AUDIT REPORT

Alibaba Qwen3.8-Max 監査報告書

「2.4兆パラメータ級・超低コストモデルの実力とエンタープライズ統合アーキテクチャ」

Target Date:	2026-08-04
Classification:	企業AI戦略・システムアーキテクト向け内部資料

最終結論：価格破壊をもたらす有望なエージェントだが、全面移行には時期尚早

CORE SPECIFICATIONS

総パラメータ:	**2.4兆（Sparse MoE）**
推論時有効パラメータ:	**95B**
コンテキスト長:	**100万トークン**
API価格:	**入力 $2 / 出力 $6** (100万トークンあたり)

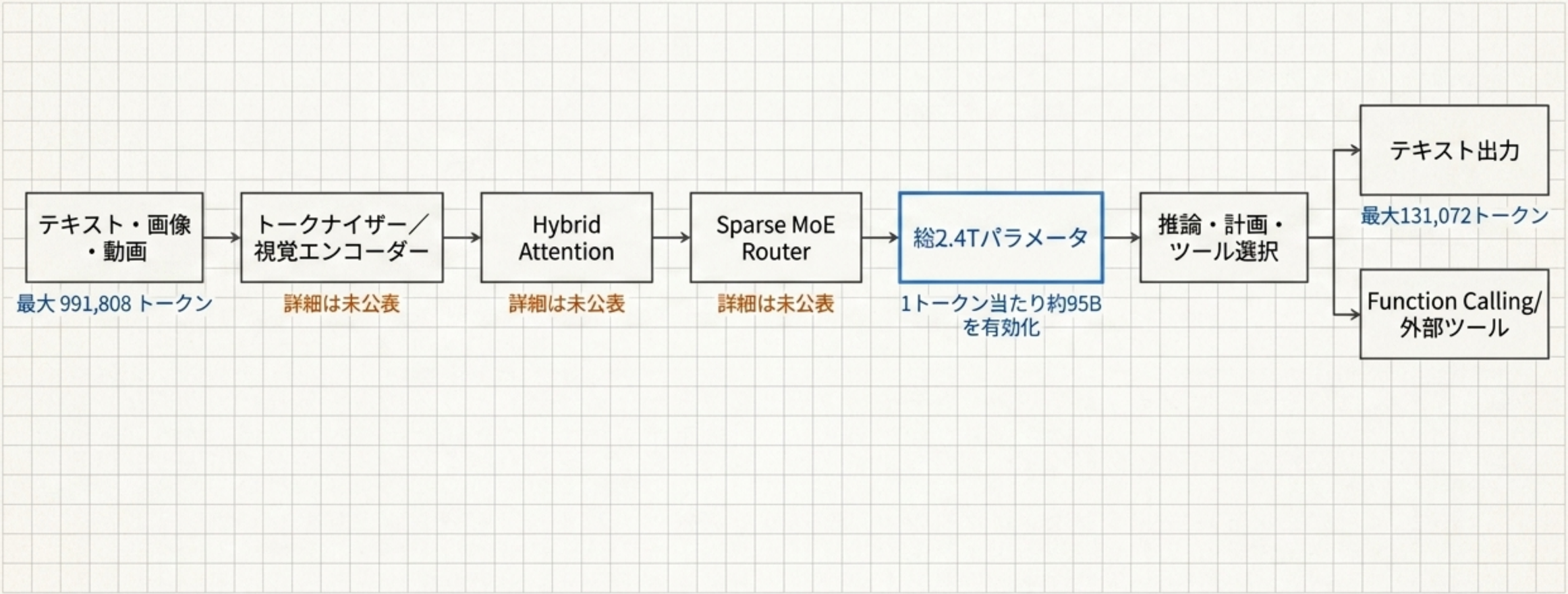
STRATEGIC VERDICT

- **価格帯が既存フロンティアモデルを圧倒。**
- **コーディング、長文解析、視覚タスクにおいて、低コストな第２モデル（ルーティング先）として極めて有望。**

CRITICAL RISKS

- **厳密なオープンウェイトモデルではなく、現時点ではAPI先行モデル。**
- **安全性、毒性、学習データに関する技術的透明性が著しく欠如。**
- **長期間の自律エージェント稼働におけるツール呼び出しの構造崩壊リスク。**

