# 戦略的技術評価レポート：リコー製マルチモーダルLLM

## 「Qwen3.6-Ricoh-27B」の図表読解性能とエンタープライズ実装への適合性

| |
|---|
| 評価対象：Qwen3.6-Ricoh-27B-20260522 / Qwen3.5-Ricoh-9B-20260522 |
| 発表日：2026年6月5日 |
| レポート発行：Claude Fable 5 (2026年6月11日) |

# エグゼクティブ・サマリー (The Bottom Line)

### 独自ベンチマークで Gemini 3 Pro級の 図表読解

**JDocQA-Reasoning スコア: 0.881**

製造業の設計図や金融約款など、複雑な日本語図表の多段推論(リーズニング)に特化。

### オンプレミスでの 完全自律稼働

**VRAM 16.8GB (4bit量子化版)**

中堅企業でも導入可能なハードウェア要件。機密データを社外ラウドに出さずに高度なAI処理を実現。

### 汎用性能の勝利ではなく「特化型」の成果

**LLM-as-a-Judge / 自社測定**

第三者検証は未了。Qwenベースであるためのライセンスやデータバイアスのデューデリジェンスが必須。

# 発表モデルの全体像とスペック

モデル　Flagship

**Qwen3.6-Ricoh-27B-20260522**

ベース: Qwen3.6-27B（27.8Bパラメータ）

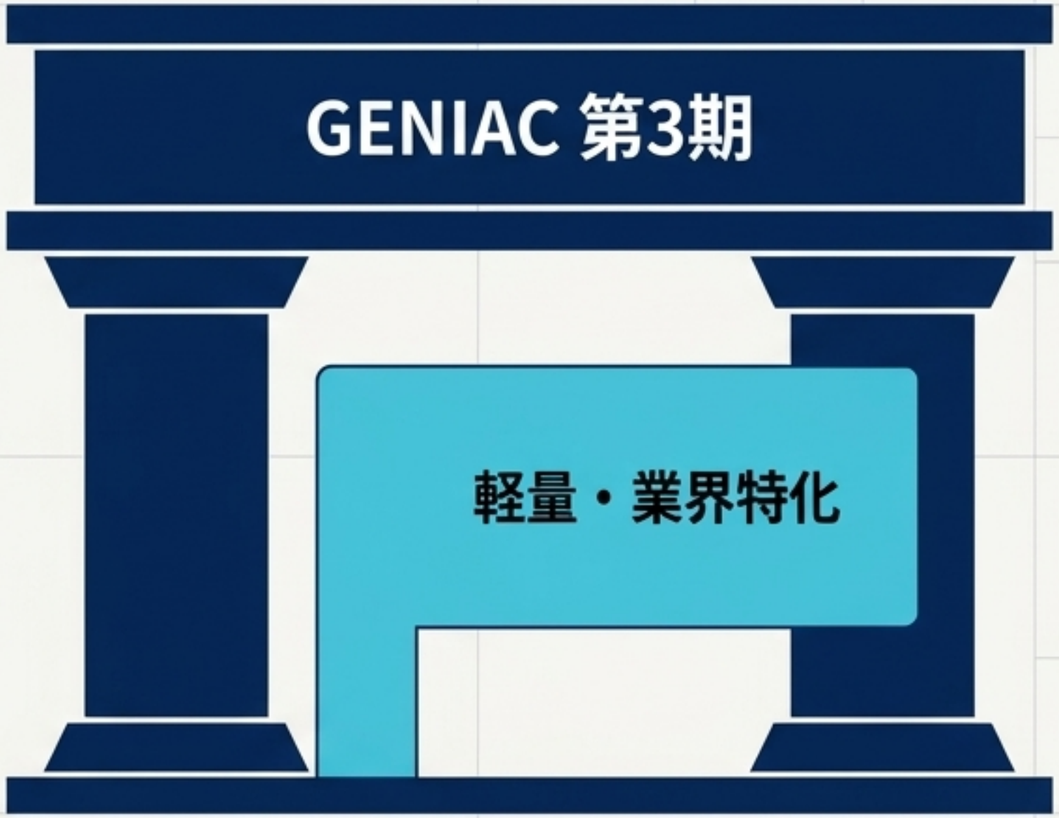
特徴: 高度なマルチモーダル・多段推論。

