Mission Control Dashboard

リコー製マルチモーダルLMM調査報告書

企業向けオンプレミスAIの実力と導入への羅針盤

Meta-data Panel

対象モデル: Qwen3.6-Ricoh-27B-20260522 / Qwen3.5-Ricoh-9B-20260522

レポート種別: アナリストグレード評価 & 導入バイヤーズガイド

基準日: 2026年6月

提供されるもの (The Offering)

コンセプト:
企業向けオンプレミス特化型LMM

コア技術:
Alibaba Cloudの「Qwen3.6-27B」「Qwen3.5-9B」をベースに日本語リーズニング性能を強化。

提供形態:
クラウドAPIではなく「RICOH オンプレLLMスターターキット」（2026年6月下旬予定）としてハード・ソフト一体提供。

最適なユースケース (The Target)

対象企業:
機密性の高い社内文書（図表・帳票・長文PDF）を扱う企業。

主要セクター:
製造業、金融・保険、公共・自治体。

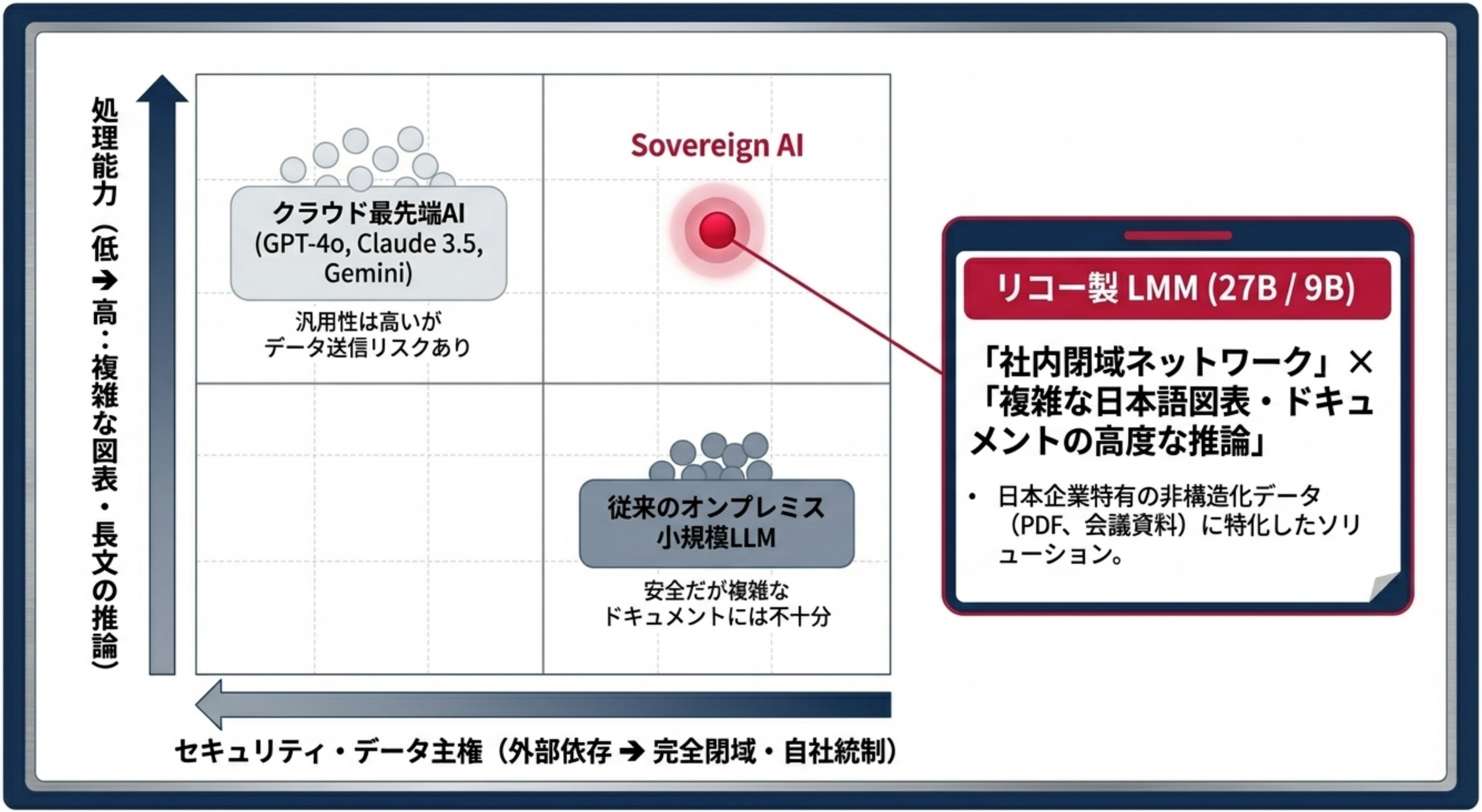
導入ドライバー:
外部にデータを出せない「閉域網要件」と「高度な日本語読解」の両立。

