ChatGPT

リコーの日本語特化型リーズニングLMM「Qwen3-VL-Ricoh-32B」の技術的評価と次世代AIインフラへの影響

エグゼクティブサマリ

本調査の結論は、**「日本語の“業務ドキュメント読解”を主戦場に定め、VL基盤（Qwen3-VL）へ“多段推論＋日本語化された思考過程”を後段学習で付与し、同時に運用コスト（画像トークン、チューニング工程）を下げる」**という設計一貫性にあります。これは、日本企業で支配的な成果物（図表・フローチャート・帳票・設計資料・IR資料など）に対する「実務の暗黙知抽出」と相性がよく、オンプレ/閉域を前提とした導入オプションを持つ点が、次世代AIインフラ（分散推論・メモリ階層化・低遅延推論・プライバシー保護）に対して具体的な要求を突き付けます。 [1]

公開されている定量評価の中核は、**JDocQA（公開）**と、図表中心に再設計した**JDocQA-Reasoning（独自、公開予定）**です。評価の重要点は「LLM-as-a-Judge」を採用しており、JDocQAはGPT-4o、JDocQA-ReasoningはGPT-4.1で審査していること、そして（参考として）32Bのスコアも併記されている点です。 [2]

一方で、ユーザーが求める「トレーニングデータの量・言語比率・ソース」「蒸留の有無」「推論レイテンシ/スループットの実測値」などは、現時点の一次情報からは**不明**です（本レポートでは不明点を明示し、必要箇所は前提を置いて概算します）。 [3]

インフラ面では、オンプレ導入パッケージ（RICOHオンプレLLMスターターキット）や、小型のGPU一体型サーバ（NVIDIA DGX Spark OEM）まで含む「配布・運用の製品化」が既に進んでおり、**“モデル単体の性能”よりも“運用可能な形で組織に入る”ことを重視する市場（日本）**に適合しています。 [4]

モデル概要と設計思想

位置づけと公開状況

「Qwen3-VL-Ricoh-32B-20260227」は、国産生成AI開発強化プロジェクトGENIAC第3期で、Qwen3-VL-32B-Instructをベースに開発完了した“リーズニング（多段推論）”志向のマルチモーダル大規模言語モデル（LMM）と説明されています。複数ページにまたがる図表を関連付けて理解し、読解難易度の高い質問へ高精度に回答することを狙っています。 [5]

ただし、**32Bモデル重みの一般公開は確認できません**。一方で、同技術を適用して開発した軽量モデル「Qwen-3-VL-Ricoh-8B-20260227」がHugging Face [6] で公開され、32Bはベンチマークの参考値としてスコアが併記されています。 [7]

ベースアーキテクチャ

ベースとなるQwen3-VLは、テキストと視覚（画像・動画）を統合して扱うVLMで、モデルカードでは「Dense/MoEの系統」「ネイティブ256Kコンテキスト（最大1Mへ拡張可能）」「OCR強化（32言語）」「視覚推論・視覚エージェント・視覚コード生成」などの機能を前面に出しています。 [8]

特にアーキテクチャ更新として、(1)Interleaved-MRoPE（時空間位置の扱いを強化）、(2)DeepStack（ViTの多段特徴を融合して画像-テキスト整合を改善）、(3)Text–Timestamp　Alignment（動画の時間位置をテキストで明示的に扱う）を挙げています。 [8]

マルチモーダル（VL）入力・実装I/F

Qwen3-VL系は、チャット入力のcontentにimage/textを混在させる形式を採り、Transformersの`AutoProcessor`と`apply_chat_template`でプロンプト化し、生成を行う流れがモデルカードに示されています。加えて、推論高速化・省メモリ化としてFlashAttention 2の利用が推奨されています。 [9]

Ricoh 8B公開モデルのvLLM実行例では、`qwen_vl_utils`で視覚入力を処理し、`multi_modal_data`として画像を渡す形（vLLM 0.11.0動作確認）になっており、画像の最小/最大ピクセル指定を通じて解像度制御も行っています。 [10]

