000 トークンまで拡張可能とされるが、**Ricoh 版がその上限を維持するかは未公開**である。 7

学習データについて Ricoh が公式に述べている範囲は、公開データ、研究機関連携データ、独自開発データ、内部データを用いていること、そして**顧客情報を学習に一切使っていない**ことである。さらに、データ品質向上のため理化学研究所 AIP の日本語インストラクションデータ作成プロジェクトにも参画している。だが、27B / 9B 各モデルごとの**データ量、トークン数、画像枚数、学習期間、学習計算資源は未公開**である。なお、参考情報として、前世代の Ricoh LMM では日本語図表に特化した合成データ約600万枚を学習した実績が公表されている。 4

マルチモーダル対応について、Ricoh の 2026年6月5日リリースは LMM を「テキスト・画像・音声・動画など複数データを扱える AI 技術」と一般定義しているが、当該 27B / 9B の個別説明で主に強調されるのは **図表・文書・画像・テキスト**である。他方、ベース Qwen3.6-27B と Qwen3.5-9B の Hugging Face モデルカードには画像入力と動画入力の API 例が示されている。したがって、**画像・図表・テキスト対応は公式確認、動画はベース由来では確認、音声の Ricoh 版正式サポート範囲は未公開**と記すのが最も保守的である。 8

軽量化については、27B モデルで **FP16、8bit、4bit** 量子化版が公式に明示されたことが重要である。しかもリコーは、量子化版でもベースモデルを上回るリーズニング性能と LLM 性能を維持すると説明している。いっぽう 9B については「より少ない GPU リソースでの運用を想定」とあるが、FP16/8bit/4bit の個別展開は公式に書かれていないため、**9B の量子化 SKU は未公開**とするのが厳密である。加えて、Ricoh は第3期の技術説明で、画像トークン圧縮によりトークン数を半減しつつ精度低下を 5% 未満に抑える技術、ならびにモ

デルマージ技術を開発したと説明している。現行 27B/9B にそれがどこまで直接適用されたかは未公開だが、**軽量化思想の技術的背景**としては重要である。 [6]

モデル仕様比較表

項目	Qwen3.6-Ricoh-27B-20260522	Qwen3.5-Ricoh-9B-20260522
位置づけ	Ricoh LMM_27B。Qwen3.6 ベースの推論強化モデル	Ricoh LMM_9B。Qwen3.5 ベースのコンパクト後継モデル
パラメータ数	27B	9B
ベースアーキテクチャ	Vision Encoder 付き causal LM、64層、hidden 5120、Gated DeltaNet / Gated Attention ハイブリッド	Vision Encoder 付き causal LM、32層、hidden 4096、Gated DeltaNet / Gated Attention ハイブリッド
日本語強化手法	Ricoh 独自の強化学習 + カリキュラム学習	同左
学習データ概要	公開データ + 研究機関連携/独自開発データ + 内部データ。顧客情報不使用	同左
マルチモーダル	図表・文書・画像・テキストは公式確認。動画はベースモデル由来では確認。音声は未公開	同左
コンテキスト長	未公開。ベースモデルは 262,144 native / 1,010,000 まで拡張	未公開。ベースモデルは同値
量子化	FP16 / 8bit / 4bit を公式公表	未公開
軽量化関連技術	画像トークン圧縮・モデルマージの技術基盤あり。現行モデルへの適用詳細は未公開	同左
推論速度	未公開	未公開
理論的重みサイズ	**推定**: BF16 約54GB / INT8 約27GB / INT4 約13.5GB	**推定**: BF16 約18GB / INT8 約9GB / INT4 約4.5GB
実運用要件	未公開。ベース Qwen3.6-27B のフル 262K 例は TP=8 を前提	未公開。ベース Qwen3.5-9B のフル 262K 例は TP=1 を前提
提供形態	2026年6月下旬からスターターキット提供予定	同左

出典: リコー公式リリース 2026-06-05、リコー技術ページ、Qwen 公式 Hugging Face モデルカード。理論的重みサイズは、公開パラメータ数に対する単純計算であり、KV キャッシュやビジョン処理の追加メモリを含まない。 [9]