評価と結論 (The Verdict)

ステータス:
有力な候補だが、**PoCが必須**。

強み:
自社評価スコアは**Gemini 3 Pro Previewに迫る高水準**。

留意点:
第三者による同一条件でのベンチマーク、正確なBOM（部品表）、正式ライセンスが未公開。即断ではなく実データでの検証を推奨。


処理能力（低 → 高・複雑な図表・長文の推論）
クラウド最先端AI
(GPT-4o, Claude 3.5, Gemini)
汎用性は高いが
データ送信リスクあり
Sovereign AI
従来のオンプレミス
小規模LLM
安全だが複雑な
ドキュメントには不十分
リコー製 LMM (27B / 9B)
「社内閉域ネットワーク」×「複雑な日本語図表・ドキュメントの高度な推論」
• 日本企業特有の非構造化データ（PDF、会議資料）に特化したソリューション。
セキュリティ・データ主権（外部依存 → 完全閉域・自社統制）

Qwen3.6-Ricoh-27B-20260522

GENIAC第3期開発「LMM_32B」の後継。
高精度な推論特化型。

ベース: Qwen3.6-27B
（64層 / Hidden 5120）

ターゲット: 複雑な図表を含む
仕様書や帳票の正確な読み取り。

特長: 4bit/8bit量子化でも
4bit/8bit量子化でも性能劣化が
極めて少ない。

Qwen3.5-Ricoh-9B-20260522

LMM_8Bの後継。コンパクト・実装重視型。

ベース: Qwen3.5-9B
（32層 / Hidden 4096）

ターゲット: 導入障壁を下げた
実務向けPoC、
軽量な検索補助。

特長:「汎用GPUサーバー1台
「汎用GPUサーバー1台構成」での
稼働を想定した設計。

DEVELOPMENT PIPELINE: RICOH LLM & LMM

01
ベースモデル

Qwen3.6-27B /
Qwen3.5-9B

02 データ整備

リコー独自の