公開情報の棚卸し

項目	一次情報で確認できる内容	備考
パラメータ規模	32B（320億）	GENIAC第3期の基本モデルとして説明。 [11]
ベースモデル	Qwen3-VL-32B-Instruct	明示。 [12]
多段推論	あり（“リーズニング性能”として強調）	強化学習・カリキュラム学習で強化。 [13]
日本語化された思考過程	8Bモデルで明示（`<think>`内も日本語化）	32Bでも同系統が示唆。 [14]
トレーニングデータ量/言語比率/ソース	**不明**	一次資料に数値・内訳なし。 [15]
蒸留の有無	**不明**	言及なし。 [16]
量子化の有無（Ricoh 32B）	**不明（少なくとも公開情報なし）**	Qwen3-VL側はFP8等の派生があるが、Ricoh 32Bの公開形態は未確認。 [17]
画像トークン圧縮	あり（推論時トークン約半減、精度劣化5%未満と説明）	コスト低減技術として明示。 [18]
ベンチマークツール	JDocQA（公開）＋JDocQA-Reasoning（独自、今後公開予定）	独自ツールの公開予定あり。 [19]

学習・ファインチューニングとデータ戦略

学習手法の全体像

“リーズニングLMM”としての性能強化について、一次情報で最も具体的なのは「学習の流れ」の説明です。そこでは、(1)高性能モデル出力からVQA要素を選別し自動で学習データ生成、(2)InstructionデータによるSFT（ファインチューニング）、(3)強化学習で推論結果を評価→モデル更新、(4)難易度を調整しつつデータ多様化・拡張（カリキュラム学習）、という段階的プロセスが示されています。 [20]

またニュースリリースでは、強化学習において独自の報酬関数を置き、学習効率を高めつつ過学習を抑制したこと、カリキュラム学習で難易度設定と学習ペースを最適化したことが明記されています。[11]

報酬設計の示唆

Ricoh　8Bモデルカードでは、強化点として「図表読解の深化（強化学習により推論プロセス導入）」「読み取った数値に基づく計算・比較分析の精度向上」「`<think>`タグ内を含め日本語化して根拠を明確化」を掲げています。これは報酬関数が、単純な正答率だけでなく「推論プロセスの形式」や「説明可能性」へも圧力を掛けている可能性を示します（ただし報酬関数の数式・重みは**不明**です）。[2]

データ戦略の“公表範囲”と不明点

ユーザー指定の「量・言語比率・ソース」は、現時点の一次資料では数値が開示されていません。したがって、**日本語比率や文書ドメイン構成は不明**です。[19]

ただし、設計思想としては「国内有数のビジネス文書データ量の活用」「設計図書やIR文書など多様な文書への対応」「日本企業特有の複雑な表・フローチャート・グラフの読解」を前面に置いており、データが“日本語×業務文書”へ強く寄っていることは一次情報から読み取れます。[21]

運用コストを下げる学習・推論側の工夫

技術ページでは、(a)推論時の画像トークンを圧縮してトークン数を約半分にし、精度劣化を5%未満に抑える、(b)用途特化LMMをマージしてチューニングコストを削減し、安価なプライベートモデル提供に繋げる、という2系統のコスト低減策が示されています。[18]

この2点は、単なる「モデル性能」ではなく「**インフラ制約（GPUメモリ/計算資源/コスト）に最適化された開発プロセス**」を同時に成果物化している点で、次世代AIインフラへの影響が大きい要素です。[22]

性能評価

評価ベンチマークの性質

JDocQA（公開）

JDocQAは、日本語文書（PDF）を対象としたドキュメントQAデータセットで、**5,504文書・11,600 QA**からなり、回答根拠ページやバウンディングボックスも付与されると報告されています。日本語特有の縦書き/横書きの混在なども課題として言及されています。[23]

論文内では、OpenAI　GPT（gpt-3.5-turbo-16k、gpt-4）をゼロショットベースラインとして評価しており、表（Table　4）としてスコアが掲載されています（ただしこのスコア体系は、Ricohが採用するLLM-as-a-Judgeの5段階系とは一致しません）。[24]