モデル　Lightweight

**Qwen3.5-Ricoh-9B-20260522**

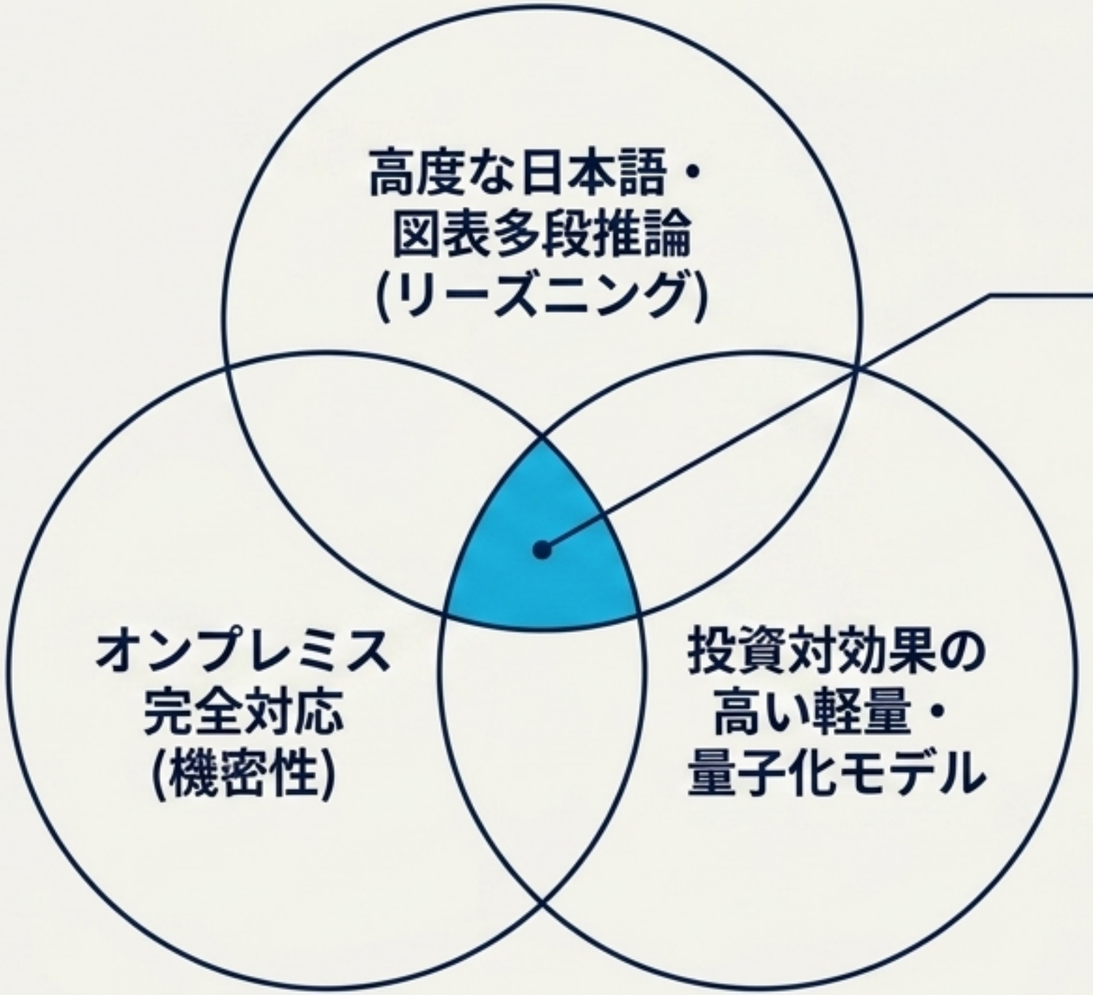
ベース: Qwen3.5-9B

特徴: リソース制約のある環境向け。




経済産業省・NEDO「GENIAC」第2期・第3期の成果。「軽量・業界特化」という国内AI開発の最新トレンドを体現。

# 独自の市場ポジショニング：3つの価値の交差点

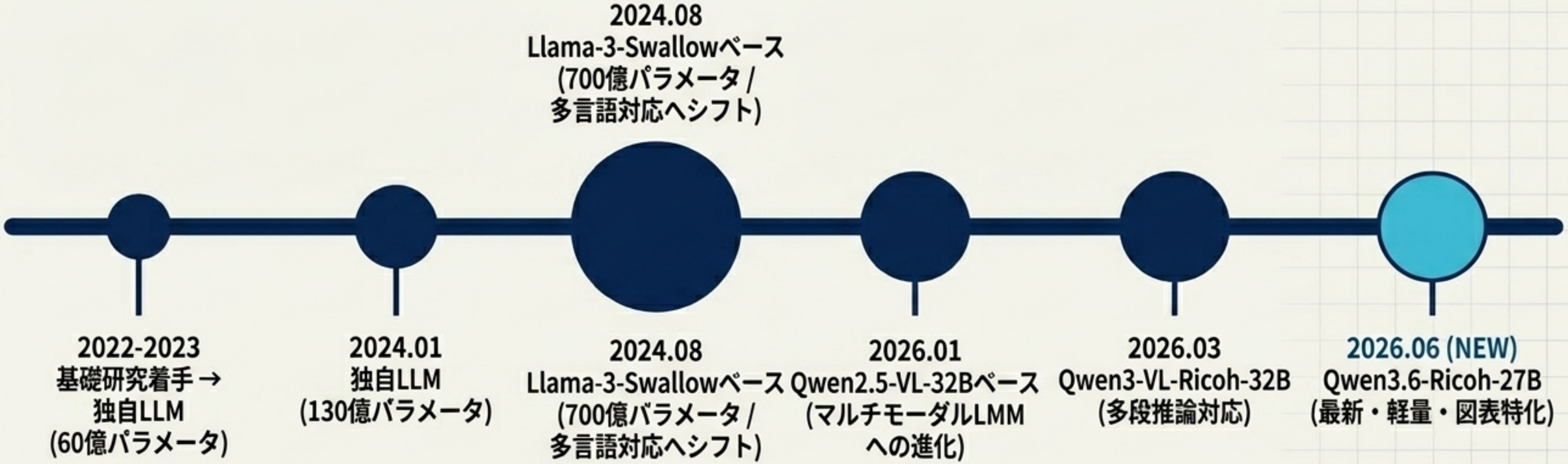



**機密性の高い
エンタープライズ
エンタント
プライズ文書の
ローカル解析**

- ✓ ・製造業(設計図)
- ✓ ・金融機関(約款)
- ✓ ・官公庁(行政文書)