VQA選定
（顧客情報は一切不使用）

03 教師あり学習

04 強化学習

GENIAC第3期
からの高度化

05

カリキュラム学習

06 日本語リーズニング強化

07 軽量化

量子化
（画像トークン圧縮・
モデルマージ技術）

08 パッケージ化

RICOH
オンプレLLMスターター
キット

09 デプロイ

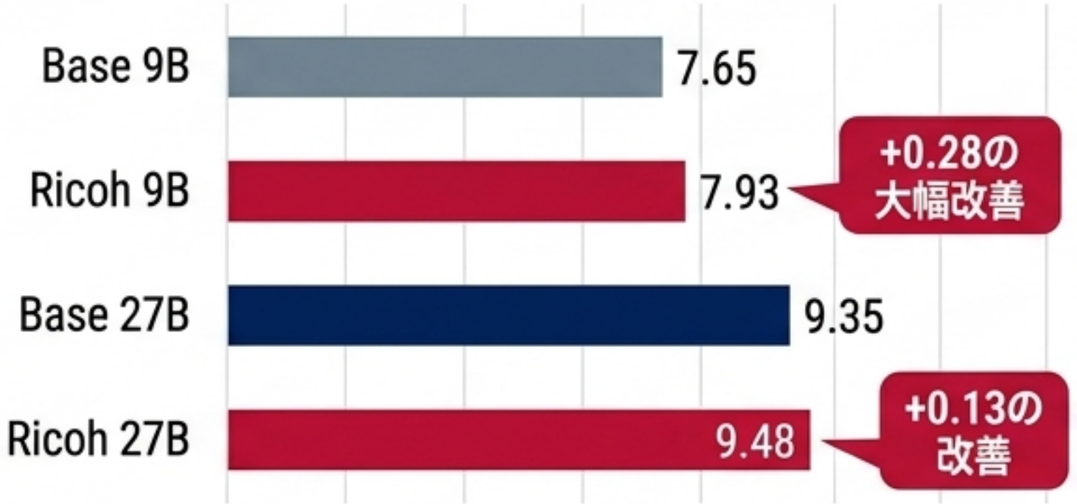
社内閉域での
業務利用

Mission Control Dashboard

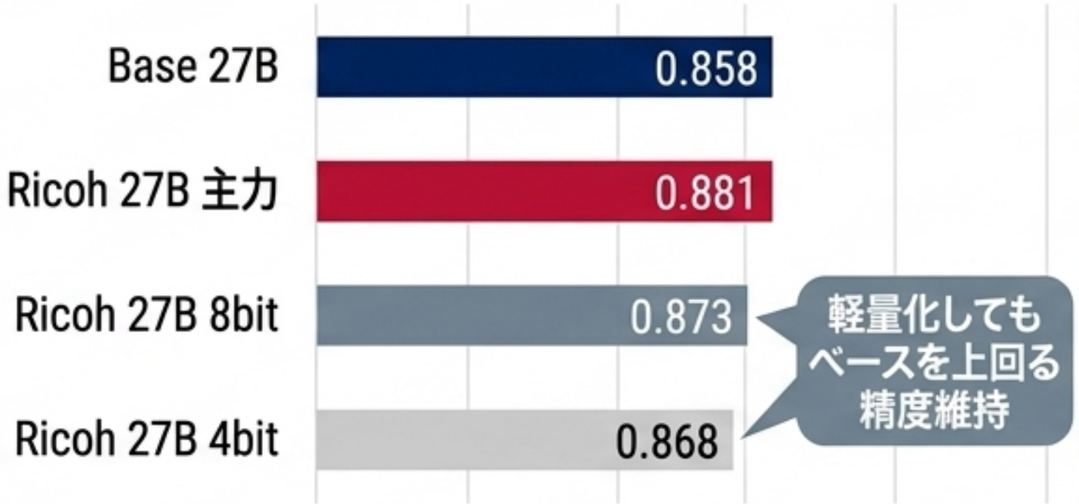
仕様項目	Qwen3.6-Ricoh-27B	Qwen3.5-Ricoh-9B
アーキテクチャ	Vision Encoder + causal LM (DeltaNet/Attention Hybrid)	Vision Encoder + causal LM (DeltaNet/Attention Hybrid)
コンテキスト長	ベースモデルは262,144トークン（最大1M拡張）。Ricoh版のRicoh版の上限は未公開。	ベースモデルは262,144トークン（最大1M拡張）。Ricoh版のRicoh版の上限は未公開。
理論的重みサイズ（推定）	BF16:約54GB / INT8:約27GB / INT4:約13.5GB	BF16:約18GB / INT8:約9GB / INT4:約4.5GB
量子化提供	FP16 / 8bit / 4bit を公式発表	未公開（汎用GPU1台構成を想定）
マルチモーダル	図表・文書・画像・テキストは公式確認。動画はベース由来。音声は未公開。	図表・文書・画像・テキストは公式確認。動画はベース由来。音声は未公開。

Mission Control Dashboard

ベースからの性能底上げ (Japanese MT-Bench)



量子化時の精度維持 (JDocQA-Reasoning)