JDocQA-Reasoning（独自・公開予定）

RicohはJDocQAのテスト画像サブセット（図表を含むこと）に対して、1,000問以上の新規QAを付け直した独自ベンチマークを作成し、抽出・計算（四則演算/比率/統計集約など）・比較・補完タスクを含むと説明しています。難易度調整後に公開予定とされています。[2]

Ricoh公表の定量比較（JDocQA / JDocQA-Reasoning）

以下は、Ricoh 8Bモデルカード内で説明されている評価（vLLMで推論、審査はAzure OpenAI Service上でJDocQA=GPT-4o、JDocQA-Reasoning=GPT-4.1のLLM-as-a-Judge）に基づくスコアです。 10
※「Qwen3-VL-Ricoh-32B-20260227」は“参考値”として掲載されています。 2

モデル	JDocQA-Reasoning	JDocQA	備考
Gemini 3 Pro Preview	0.880	4.241	Ricoh評価（審査: GPT-4.1/4o） 10
Gemini 2.5 Pro	0.838	4.077	同上 10
Qwen3-VL-Ricoh-32B-20260227	0.826	4.076	同上（参考値） 2
Qwen3-VL-235B-A22B-Thinking	0.827	4.03	ニュースリリース側の比較表にも掲載 16
Qwen3-VL-32B-Thinking	0.799	3.99	同上 16
GPT-5.2	0.731	3.928	同上（注: Ricoh評価での比較） 2
Qwen-3-VL-Ricoh-8B-20260227	0.718	3.998	公開モデル 10
Qwen3-VL-8B-Thinking	0.699	3.890	公開モデル 10
GPT-4o	不明	不明	本枠組みではJDocQAの“審査役”として使用（被評価モデルでは未公表） 10
Llama 3（各派生含む）	不明	不明	同評価での被評価結果は未公表 2

この表が示す最重要点は、**32Bが（少なくとも当該条件下で）Gemini 2.5 ProとJDocQAで同等値（4.076≒4.077）に到達**し、JDocQA-Reasoningでも商用上位モデルに近接する、とRicohが主張していることです。 2

日本語理解・生成・推論・数学・コードの観点

Ricohの一次情報で「日本語総合（数学/コード含む）」を網羅する公表スコアは限定的です。その代替として、本モデル群が参照している日本語評価の枠組み（Japanese MT-Bench、ELYZA-tasks-100等）をRicoh自身が挙げている事実、ならびに外部の日本語評価プロジェクトがGPT-4oやLlama/Qwen系を比較している事実を整理します。 25

- Ricohは自社ラインアップ性能表記について、Japanese MT-Bench、ELYZA-tasks-100、JDocQA等に基づく「当社評価による性能水準の比較」であり、各社公式モデルの利用や同一性を示すものではない、と注記しています（数値は非開示）。 26
- Swallowプロジェクトの公開ページでは、日本語MT-Bench等でGPT-4oやLlama系・Qwen系の比較例が示されています（例：Llama 3.1 Swallow 70B Instruct v0.3の日本語MT-Bench平均0.7115、GPT-4o 0.7791等）。 27
- Qwen3 Swallowのページでは、日本語タスク平均スコア（例：0.609）が提示され、Qwen3 32Bを継続事前学習して日本語性能を押し上げた旨が述べられています。 28

従って、**「日本語の数学・コードまで含む網羅ベンチ」でRicoh 32Bがどの位置にいるか**は現時点では**不明**であり、一次資料上は「ドキュメント図表読解（JDocQA系）」の優位をもって技術的価値が主張されています。 19

推論特性と最適化

メモリ/VRAM要件の概算

前提として、32B級Transformerの重みメモリは、量子化なし（BF16/FP16）では概算で次式になります。

重みメモリ（概算）
- BF16/FP16: 32B params × 2 bytes ≒ **64GB**（※重みのみ。実際は各種バッファ等が上乗せ）
- INT8: 32B × 1 byte ≒ **32GB**相当
- INT4: 32B × 0.5 byte ≒ **16GB**相当