性能評価と比較

性能評価で最も重要なのは、**どのスコアが、どの条件で得られたか**である。Ricoh 公式によれば、VLM ベンチマークの JDocQA-Reasoning / JDocQA は vLLM 0.19.0、LLM ベンチマークの ELYZA-tasks-100 / Japanese MT-Bench は vLLM 0.20.1 で推論し、評価は Azure OpenAI Service の gpt-4.1、ただし JDocQA のみ gpt-4o

を用いた LLM-as-a-Judge 方式で、各ベンチマークを 5 回実施した平均値である。さらに Gemin 2.5 Pro / Gemini 3 Pro Preview の値は 2026年3月30日リリース時点の参考値であり、**同日同条件の再走査ではない**。この注記は、スコアを読む際の最重要留保である。 [2]

公式ベンチマーク比較表

モデル	JDocQA-Reasoning	JDocQA	ELYZA-tasks-100	Japanese MT-Bench	備考
Qwen3.6-27B（ベース）	0.858	4.15	4.58	9.35	比較基準
Qwen3.6-27B-FP8	0.856	4.13	4.56	9.34	ベース量子化
Qwen3.6-Ricoh-27B-20260522	**0.881**	**4.22**	**4.64**	**9.48**	公式主力モデル
Qwen3.6-Ricoh-27B-20260522-AWQ-W8A16	0.873	4.21	4.65	9.47	量子化版
Qwen3.6-Ricoh-27B-20260522-AWQ-W4A16	0.868	4.20	4.62	9.35	量子化版
Qwen3.5-9B（ベース）	0.762	3.89	3.76	7.65	小型基準
Qwen3.5-Ricoh-9B-20260522	**0.782**	**4.00**	**3.95**	**7.93**	軽量モデル
Gemini 3 Pro Preview	0.880	4.24	—	—	前作時点参考値
Gemini 2.5 Pro	0.838	4.08	—	—	前作時点参考値
GPT-5.2	0.731	3.93	—	—	前作時点参考値
前作 Qwen3-VL-Ricoh-32B-20260227	0.826	4.08	—	—	27B の前世代比較
前作 Qwen-3-VL-Ricoh-8B-20260227	0.718	4.00	—	—	9B の前世代比較

出典: リコー公式リリース 2026-06-05。Gemini / GPT-5.2 / 前作値は同リリース内の注記どおり参考値。 [2]

上表から読み取れる核心は三つある。第一に、27B リコー版はベース 27B に対し、JDocQA-Reasoning で **+0.023**、JDocQA で **+0.07**、ELYZA-tasks-100 で **+0.06**、Japanese MT-Bench で **+0.13** 改善している。第二に、9B リコー版も、JDocQA-Reasoning で **+0.020**、JDocQA で **+0.11**、ELYZA-tasks-100 で **+0.19**、Japanese MT-Bench で **+0.28** 改善しており、**小型モデルでも日本語推論強化の効果は大きい**。第三に、27B の 8bit / 4bit 量子化版は精度劣化が限定的で、少なくとも Ricoh の自社ベンチでは「軽量化と性能維持の両立」がかなり意識されている。 [2]

ただし、これをもって「GPT-4o / GPT-4 / Claude / Llama 3 を厳密に上回る」と読むのは早計である。Ricoh が同一ハーネス・同日再評価で公開している相手は、JDocQA 系では主として Gemini 2.5 Pro / Gemini 3 Pro Preview / GPT-5.2 であり、**ユーザーが比較対象として挙げた GPT-4o / GPT-4 / Claude / Meta Llama 3 に対する同一条件の公開比較は見当たらない**。そのため、以下の比較はあくまで**別ハーネス・別時点・別 judge を含む参考比較**である。 [2]