リコーの独自学習（強化学習+カリキュラム学習）により、両モデルともベースラインから明確なリーズニング性能の底上げを達成。特に量子化版での性能維持がオンプレ運用において極めて有利。

Mission Control Dashboard

The Claim（公式発表）

“Gemini 3 Pro Previewに近い性能水準”

(JDocQA: 4.22 vs Gemini’s 4.24)

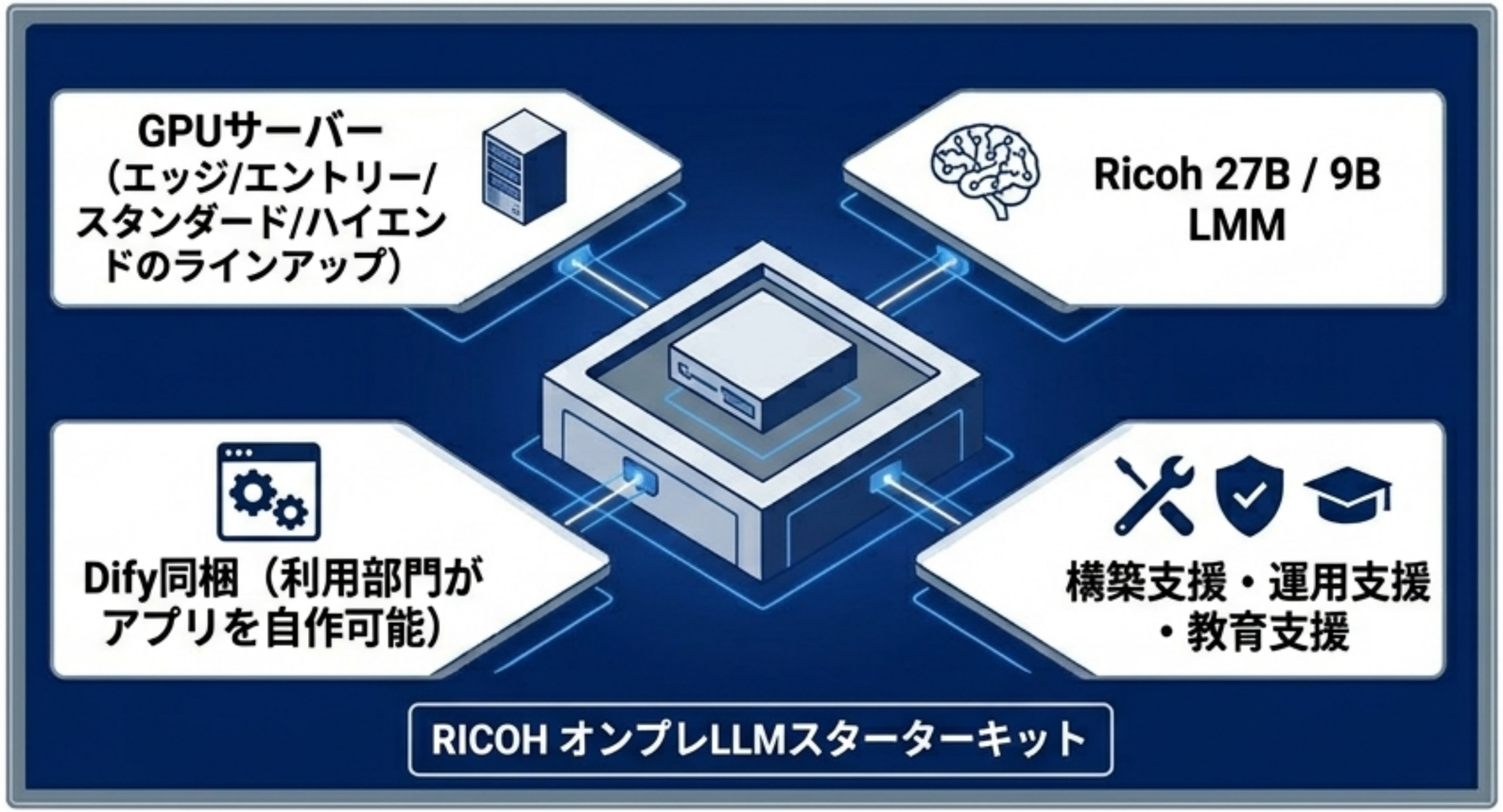
The Analyst Context（スコアの厳密な解釈）

1. **LLM-as-a-Judgeの偏り:** 評価はリコー自身がgpt-4.1 / gpt-4oを用いて実施した自社評価である。
2. **時点のズレ:** Gemini 2.5 Pro / 3 Pro Previewのスコアは「2026年3月30日時点の参考値」の再掲であり、同日・同一条件での厳密な再評価ではない。
3. **比較対象の不在:** GPT-4o, Claude 3.5 Sonnet, Llama 3など最新鋭モデル群との同一ハーネスでの公式横並び比較は未公開。

非常に有望な性能を示すが、第三者による独立したベンチマーク検証が待たれるフェーズ。「自社評価中心」であることを前提としたPoCが必須。

Mission Control Dashboard

RICOH オンプレLLMスターターキット: 製品構成とエコシステム