このため、「単一GPUにBF16で載せる」だけでも80GB級GPUが現実的な下限になりやすく、さらに長いコンテキスト（KVキャッシュ）やマルチモーダル入力（画像トークン増）を考えると、**推論エンジンのメモリ管理**がボトルネックになります（後述の画像トークン圧縮は、このボトルネックに直接効きます）。 29

長文脈・マルチモーダルが引き起こす“インフラ課題”

Qwen3-VLはネイティブ256Kコンテキストで、1Mへ拡張可能とされます。GitHub上ではvLLMサーブ時のRoPEスケーリングの例や、Interleaved-MRoPEでは位置IDの増え方が通常RoPEより遅いのでスケーリング係数を小さくする、といった運用上の注意も明示されています。 30

この長文脈は、マルチページPDFや長尺動画に効く一方で、推論時のメモリ（特にKVキャッシュ）とI/Oを急増させます。従って次世代AIインフラでは、(1)KVキャッシュのページング/階層化、(2)分散推論（テンソル並列・パイプライン並列）と通信最適化、(3)前処理（OCR/レイアウト解析/画像縮退）を含むパイプライン最適化、が不可避になります。 31

推論高速化の“公開ベストプラクティス”

Qwen3-VL 32Bのモデルカードは、FlashAttention 2を有効化すると「加速と省メモリ」に有効、と明記しています。 9

また、Ricoh 8B公開モデルはvLLM 0.11.0で動作確認し、推論をvLLMで実施しています。さらにMLPerf推論ベンチの文脈では、vLLMベースでOpenAI互換のChat Completions APIで任意推論システムをベンチ可能、という記述もあり、**OpenAI互換I/F＋vLLM**が実運用の共通化レイヤになりつつあることが読み取れます。 32

画像トークン圧縮の意味

Ricohは、画像トークン圧縮により「推論時トークン数を約半分」「精度劣化5%未満」と説明しています。 18

技術的に重要なのは、画像トークン圧縮が単に“視覚エンコーダの計算”を減らすだけでなく、**（長文脈で支配的になりがちな）KVキャッシュ成長を抑える方向に効く**点です。マルチページ文書でページ数・図表数が増えるほどマルチモーダルトークンが増えやすく、この圧縮はインフラコストの一次近似（計算量・メモリ量）に直結します。 33

量子化・蒸留の有無と効果

Ricoh　32Bモデルについて、量子化（INT8/INT4/FP8）や蒸留を適用して配布しているという一次情報は**不明**です。 [16]

一方、Qwen3-VL側はFP8版をリリースしている旨がGitHub上で示され、Hugging　Face上でも量子化派生が多数存在することが確認できます（ただし“Ricoh 32B”に直接は結びつきません）。 [30]

安全性・倫理と日本語固有リスク

モデル利用条件とガバナンス設計

Ricoh　8B公開モデルはApache-2.0に加えて追加利用規約を定め、(1)入力/追加学習/評価/出力利用の適法性と第三者権利侵害の回避を利用者責任とする、(2)医療・法律・金融・採用など説明責任が大きい分野で“唯一/主要根拠としての自動判断”を禁止し人的確認を要求する、(3)無保証・責任制限、などを明記しています。 [34]

これは「モデル単体の安全性」だけでなく、「**利用ポリシーを含む設計**」として、企業展開を前提にしたガバナンスの一部と評価できます。 [35]

セーフガードモデルの併用

技術ページでは、生成AI利用時の安全性確保のために、入力・出力の安全性チェックに対応する“セーフガード（ガードレール）モデル”を開発し、不適切・有害な入出力の自動検知や、暴力/犯罪/差別/プライバシー侵害などのカテゴリー判別を行う、と記載されています。 [36]

日本語固有のリスク論点

日本語で企業文書を扱う場合、一般的な有害性（差別・暴力等）に加えて、(1)敬語/婉曲表現による責任回避や誤解、(2)社内規程・契約書の“もっともらしい誤要約”、(3)固有名詞・組織名・人名の誤同定による風評/名誉毀損、(4)縦書き/表組みの読み違いに起因する数値ミス、が実害に繋がりやすい領域です。これを抑えるには「事前のリスク分類」「人手レビュー」「監査ログ」「評価の継続更新（Living）」が重要になります。 [37]

