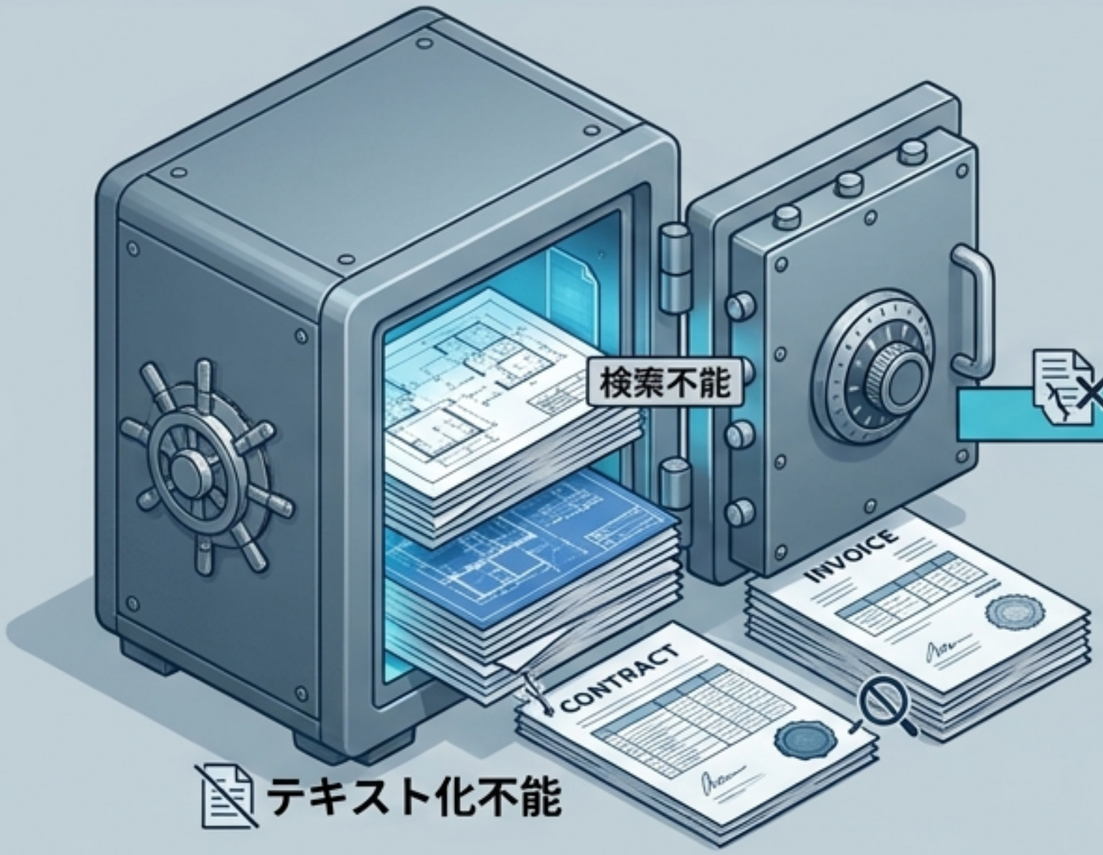
エンタープライズAIのパラダイムシフト

汎用クラウドAIの限界を超え、「知の金庫」で駆動する実務特化型オンプレミスAIの夜明け

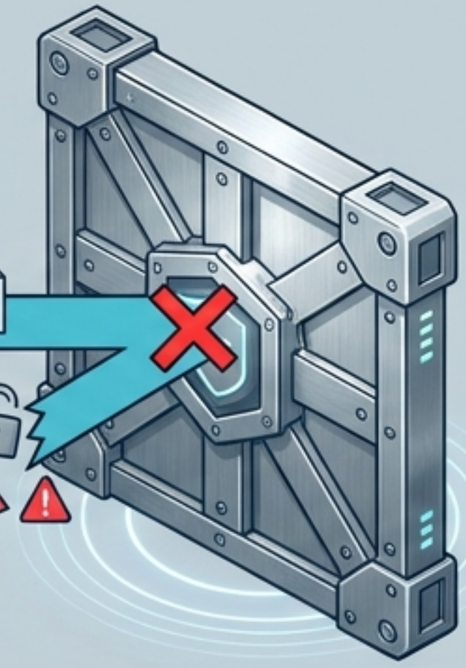
GENIACプロジェクトの成果がもたらす、日本企業のデータガバナンスと業務革新の両立