Ecosystem Partners

- エフサステクノロジーズ（Private AI Platform on PRIMERGY）
- 伊藤忠テクノソリューションズ / CTC（超小型デスクサイドAI用サーバーOEM）

リコーはAPIやモデル単体ではなく、「社内インフラとして「社内インフラとして使い放題」のローカル環境パッケージを販売している。

Mission Control Dashboard

Qwen3.6-Ricoh-27B
（高精度・高要件）

推論速度: 高精度寄り。

ハードウェア要件: ベースモデルのフル262K文脈処理には Tensor Parallel = 8 (TP=8) が前提。長文脈を維持するには**大規模なGPU投資が不可欠**。

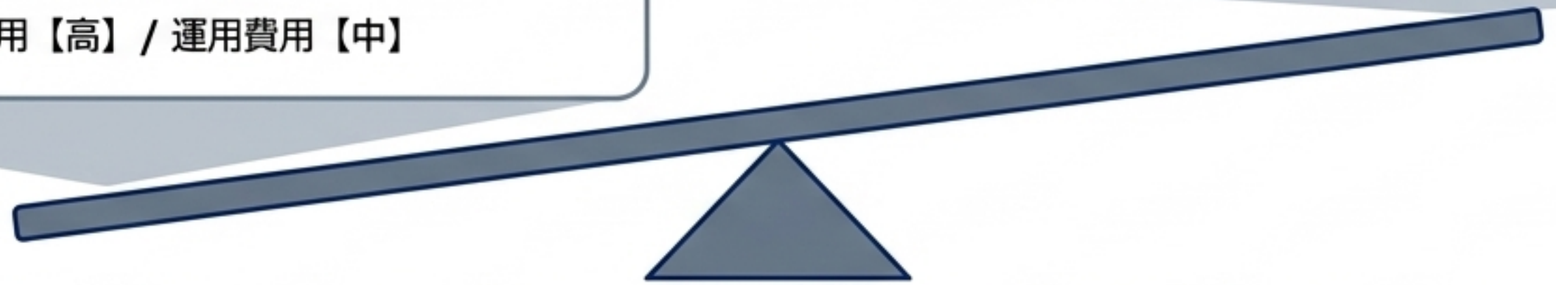
コスト推定: 初期費用【高】/ 運用費用【中】

Qwen3.5-Ricoh-9B
（スループット・高実装力）

推論速度: スループット寄り。

ハードウェア要件: Tensor Parallel = 1 (TP=1) 対応。汎用GPUサーバー1台構成での稼働を想定しており、導入障壁が極めて低い。

コスト推定: 初期費用【中】/ 運用費用【低〜中】



正式な価格・ハードウェアBOMは未公開。上記は仕様に基づく推計。長文脈処理とGPUコストのトレードオフ設計が運用の鍵を握る。

Governance & Security

Data Privacy
(データ主権)