他社モデルとの参考比較

比較対象	公開された日本語系指標	数値	出典・条件	厳密比較可否
OpenAI GPT-4o	Japanese MT-Bench	8.560	METI / GENIAC 第2期結果詳細	不可
OpenAI GPT-4-0613	Japanese MT-Bench	7.563	METI / GENIAC 第2期結果詳細	不可
Anthropic Claude 3.5 Sonnet	Japanese MT-Bench	8.635	METI / GENIAC 第2期結果詳細	不可
Meta Llama 3系 参考	Llama 3.3 Swallow 70B の Japanese MT-Bench	7.72	Tokyo Tech Swallow 公式	不可
Meta Llama 3系 参考	Llama-3-ELYZA-JP-70B は GPT-4 / Claude 3 Sonnet 超えと公表	定量表は別出典	ELYZA 公式	不可

出典: METI / GENIAC の性能評価結果詳細、Tokyo Tech Swallow 公式、ELYZA 公式。評価条件が一致しないため、この表は**意思決定の補助材料**にとどまり、公式横並び比較表としては扱えない。 [10]

第三者ベンチマークや独立レビューの観点では、2026年6月11日時点で本調査が確認できた主要公開記事は、AI Watch や note の週次解説を含めて **公式発表の要約・紹介が中心**であり、独立した再計測や実機レイテンシ評価、長時間運用レビューは確認できなかった。他方で、Ricoh は JDocQA-Reasoning ベンチマーク自体を Hugging Face で公開しており、今後は第三者再評価が可能な環境は整いつつある。現段階の評価の重みは、依然として Ricoh 自己申告が大きい。 [11]

導入運用セキュリティコスト

Ricoh の導入戦略は、モデル単体よりも **「RICOH オンプレLLMスターターキット」**というパッケージで理解したほうが正確である。製品ページによれば、このスターターキットは GPU サーバー、Ricoh 製 LLM、必要ソフトウェア、構築支援、運用支援、教育支援を一体で提供するローカル LLM パッケージであり、Dify も同梱される。公式には「GPU サーバ1台で稼働するお手軽な設備」とされ、利用規模に応じてエッジ、エントリー、スタンダード、ハイエンドのラインアップがある。 [12]

オンプレ導入の利点として Ricoh が繰り返し強調するのは、**機密データを外部に出さない構成、従量課金ではなく社内インフラとして使い放題に近い運用、利用部門がアプリを自作できる自由度**である。これは、製

造業・金融・行政で図表を含む社内文書を扱うケースでは説得力が強い。実際、Daifuku への導入事例では「機密性の高い顧客情報を扱うためセキュリティやガバナンスに課題があった」ことがオンプレ採用理由として明示されている。[13]

他方、制約も明確である。オンプレ LLM はクラウド従量課金より予測可能な面がある一方、**初期投資が先に立つ**。さらに、27B 級はベース Qwen3.6-27B のフル 262K 文脈例で tensor parallel 8 を前提としており、長い文脈をそのまま維持する場合には相応の GPU リソースが必要になる。逆に 9B はベースモデル側が TP=1 での提供例を持ち、Ricoh 技術ページでも「汎用 GPU サーバー1台構成でも活用できる」とされるため、**9B のほうが明らかに導入障壁が低い**。また Qwen 側は、メモリ不足時には文脈長を下げるよう案内しており、長文脈と運用コストのトレードオフは無視できない。[14]

推論速度そのものは Ricoh 版について未公開である。ただし、近傍サイズの Qwen3 ベンチマークでは、H20 96GB・SGLang 条件で Qwen3-32B BF16 が 20.72 tokens/s、AWQ-INT4 が 47.67 tokens/s、Qwen3-8B BF16 が 81.73 tokens/s、AWQ-INT4 が 144.11 tokens/s と報告されている。**これは Ricoh 27B / 9B そのものの速度ではない**が、「27B は高精度寄り、9B はスループット寄り」という大まかな運用感をつかむ参考にはなる。Ricoh 版の実測値は未公開として扱うべきである。[15]

コストについて価格表は公開されていない。したがって、ここではユーザー要望どおりレンジで表す。**これは公式価格ではなく、本報告書の推定**である。

モデル	初期費レンジ	運用費レンジ	判断根拠
Qwen3.6-Ricoh-27B-20260522	高	中	27B 級、フル文脈では高い GPU 要件、ただし 8bit/4bit 量子化あり
Qwen3.5-Ricoh-9B-20260522	中	低〜中	汎用 GPU サーバー1台構成を想定、より広い環境に導入可能