準拠しやすいフレームワーク

NIST [38] のAI　　RMFは、AIに伴うリスクを個人・組織・社会の観点で管理するための枠組みを提供しています。 [39]

日本国内では、経済産業省 [40] の「AI事業者ガイドライン」が版管理され（1.0→1.01→1.1→1.2など）、リスクベースアプローチやLiving Documentとしての継続更新が強調されています。 [41]

さらに内閣府 [42] も、AIの研究開発・活用における適正性確保の指針（生成AIの利活用促進とリスク管理を表裏一体で進める、継続的見直し等）を提示しています。 [43]

国際原則としてはOECD [44] のAI勧告が、信頼できるAIの責任ある取扱いと人権・民主主義的価値の尊重を掲げています。 [45]

運用・インフラ・エコシステムへの影響と将来展望

オンプレ/クラウドのデプロイ要件

Ricohは「オンプレ環境で個別カスタマイズ（プライベート化）が可能」「ラックマウント型サーバから小型PCサーバまで搭載可能なラインアップを開発」とし、閉域運用を強く意識しています。[46]

「RICOH オンプレLLMスターターキット」は、社内環境で動くローカルLLMパッケージで、GPUサーバ1台で稼働し、遮断環境で機微業務に利用できる旨が公式ページで説明されています。[47]

さらにスターターキットには、生成AIアプリ開発プラットフォームのプリインストール等を含むパッケージ提供が説明されており、“インフラ＋モデル＋アプリ基盤＋支援”としての提供形態が見えます。[48]

小型オンプレ推論の現実解

小型オンプレAIサーバとして言及されるのがDGX　Sparkです。NVIDIA[49]の公式情報では、DGX　Sparkは150×150×50.5mmの小型フォームファクタで、電源240W等が示されています。[50]

またデータシートでは、GB10 Grace Blackwell Superchip、128GB統合システムメモリ、最大200Bパラメータ級モデルをローカルで扱える、といった主張が示されています（ただし“最大200B”は量子化前提の可能性が高く、32BをBF16で高速に回すことを保証するものではありません）。[51]

参考コスト試算（推論TCOの概算）

ここでは「値が変動しやすい」前提を明示し、**代表例**として概算します。

クラウドGPU時間単価の例

Amazon Web Services[52]のEC2 Capacity Blocks for MLの価格表では、p5.48xlarge（8×H100）の実効時間単価が掲載されています（例：US　East等で$31.464/時・インスタンス、括弧内に$3.933/時・アクセラレータ）。[53]

よって、H100を1基ぶん確保する単純換算は **約$3.933/時** になります（Capacity Blockの枠組みである点に注意）。[54]

小型オンプレ（DGX Spark）の例

DGX　Sparkは国内販売例として税込90万円という提示が確認できます（小売価格例であり、調達条件で変動）。[55]
消費電力は240W（電源）とされています。[56]

電力単価は契約や時間帯で変動しますが、日本卸電力取引所の月次スポット平均（JPY/kWh）データが公開されています。Japan Electric Power Exchange[57][58]

例えば電力単価を **25円/kWh（仮定）** とすると、240Wの連続稼働電力コストは
0.24kWh × 25円≒**6円/時** となります（電力のみ）。
ハードウェア償却（例：90万円を3年・24/7で均す）を加えると、
900,000円 ÷ (3年×365日×24h)≒**34円/時**。
合算で **約40円/時（電力＋償却の超概算）** になります。

この比較は「同一性能・同一スループット」を仮定していないため、結論は“どちらが安い”ではなく、**オンプレは固定費構造、クラウドは可変費構造**であり、用途（常時稼働か、波動か、データ持ち出し可否か）で最適解が変わる、という点にあります。 59

スケーラビリティとハードウェア依存性

大規模側のスケールでは、AWSのP5インスタンスはEC2 UltraClustersで最大20,000 H100/H200 GPUまで相互接続可能と説明されています。これは学習だけでなく、推論を分散配置する際の「上限の大きさ」を示す一次情報です。 60