企業の「知の結晶」とクラウドAIの限界

企業の内部



外部クラウド



GeminiやGPT-4/5クラスの
圧倒的推論力

企業の最も価値あるデータは視覚的（図表・レイアウト）であり、機密性が高い。
強力なクラウドAIが存在しても、ガバナンスの壁により『分析不能なダークデータ』として死蔵されている。

リコー・エンタープライズ特化型LMMの誕生

完全オンプレミス稼働

外部通信ゼロ。機密データの社外流出リスクを根本から排除。

超高度な図表推論

日本語ドキュメントの複雑な図表、論理的推論においてクラウド巨大AIに肉薄。

現実的な導入コスト

軽量化技術により、超高額なサーバー専用GPUへの依存を脱却。

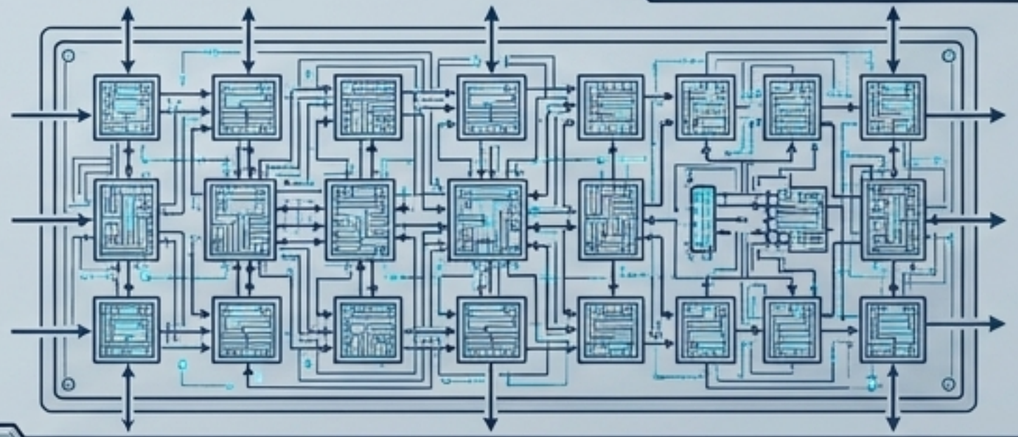
GENIAC
プロジェクト成果

戦略的ポジショニング：巨大モデルから特化モデルへのシフト

評価軸	汎用巨大クラウドAI （AGI志向）	リコー 実務特化型オンプレミスAI
データセキュリティ	外部送信必須 （ガバナンス上の懸念）	完全閉域網内でのローカル処理 （流出リスクゼロ） ✓
日本語図表推論性能	世界最高水準	ターゲット領域にリソースを集中し、 同等以上のスコアを達成 ✓
インフラ・運用コスト	継続的なAPI課金と 膨大な計算資源	量子化技術による市販ハイエンドPC レベルでのオフライン運用 ✓
カスタマイズ性 （暗黙知の学習）	制限あり	企業特有の専門用語や社内スラングの 個別ファインチューニングが可能 ✓