事実: 顧客情報は学習データに一切不使用。

ポリシー: リコーグループのデータガバナンスに準拠。暗号化、権限管理による機密性・完全性の担保。

Technical Ethics
(技術倫理と安全性)

事実: 人権侵害やバイアスを抑制する技術倫理憲章。

セーフガード: 別途開発された14種類の有害性判定ラベルを持つセーフガードモデルが存在（※本作への組み込み詳細は未公開）。

Licensing Strategy
(ライセンスと法的要件)

ベース: QwenモデルはApache-2.0。Ricoh版の正式配布条件は未公開。

前例（重要）: 前作8Bモデルでは**「高リスク分野での単独自動判断の禁止・人的確認（Human-in-the-loop）の要求」**が利用規約に明記。本作でも同様の運用が予想される。

Market Sentiment Ledger

Positive Reception (期待と支持)

> “Sovereign AI の具体例” (note言説)

- Fit for Reality: 汎用チャットではなく、PDFや会議資料など**「日本企業特有の非構造化データ」**に焦点を絞った点が実務層から高評価。
- Pragmatic Lineup: 27Bと9Bの二段構え、及び4bit/8bit量子化の公式提示が、**「PoCから本番への梯子」**として現実的。

Cautious Reception (懸念と慎重論)

- Lack of Independence: 第三者の同一条件検証（独立ベンチマーク）が不足している点に慎重論が集中。
- **Opaque ROI**: 価格表と**正確な要求ハードウェア構成（BOM）**が**未開示**のため、情シス視点での初期投資（ROI）計算が困難。

Professional Analyst Report

The Ultimate Decision Matrix

主なユースケース / セキュリティ要件 / ハードウェア投資 / 成熟度

27B を優先すべきケース (The Precision Route)

- 図表付き仕様書・長文PDFの高度な推論と審査。
- 機密データを絶対に外部に出せない（完全閉域）。
- 中〜大規模なGPU投資（複数基構成）が許容できる。

9B を優先すべきケース (The Agile Route)

- 社内FAQ、要約、軽量な社内ドキュメント検索。
- 機密データを扱いつつ、小さくPoCから始めたい。
- 汎用GPU1台など、インフラ制約が強い環境。

クラウドAI併用を優先すべきケース (The Hybrid Route)

- 広範囲な一般知識、外部APIとの連携が前提。
- 外部へのデータ送信（セキュアなAPI経由）が許容可能。
- 自社でGPUインフラを持ちたくない（運用負荷ゼロ）。

Strategic Roadmap: Four-Step Implementation Path

Step 1: 実データ検証 (Own Data Test)

公式の汎用ベンチマークを鵜呑みにせず、自社の実際のPDF、図表、表計算データで正答率と根拠提示率を測る。

Step 2: 量子化モデルの優先評価 (Quantization First)

コスト削減のため、27BのFP16よりも、性能劣化の少ない「8bit / 4bit版」を実運用の中心に据えて初期設計を行う。

Step 3: 人間関与の組み込み (Human-in-the-Loop)

リコーの技術倫理と前作ライセンス方針に倣い、品質保証や法務・人事等の高リスク業務では自動確定させないワークフローを組む。

Step 4: クラウドの「逃げ道」確保 (Cloud Fallback)

独立評価が未成熟な現段階では、**境界事例**や**難問**のみを**クラウド上の最先端モデル（GPT-4o等）へエスカレーション**する**二層アーキテクチャ**を構築する。

社内閉域・日本語文書・図表推論が中核要件ならば、リコー製27B/9Bは「真面目にPoCを実施すべき最有力モデル群」である。

1　評価の留保

即断で全社標準にする完成済みパッケージではなく、まだ性能検証の余地を残す段階。

2　最適解の探求

27Bを本命、9Bをスモールスタート用とし、クラウドAIと比較対照させるアプローチが最も合理的。

3　今後の焦点

正式リリース（2026年6月下旬予定)における、ライセンス条件、第三者評価、そして具体的なハードウェア価格の開示を注視すること。