一方、Ricoh側は「小型PCサーバまで搭載可能」「閉域運用」「プライベート化」を全面に置くため、現実のシステム設計は**ハイブリッド（オンプレ＋必要に応じクラウド）**になりやすいと考えられます。 61

また、Ricohのラインアップ表では提供方法として、RICOHオンプレLLMスターターキットに加え、「Private AI Platform on PRIMERGY」や「NVIDIA DGX Spark OEMモデル」が併記されています。これは、GPU/サーバ調達の選択肢を複線化し、企業導入障壁（調達・設計・運用）を下げる戦略に見えます。 62

エコシステム影響と企業導入シナリオ

企業導入シナリオとして一次情報が示唆する中心は、(1)社内Wiki連携、(2)大量文書からの要点抽出、(3)Q&Aチャットボット、(4)営業支援エージェント等の社内情報活用です。オンプレ導入事例として、機密情報を扱うためセキュリティ/ガバナンス観点でスターターキットを選んだ、という説明もあります。 63

API/互換性の観点では、Ricoh　8Bモデルがv LLMでの動作確認とコード例を示し、MLPerf側でもOpenAI互換Chat Completions APIで推論システムをベンチ可能とされているため、**OpenAI互換I/Fを“社内標準API”として採用し、背後のモデルを差し替える**運用が合理的です。 32

将来展望と推奨アクション

次世代AIインフラへの具体的影響

1. **分散推論の標準化圧力**
   256K〜1M級の長文脈は、単一GPUでのKVキャッシュ保持を困難にし、GPU間分割やメモリ階層化を前提にした推論基盤へ移行させます。Qwen3-VL側がvLLMサーブ設定（RoPEスケーリング等）を具体例として提示している点は、運用が既に“分散・長文脈対応”へ踏み込んでいることを示します。 30

2. **前処理パイプラインのインフラ化**
   RicohはJDocQA画像生成時に文字が読めるサイズまで拡大して評価したと明言しており、入力前処理（PDFレンダリング、解像度管理、図表抽出、画像トークン圧縮）が品質・コスト双方の支配要因になります。 64

3. **プライバシー保護・データ主権の強化**
   オンプレ設計が前提となることで、「暗号化・アクセス制御・監査ログ・データ持ち出し禁止」をコンピュートの設計要件に織り込む必要が高まります。Ricohのスターターキットは完全ローカル・遮断環境での利用を強調しており、この方向性を後押しします。 65

4. **小型AIサーバ（統合メモリ）の再評価**
   DGX Sparkのような統合メモリ型は、単一の大容量メモリ空間でモデル/キャッシュを抱える設計を可能にします（性能・熱・帯域の制約は別途）。これは“開発・検証・小規模本番”のインフラ形態を変えます。 66

推奨アクション

企業側の推奨アクションは、(a)業務ドキュメントQAの評価セットを自社で整備し、JDocQA系（図表・複数ページ）と同型の“失敗パターン”を定義する、(b)OpenAI互換APIで推論ゲートウェイを作りモデル差し替え可能にする、(c)オンプレ運用では「監査ログ＋人手レビュー＋ガードレールモデル」の三点を最初からアーキテクチャ要件に入れる、(d)長文脈のコストが支配的になるため、画像トークン圧縮・RAG・要約・段階的推論（必要なときだけ深い推論）を組み合わせて“推論予算”を制御する、の4点です。 [67]

```
flowchart TB
  subgraph Input[企業ドキュメント入力]
    A[PDF/画像/スキャン] --> B[レンダリング・解像度制御]
    B --> C[図表/レイアウト抽出]
  end

  subgraph Inference[推論パイプライン]
    C --> D[画像トークン圧縮]
    D --> E[Qwen3-VL系VLM\n(32B/8B)]
    E --> F[多段推論 + 日本語<think>]
  end

  subgraph Safety[安全性/統制]
    F --> G[ガードレール\n(入力/出力判定)]
    G --> H[人手レビュー/監査ログ]
  end

  H --> I[業務アプリ\n(検索/要約/QA/エージェント)]
```

```
flowchart LR
  subgraph Cost[コスト構造の比較イメージ]
    P[オンプレ\n固定費中心\n(償却+保守+電力)] --> R[常時稼働で有利]
    Q[クラウド\n従量/可変費中心\n(GPU時間課金)] --> S[需要変動で有利]
    T[長文脈/図表増] --> U[KV/トークン増\n=コスト増]
    V[画像トークン圧縮] --> W[トークン削減\n=コスト抑制]
  end
```