エンジン・コアの解剖：ベースモデル「Qwen」のアーキテクチャ優位性

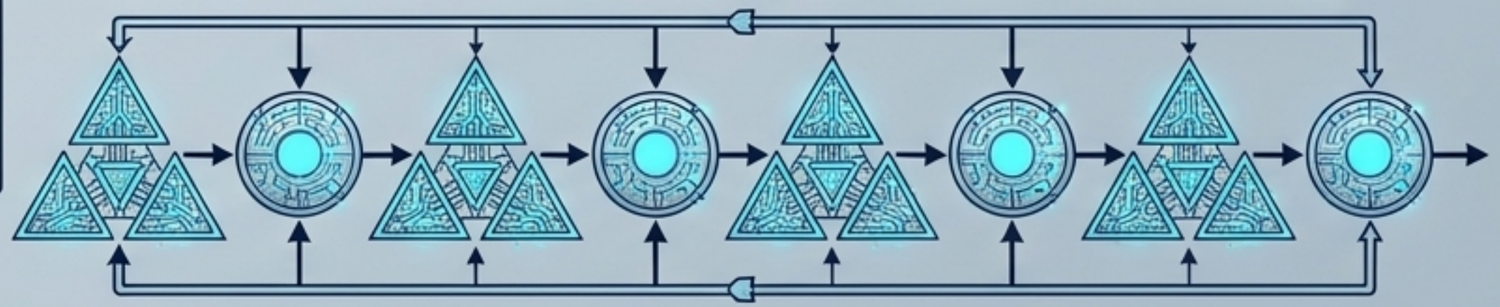
Qwen3.6-27B 高密度アーキテクチャ



- 270億の全パラメータが推論に参加する堅牢なブロック構造
- **Agentic Coding**（広範なコンテキストの横断理解）
- **Thinking Preservation**（思考プロセスの保持によるハルシネーション低減）

Qwen3.5-9B ハイブリッド構造

$8 \times (3 \times$ (Gated DeltaNet) $\rightarrow 1 \times$ (Gated Attention)$)$



状態空間モデルとTransformerの融合

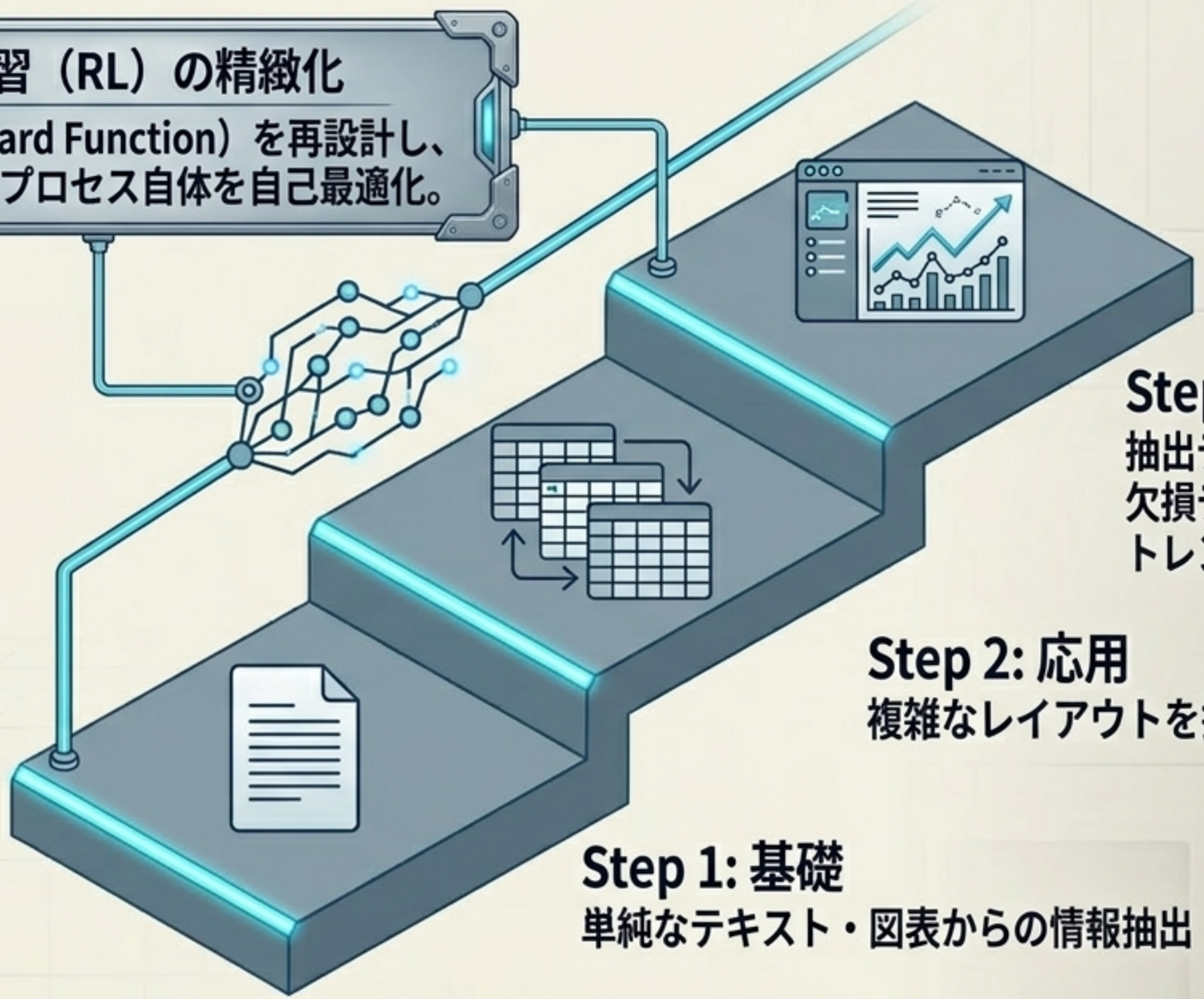
Early Fusion（視覚・言語の初期融合）により、エッジ環境でも10〜15秒でタスクを完了する圧倒的スピード。