この推定レンジは、Ricoh の 27B/9B の位置づけ、量子化有無、スターターキット構成、ベース Qwen の運用例を総合したものにすぎず、正式見積りの代替にはならない。[16]

パートナー面では、Ricoh 技術ページに **エフサステクノロジーズの「Private AI Platform on PRIMERGY」**と **伊藤忠テクノソリューションズの NVIDIA DGX Spark OEM モデル**が明記されている。加えて CTC・Ricoh・Ricoh Japan は、27B 級 Ricoh LLM を搭載した「超小型デスクサイド AI 用サーバー」を2026年3月に提供開始している。したがって、Ricoh のオンプレ戦略は、**自社一社完結ではなく、サーバーベンダーと SI パートナーを伴う実装エコシステム**として理解すべきである。[17]

世間の評判と公開言説

2026年6月11日時点での世間の反応は、**好意的だが、まだ初期段階**という評価が妥当である。公開記事で最も目立つ評価軸は、「オンプレで動く」「日本語の図表・文書推論に強い」「Gemini 級に迫る」という三点であり、逆に言えば、実機検証や価格・SLA・長期運用レビューまで踏み込んだ議論はまだ少ない。AI Watch は見出しで「軽量・オンプレミス環境での運用に対応、商用クラウドAIに迫る性能」と整理し、note の週次 AI 解説記事はこれを「Sovereign AI の具体例」と位置づけた。[18]

この温度感は、はてなブックマークでも確認できる。6月11日時点で当該リリースには 7 users の反応がついていたが、爆発的な話題化というよりは、**AI 実務層が着実に注目し始めた水準**と読むのが自然である。[19]

代表的な公開言説を短く引けば、次の三つに要約できる。

「軽量・オンプレミス環境での運用に対応、商用クラウドAIに迫る性能を実現」—— AI Watch, 2026-06-05. [20]

「Sovereign AI の具体例」—— HIROE, note, 2026-06-06. [21]

「Gemini 2.5 ProおよびGemini 3 Pro Previewのスコアは…参考値として記載」—— リコー公式注記, 2026-06-05. [2]

肯定的評価の要約

- **日本企業の現実に合う**。社内文書、図表、PDF、会議資料など、実務で多い非構造データに焦点が定まっており、汎用チャットより刺さるという受け止めが強い。 [22]
- **オンプレ適性が高く評価されている**。とくに製造、自治体、金融など「データを外に出しづらい」部門にとって、クラウド依存を減らせる点が支持されている。 [23]
- **27B と 9B の二段構えが現実的**。27B を主力、9B を導入障壁の低い選択肢として用意したことにより、「PoC から本番までの梯子」がわかりやすい。 [1]
- **量子化まで公式に見せた点は好印象**。多くの企業向け発表は“精度だけ”で終わるが、Ricoh は 8bit / 4bit まで示し、運用現場を意識している。 [2]

否定的評価の要約

- **評価の中心は Ricoh 自社ベンチと自社運用**であり、第三者の同一条件検証がまだ不足している。最も大きな慎重論はここに集中する。 [24]
- **他社比較の“見え方”に注意が必要**。Gemini の数値は前作時点の参考値であり、同日の apples-to-apples 比較ではない。 [2]
- **価格と正確なハードウェア構成が未開示**である。導入検討の実務に直結するため、CIO / 情シス視点ではまだ情報不足が残る。 [25]
- **公開ライセンスと安全性文書の現行版が未公開**である。前作 8B の公開条件は確認できるが、今回の 27B / 9B の配布条件はまだ見えない。 [26]

総じて、公の評価は「かなり有望だが、まだ Ricoh の主張を検証しきる段階にはない」というものである。これは否定というより、**企業向け調達で当然求められる厳密性**がまだ追いついていない、という意味である。 [24]