---

[1] [3] [5] [7] [11] [12] [13] [15] [16] [19] [29] [52] [57] https://jp.ricoh.com/release/2026/0330_1
https://jp.ricoh.com/release/2026/0330_1

[2] [10] [14] [32] [34] [35] [49] [64] [67] https://huggingface.co/ricoh-ai/Qwen-3-VL-Ricoh-8B-20260227
https://huggingface.co/ricoh-ai/Qwen-3-VL-Ricoh-8B-20260227

[4] [47] [59] [65] https://www.ricoh.co.jp/products/list/ricoh-on-premises-llm-starter-kit
https://www.ricoh.co.jp/products/list/ricoh-on-premises-llm-starter-kit

[6] [38] [48] https://prtimes.jp/main/html/rd/p/000000172.000043114.html
https://prtimes.jp/main/html/rd/p/000000172.000043114.html

8 9 https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct
https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct

17 30 31 42 https://github.com/QwenLM/Qwen3-VL
https://github.com/QwenLM/Qwen3-VL

18 20 21 22 25 26 33 36 40 44 46 61 62 https://jp.ricoh.com/technology/ai/LLM
https://jp.ricoh.com/technology/ai/LLM

23 24 https://aclanthology.org/2024.lrec-main.830.pdf
https://aclanthology.org/2024.lrec-main.830.pdf

27 https://swallow-llm.github.io/llama3.1-swallow.ja.html
https://swallow-llm.github.io/llama3.1-swallow.ja.html

28 https://swallow-llm.github.io/qwen3-swallow.en.html
https://swallow-llm.github.io/qwen3-swallow.en.html

37 41 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/index.html
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/index.html

39 https://www.nist.gov/itl/ai-risk-management-framework
https://www.nist.gov/itl/ai-risk-management-framework

43 https://www8.cao.go.jp/cstp/ai/ai_guideline/ai_gl_2025.pdf
https://www8.cao.go.jp/cstp/ai/ai_guideline/ai_gl_2025.pdf

45 https://www.oecd.org/en/topics/ai-principles.html
https://www.oecd.org/en/topics/ai-principles.html

50 56 66 https://www.nvidia.com/en-us/products/workstations/dgx-spark/
https://www.nvidia.com/en-us/products/workstations/dgx-spark/

51 https://www.elsa-jp.co.jp/wp-content/uploads/2025/06/NVIDIA_DGX_Spark_A4_1P_ELSA.pdf
https://www.elsa-jp.co.jp/wp-content/uploads/2025/06/NVIDIA_DGX_Spark_A4_1P_ELSA.pdf

53 54 https://aws.amazon.com/ec2/capacityblocks/pricing/
https://aws.amazon.com/ec2/capacityblocks/pricing/

55 https://shop.elsa-jp.jp/view/item/000000001031?srsltid=AfmBOoo7bslVXmcgHn8hB1QdxabCrA_2K_oOMAu3RevJz3QDdeZmiPKF
https://shop.elsa-jp.jp/view/item/000000001031?srsltid=AfmBOoo7bslVXmcgHn8hB1QdxabCrA_2K_oOMAu3RevJz3QDdeZmiPKF

58 https://www.jepx.jp/en/electricpower/market-data/spot/ave_month.html
https://www.jepx.jp/en/electricpower/market-data/spot/ave_month.html

60 https://aws.amazon.com/ec2/instance-types/p5/
https://aws.amazon.com/ec2/instance-types/p5/

63 https://digitalpr.jp/r/127989
https://digitalpr.jp/r/127989