GENIACプロジェクト成果

リコー独自のチューニング：論理的推論力を鍛え上げる「カリキュラム学習」

強化学習（RL）の精緻化

報酬関数（Reward Function）を再設計し、
「多段推論」のプロセス自体を自己最適化。



Step 3: 高度推論

抽出データに基づく四則演算、
欠損データの論理的補完、
トレンド推論

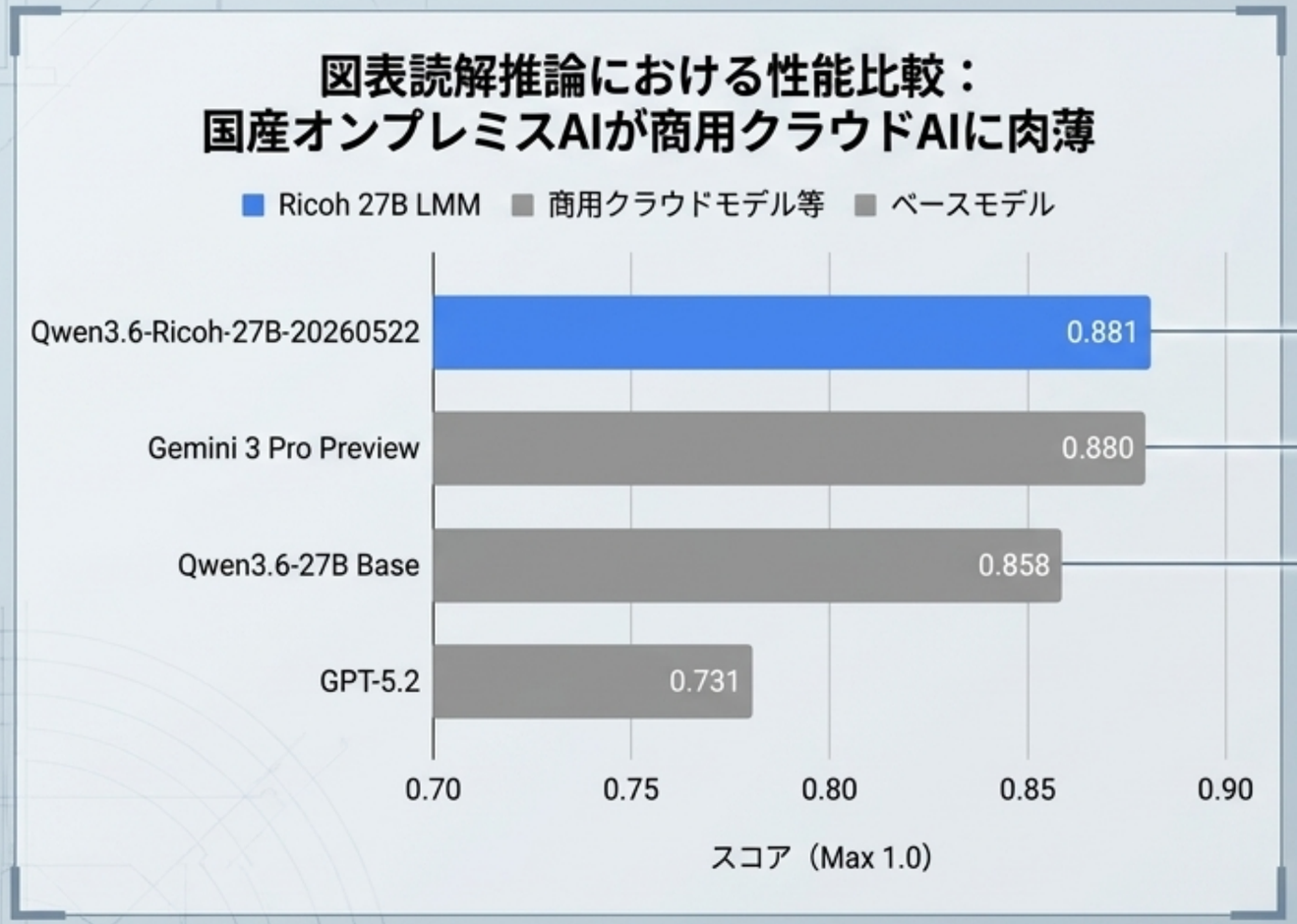
Step 2: 応用

複雑なレイアウトを持つ複数表の読解と連携

Step 1: 基礎

単純なテキスト・図表からの情報抽出

決定的な証拠（エビデンス）：「逆転現象」の証明



事実1

リコー27Bモデル（0.881）が、兆パラメータ規模と推測されるGemini 3 Pro Preview（0.880）と事実上同等に並び、GPT-5.2（0.731）を大きく凌駕。

事実2

ベースモデル（0.858）からの明確なジャンプアップ。リコーのチューニング技術の有効性を実証。

結論

「あらゆる知識を網羅する巨大AGI」ではなく、「特定ドメイン（日本語図表推論）にリソースを集中した中型特化モデル」が実務において圧倒的な費用対効果を生み出すことの証明。