単なる汎用LLMではなく、クラウドに送信できない「図表入りビジネス文書」の処理に特化。

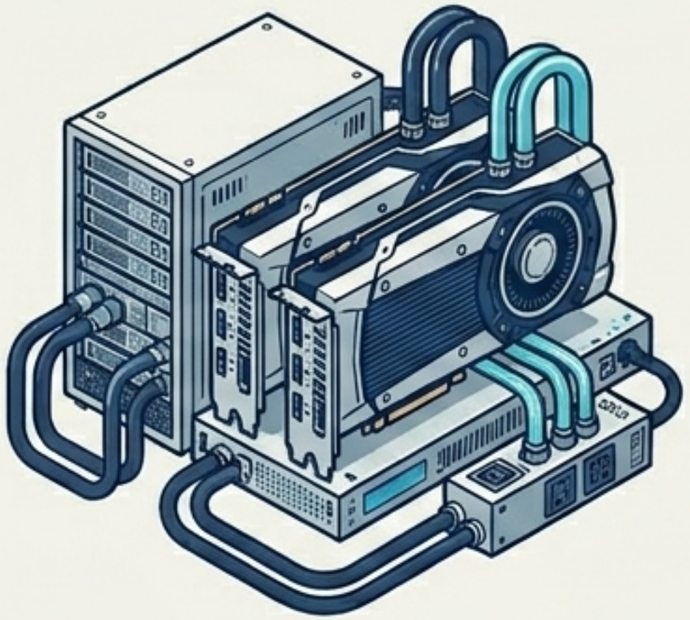
# リコーLLM開発の系譜：独自開発からQwenベースへの戦略的シフト




開発の歴史から見えるのは、フルスクラッチの固執を捨て、「最高峰の海外オープンモデル」＋「独自の日本語・図表学習」へとリソースを集中させた現実的かつアジャイルな戦略。

# 実装要件とハードウェア・コスト（推計）

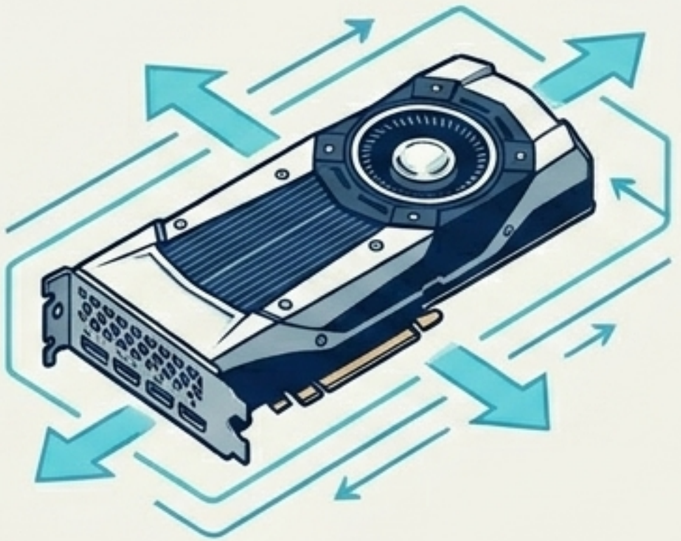
## The Heavy Configuration (FP8)