法的・倫理的観点

ライセンス面では、**現行 27B / 9B の公開ライセンスは未公開**である。現時点の提供形態はスターターキット経由の商用提供予定であり、Hugging Face 上での一般公開は技術ページでも「予定」とされていない。他方で、ベースとなる Qwen3.6-27B と Qwen3.5-9B は、ともに Hugging Face 上で Apache-2.0 ライセンスが明示されている。したがって、ベースモデルはオープンだが、**Ricoh 派生モデルそのものの配布条件はまだ見えない**。 [27]

ただし、Ricoh の前作公開モデル「Qwen-3-VL-Ricoh-8B-20260227」は法的運用の前例として重要である。このモデルは Apache-2.0 を採用しつつ、追加利用規約で、入力データの適法性確認、第三者権利侵害の回避、出力の正確性・合法性・安全性の自己確認、そして医療・法律・税務・会計・金融・人事・採用・与信・公共サービスなど高リスク分野では**唯一または主要な判断根拠として自動使用してはならず、人的確認を要する**と明記している。したがって、Ricoh は少なくとも公開モデルにおいて、**高リスク用途の人間関与をライセンス文書に埋め込む方針**をとっている。現行 27B / 9B で同条件が採用されるかは未公開だが、前例として重い。 [28]

データ利用とプライバシー面では、Ricoh は技術ページで、技術倫理とデータガバナンスポリシーを遵守し、顧客情報を学習に一切使っていないと明記している。さらに、グループのデータガバナンスポリシーでは、暗号化、権限管理、脆弱性管理などにより機密性・完全性・可用性を確保し、個人情報保護を重視するとしている。オンプレミス提供とあわせて、**データ主権とプライバシー保護を強く打ち出す設計**である。 29

バイアス・安全性対策については、Ricoh グループ技術倫理が、AI が差別・偏見・格差を助長するリスクや人権侵害リスクを明示的に認め、これを抑制するための技術倫理憲章を定めている。また、Ricoh は別建てでセーフガードモデルを開発しており、入力・出力に対して暴力、犯罪、差別、プライバシー侵害など14種類のラベルで有害性判定を行うとしている。ただし、**そのセーフガードが今回の 27B / 9B にどのように統合されるかは未公開**であり、現行2モデルの個別セーフティ評価スコアも開示されていない。 30

規制対応の観点では、Ricoh の公開文書から EU AI Act や日本の AI 事業者ガイドラインへの個別適合表が確認できたわけではない。現時点で確認できるのは、データガバナンス、技術倫理、情報セキュリティ体制を会社として整備していること、そして前作公開モデルの追加規約で高リスク用途の人的確認を要求していることである。したがって、**制度適合の姿勢は見えるが、モデル単位のコンプライアンス資料としてはまだ不足**している、というのが妥当な評価である。 31

結論と推奨

結論を先に述べる。**Qwen3.6-Ricoh-27B-20260522 は、2026年6月時点の日本市場において、オンプレミス条件下で日本語の図表・文書推論を重視する企業にとって有力な選択肢である。Qwen3.5-Ricoh-9B-20260522 は、その導入障壁を大きく下げる実務向け小型モデルである。** ただし、現段階での強みは「Ricoh が定めた評価軸で強い」ことであり、同時に限界は「第三者が同じ条件でまだ十分に検証していない」ことにある。 32

企業が導入を判断する際の基準は、次のように整理できる。

判断基準	27B を優先すべきケース	9B を優先すべきケース	クラウド AI 併用を優先すべきケース
主業務	図表付き文書の高精度読解、仕様確認、複雑な照会応答	FAQ、要約、抽出、軽量な社内検索補助	広範囲な一般知識、最新性、外部接続前提の業務
セキュリティ要件	機密データを外部送信できない	同左	外部 API 利用が許容される
ハードウェア制約	中〜高い GPU 投資が可能	1台構成など制約が強い	自前 GPU を持ちたくない
成熟度要件	自社 PoC と評価設計を回せる	小さく始めて段階拡張したい	既にクラウド運用と MLOps が整っている