包括的推論能力：視覚（VLM）と言語（LLM）の高次元な両立

視覚能力（VLM）

複雑なビジネス図表、決算書、路線図
などの正確な解析能力。

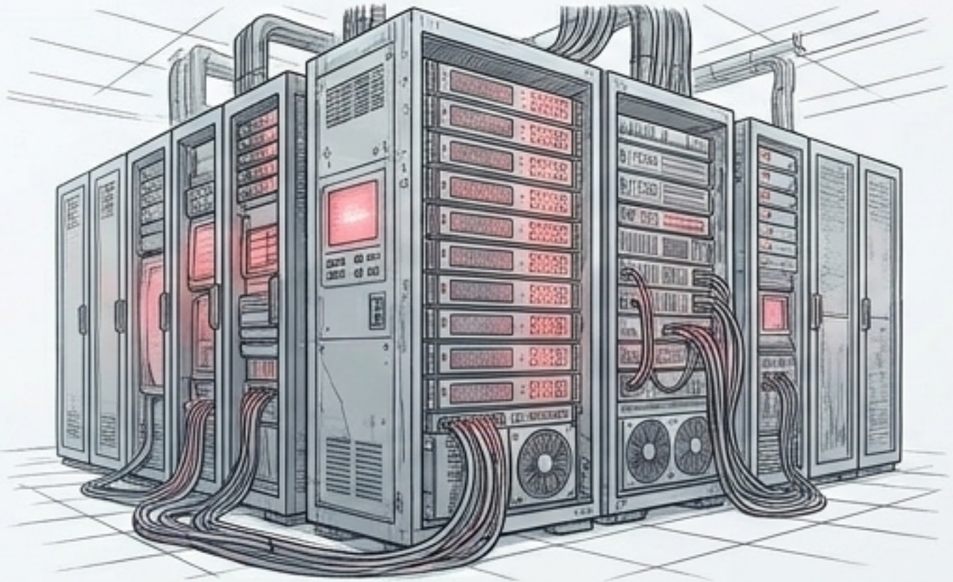
言語能力（LLM）

ベースモデルを凌駕する高水準の
自然言語処理・テキスト論理構築力。

9Bの軽量モデルに関しても、推論スコア0.782を達成しベースモデル (0.762) を凌駕。
スマートフォンやエッジデバイスでの高度運用への道を拓く。

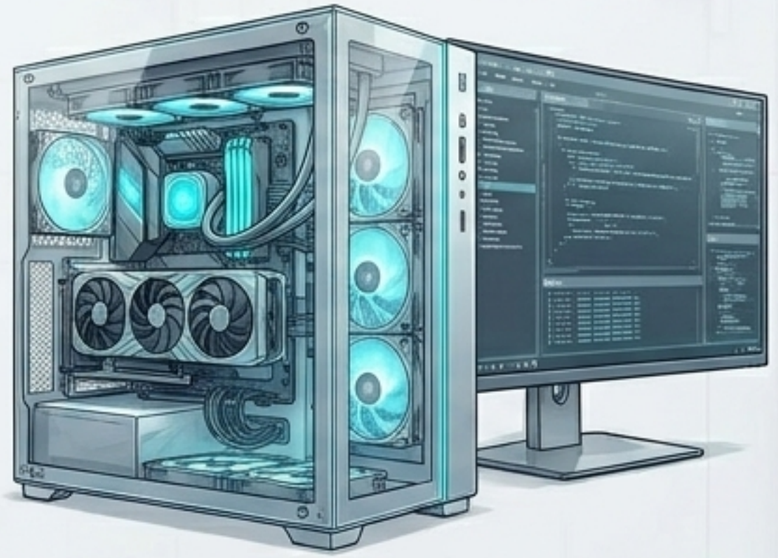
コスト革命：オンプレミス実装のボトルネックを破壊する「量子化技術」

従来のアプローチ（非量子化）



- H100 / A100クラスのデータセンター向けGPU（数百万〜一千万円規模）が複数必須
- 中堅企業や部門導入において致命的なコスト障壁

リコーのアプローチ（量子化）

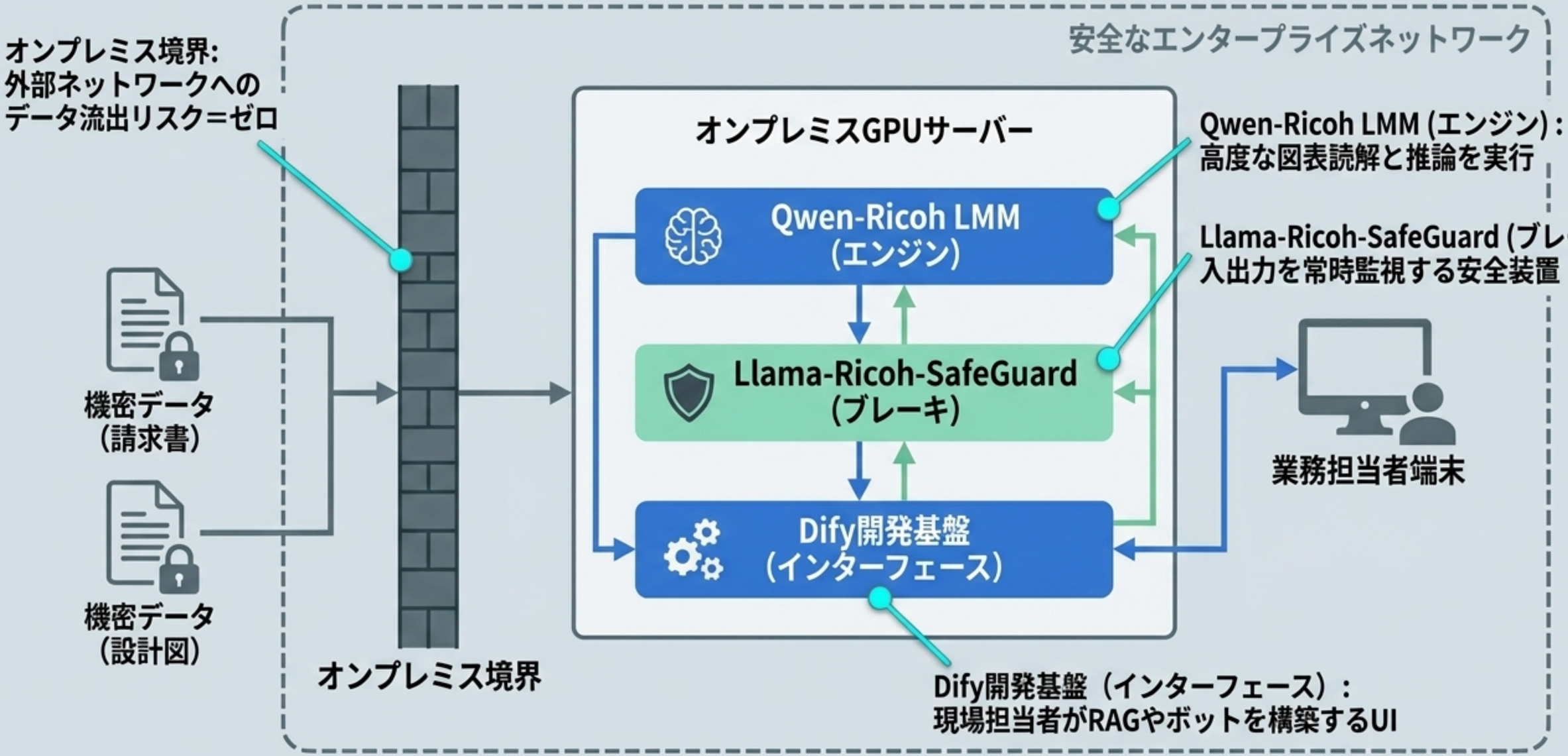


- AWQ-W4A16（4bit整数表現）による極限の圧縮

JDocQA-Reasoning: 0.868

数千万円の初期投資を不要にし、市販ハイエンドPCクラスの環境で、クラウド最高峰と同等のAIを完全オフライン駆動させるコストパフォーマンス革命。