**H200 1基 または A100 2基**

ベースモデルの標準的な動作環境。

## The Lightweight Configuration (4bit AWQ-W4A16)



**VRAM 約16.8GB**
**(16GB級GPUで動作可能)**

量子化により大幅なコスト削減。中堅企業のオンプレミス環境でも現実的な選択肢に。

## RICOH オンプレ LLM スターターキット

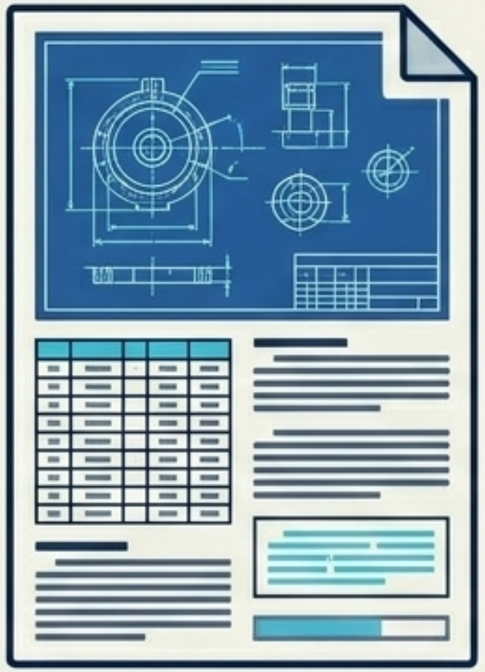
- GPUサーバ 1台
- リコー製LLM
- ノーコード開発基盤「Dify」

**提供開始予定: 2026年6月下旬 / 一部門(30名程度)想定**

参考：2025年モデル提供時は「1500万円から」

# マルチモーダル多段推論（リーズニング）の仕組みと価値

## Step 1: Input



**複雑な機密文書**
（製造設計図面、金融稟議書など）

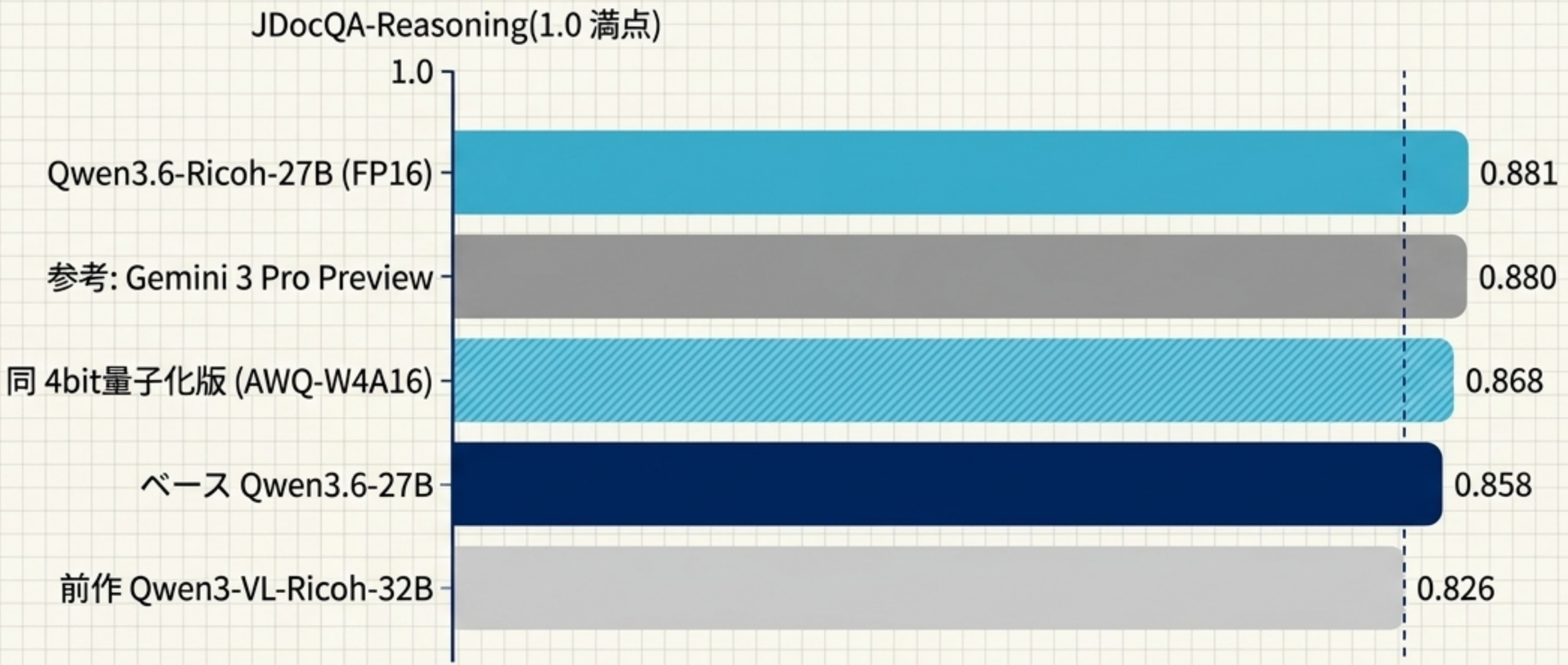
## Step 2: Processing (Qwen3.6-Ricoh-27B)

- 図表から数値を読み取る (Vision)
- テキストの要求仕様と照らし合わせる (Text Comprehension)
- 矛盾や適合性を論理的に判定する (Reasoning)

## Step 3: Output


**適合確認・要点抽出の完了**
（クラウドへのデータ送信ゼロ）

# 図表読解性能：独自ベンチマーク「JDocQA-Reasoning」スコア




**Key Takeaway Box**

最新クラウドモデル(Gemini)に匹敵するスコアを記録。
特に注目すべきは「4bit量子化版(0.868)での性能劣化が極めて小さい」点。ハードウェア制約の厳しいオンプレミス運用において強力な後押しとなる。