上表の実務的含意は明快である。**製造、金融・保険、公共・自治体、技術文書審査**のように、図表・帳票・長文 PDF を社内閉域で扱うケースなら、27B は強く検討に値する。逆に、部門 PoC、検索補助、定型抽出、閉域チャットボットから始めるなら 9B のほうが現実的である。音声や広い汎用マルチモーダル、外部情報との連携、あるいは第三者ベンチで確立済みの最先端性能が最優先なら、OpenAI / Anthropic / Google 系クラウド AI を併用または主軸とするほうが安全である。これは公式仕様の再記述ではなく、上記公開情報に基づく**推奨判断**である。 33

リスク緩和策としては、導入前に最低でも次を要求すべきである。第一に、**自社データでの再評価**。Ricoh 公式ベンチに加えて、自社の PDF、図表、表計算、議事録、社内規程を用いた正答率・再現率・根拠提示率を

測るべきである。第二に、**量子化別の比較**。27B は 8bit / 4bit でも公開ベンチの低下が小さいため、実運用は量子化版を中心に設計し、FP16 は検証用にとどめるのが合理的である。第三に、**人間確認の組み込み**。前作ライセンス前例と Ricoh の技術倫理方針を踏まえ、医療・法務・採用・与信・品質保証などでは生成結果を自動確定しないワークフローが最低条件である。第四に、**クラウド併用の逃げ道**を残すこと。独立評価がまだ薄い以上、難問や境界事例だけを frontier モデルにエスカレーションする二層構成は合理的である。 [34]

最終的な推奨は次の一文に尽きる。**社内閉域・日本語文書・図表推論が中核要件なら、Ricoh 27B/9B は「真面目に PoC すべきモデル群」である。しかし、現時点ではまだ「即断で全社標準に決めるべき完成済み製品」とまでは言い切れない。** 企業導入の最善手は、27B を本命、9B をスモールスタート、クラウド AI を比較対照として並走させ、自社データで三者比較したうえで決めることである。 [3]

---

[1] [2] [3] [6] [8] [9] [16] [22] [24] [32] [33] [34] https://jp.ricoh.com/release/2026/0605_1
https://jp.ricoh.com/release/2026/0605_1

[4] [5] [17] [29] “はたらく”を支えるリコーの大規模言語モデル（LLM） | リコーグループ 企業・IR | リコー
https://jp.ricoh.com/technology/ai/LLM

[7] [14] [27] https://huggingface.co/Qwen/Qwen3.6-27B
https://huggingface.co/Qwen/Qwen3.6-27B

[10] https://www.meti.go.jp/policy/mono_info_service/geniac/selection_2/result_2/result_details_2/index.html
https://www.meti.go.jp/policy/mono_info_service/geniac/selection_2/result_2/result_details_2/index.html

[11] [18] [20] https://ai.watch.impress.co.jp/docs/news/2114771.html
https://ai.watch.impress.co.jp/docs/news/2114771.html

[12] [13] [23] [25] https://promo.digital.ricoh.com/ai/service/ricoh-on-premises-llm-starter-kit/
https://promo.digital.ricoh.com/ai/service/ricoh-on-premises-llm-starter-kit/

[15] https://qwen.readthedocs.io/en/latest/getting_started/speed_benchmark.html
https://qwen.readthedocs.io/en/latest/getting_started/speed_benchmark.html

[19] https://b.hatena.ne.jp/entry/s/jp.ricoh.com/release/2026/0611_1
https://b.hatena.ne.jp/entry/s/jp.ricoh.com/release/2026/0611_1

[21] https://note.com/hiroe28/n/n611d35ffeb95?hl=en
https://note.com/hiroe28/n/n611d35ffeb95?hl=en

[26] [28] https://huggingface.co/ricoh-ai/Qwen-3-VL-Ricoh-8B-20260227
https://huggingface.co/ricoh-ai/Qwen-3-VL-Ricoh-8B-20260227

[30] https://jp.ricoh.com/technology/rd/technology_ethics
https://jp.ricoh.com/technology/rd/technology_ethics

[31] https://jp.ricoh.com/technology/rd/data_governance
https://jp.ricoh.com/technology/rd/data_governance