セキュア・アーキテクチャ：完全閉域網の統合AIエコシステム



リスク・コントロール：AIの暴走を防ぐ強固なガードレール

データ漏洩だけでなく、倫理的リスク・コンプライアンス違反をも完全に自社のコントロール下に置く『二段構えの防壁』が、経営陣の導入ハードルを撤廃する。

実装の民主化と圧倒的な透明性

RICOH オンプレLLMスターターキット

「2025年日経優秀製品・サービス賞 最優秀賞」受賞

GPUサーバーの設置から、Difyによるノーコードアプリ構築、運用時の伴走支援までをワンストップで提供。医療・金融向けのテンプレートも実装。

開発コミュニティへの透明性

独自ベンチマーク「JDocQA-Reasoning」のデータセットと評価コードを完全無償公開。
ブラックボックスを排除し、第三者の検証可能性を担保することで日本のAI基盤モデル開発に貢献。

産業別ユースケース：「紙と図表」の制約から解放されるコア業務

1. 製造・エンジニアリング

入力: CAD図面画像 ＋ 要求仕様書テキスト → 処理: 寸法・材質・公差の適合性を自動確認

▶ 成果: 設計部門の確認作業の自動化、トラブルシューティングの瞬時特定

2. 金融・保険コンプライアンス

入力: 微小文字の約款 ＋ 多様なフォーマットの決算書 → 処理: クラウドに出さずに矛盾検知・サマリー作成