# 日本語テキスト処理性能の向上

## ELYZA-tasks-100(5 点満点)


**リコー版 (27B): 4.64**

## Japanese MT-Bench(10 点満点)


**リコー版 (27B): 9.48**

ベースモデルの時点で高い性能を持つQwen3.6に対し、リコー独自の強化学習 (報酬関数設計の精緻化) とカリキュラム学習を適用。過学習を抑えつつ、日本語の論理的推論能力を確実に底上げしている。

# 競合ランドスケープ：国産/オンプレミスLLM市場での立ち位置

| | リコー (Qwen3.6-Ricoh) | NTT (tsuzumi 2 Vision) |
|---|---|---|
| 開発アプローチ | 海外強力モデル(Qwen)の追加学習 | 純国産・フルスクラッチ開発 |
| パラメータ/構成 | 27B (マルチモーダル密モデル) | 30B (テキスト基盤＋図表アダプタ) |
| データ主権・リスク | ベースモデルの海外依存<br>(ライセンス変更リスク等) | 完全な国内主権<br>(リスク極小) |
| 動作環境 | 両者とも: 1GPU (A100/H200クラス)でのオンプレミス稼働を想定 | |

「海外最先端アーキテクチャの恩恵を享受するリコー」か、「完全な技術主権と安心を担保するNTT」か。エンタープライズの選択はこの二極化に向かっている。

# クリティカル・レビュー：ベンチマーク評価の留意点 (The Reality Check)

## 1 評価基盤の偏り

「JDocQA-Reasoning」はリコー自身が開発したベンチマークであり、自社測定値である。

## 2 LLM-as-a-Judgeの限界

採点役はAzure OpenAI（GPT-4.1 / GPT-4o）。AIによる評価であり、人間による厳密な定性評価ではない。

## 3 比較対象の時期ズレ

Gemini 3 Pro Previewのスコアは2026年3月時点の参考値。同一条件での第三者独立検証（Nejumi Leaderboard等）は未了。

**「Geminiを凌駕した汎用AI」ではなく、「特定の図表タスクにおいて極めて高いチューニングが施された特化型AI」として正しく評価する必要がある。**

# メディア・技術コミュニティの反応

## 一般メディア報道

概ねプレスリリースの要約報道（AI Watch, ZDNET Japanなど）にとどまる。 ZDNETは「従来のオープン系モデルを上回る性能」と評価。

> ベンチマークの自社測定や参考値比較は割り引いて見るべきだが、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は『誠実なアプローチ』である。

**データセット(1,362問)の公開は、自社の性能主張に対する強い自信の表れであり、エコシステム全体への貢献として高く評価されている。**

# 導入へ向けたデューデリジェンス・チェックリスト

### 自社データによるPoC

Hugging Faceの公開データだけでなく、自社の実文書（設計図・約款）を用いた推論精度の独自検証を実施する。

### TCO分岐点の算出

スターターキット価格と、クラウドAPI（GPT-4o/Gemini）を使い続けた場合の従量課金コストの損益分岐点を計算する。

### ライセンス動向の監視

Qwenのライセンスは現状Apache 2.0だが、将来的な方針変更リスク（技術コミュニティでの懸念）に備え、契約時の権利関係を明確化する。

### 出力バイアスの確認

中国製ベースモデルに起因する回答傾向やバイアスの残存がないか、出力テストを徹底する。

Strategic Conclusion

# リコーの「Qwen3.6-Ricoh-27B」は、汎用LLM競争から距離を置き、「製造・金融・公共における機密図表の処理」というエンタープライズの真のペインポイントに照準を合わせた極めて実用的なソリューションである。

4bit量子化によるオンプレミス実装の現実性と、商用クラウドに肉薄する特化性能。純国産モデル（NTT等）との比較検討を経た上で、自社のセキュリティ要件に合致 すれば 、強力なDX推進エンジンとなる。