▶ 成果: 厳格な情報管理下での審査業務の超高速化

3. 公共・行政DX

入力: 手書き申請書 ＋ 複雑な統計図表 → 処理: 高精度な読解とデータ構造化 → 成果: 職員の入力・確認負荷の劇的低減と、行政サービスのスピード向上

4. 全産業ナレッジマネジメント

入力: 過去数十年のスライド資料 ＋ PDFレポート → 処理: グラフの売上トレンドの視覚的解釈とインデックス化 → 成果: 企業内の「集合知」の自律的な発掘と再利用

日本のエンタープライズAI戦略の「新たな羅針盤」

パラメータ規模をひたすら追及する汎用巨大AIの時代から、
特定のドメイン理解、運用コスト、セキュリティのバランスを極限まで最適化した
『実務特化型AIの社会実装』へと、世界はシフトしています。

卓越したアーキテクチャ、精緻な独自チューニング、そして透明なベンチマークという
『三位一体のアプローチ』により、リコーはかつてない事業価値の創出を証明しました。

Qwen-Ricoh LMMとスターターキットは、データガバナンスの壁に直面するすべての企業にとって、
次世代の知の利活用へ向けた最も現実的かつ強力なインフラとなります。

■ 次世代ナレッジマネジメントプラットフォーム「Hi.DEEN」への統合と、継続的な進化へ