Qwen3.8-Max アーキテクチャ解剖図



情報公開監査：充実するAPI仕様、欠落する内部・安全性情報

VERIFIED

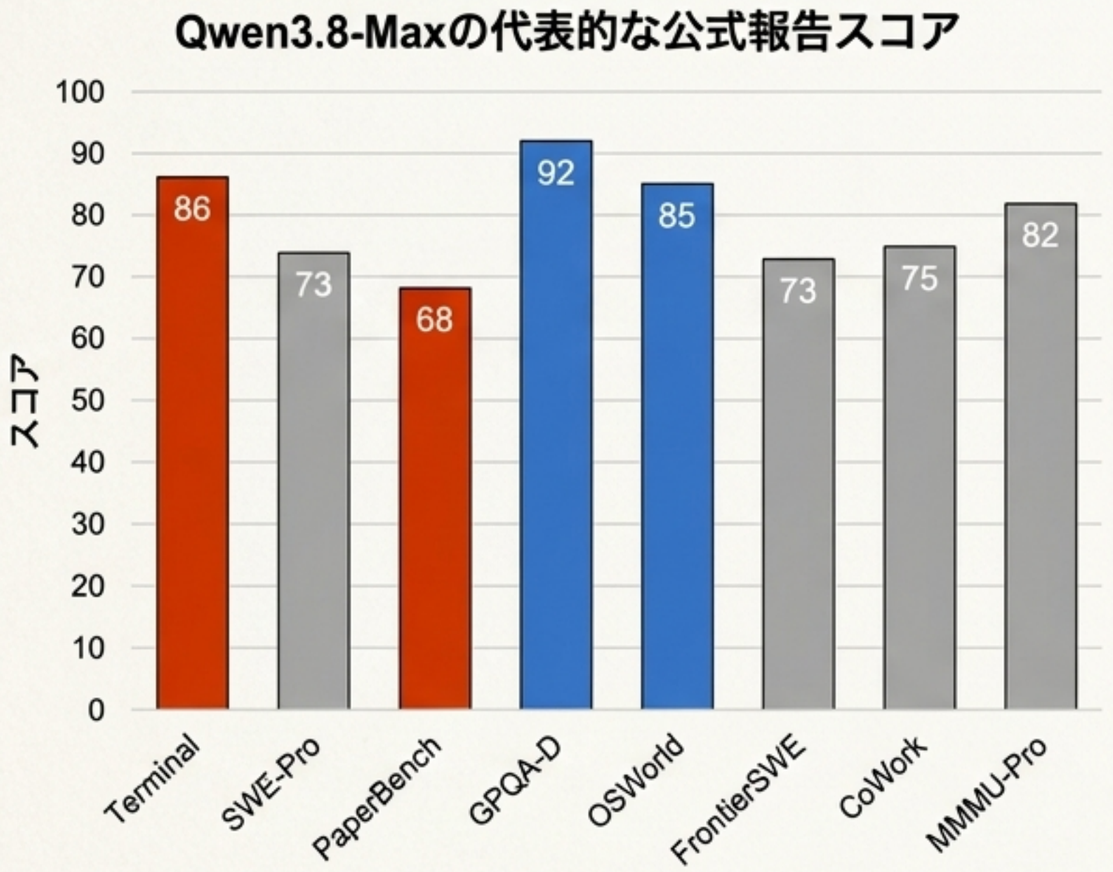
- 基本アーキテクチャ
  (Sparse MoE + Hybrid Attention)
- API仕様と機能
  （コンテキストキャッシュ、Function Calling等）
- 公式オンラインデモへの導線

MISSING

- モデル内部: レイヤー数、専門家数、
  学習データ内訳、トークナイザー
- **ライセンス:**
  商用利用条件、再配布条件（推定不可）
- **安全性評価:**
  毒性、脱獄、サイバー・生物化学リスク評価の欠如
- **基礎ベンチマーク:**
  MMLU、HELM、HumanEvalの正式結果

結論: 2026年8月4日現在、Qwen3.8-Maxは「再現可能なオープンソース」として評価できる段階にはない。

性能評価：研究再現・指示追従に優れるが、実務コード修正では後れを取る



Data Highlights
強み (Outperforms)
• PaperBench: 93.0 (Qwen3.7-Maxの64.8から劇的進化) • OSWorld-Verified: 86.1 (公式比較対象中で最高) • IFBench: 82.8
弱み (Underperforms)
• SWE-bench Pro: 67.7 (Fable 5の80.0に大きく劣る) • Terminal-Bench 2.1: 86.6 (GPT-5.6 Solの88.8に及ばず)

インサイト: これらのスコアはAlibaba自身の選択・実行によもの。コーディング全般で「最強」ではなく、長期的ワークフローやフロントエンドに強い一方、厳格なツール課題や実リポジトリ修正には課題を残す。

独立評価の現実：暫定的な上位入りと、エージェント挙動の死角

LMARENA 暫定順位表

VISION ARENA
Rank: 2位
Score: 1305±9

WEBDEV ARENA
Rank: 4位
Score: 1668±18
※投票数1563と少なく信頼区間が広い

TEXT ARENA
Rank: 5位
Score: 1496±10
※4位Claude Opus 4.6(1497±4)と統計的有意差なし

TRILOGY AI ブラインド評価

検索・ツール実行力: 優秀

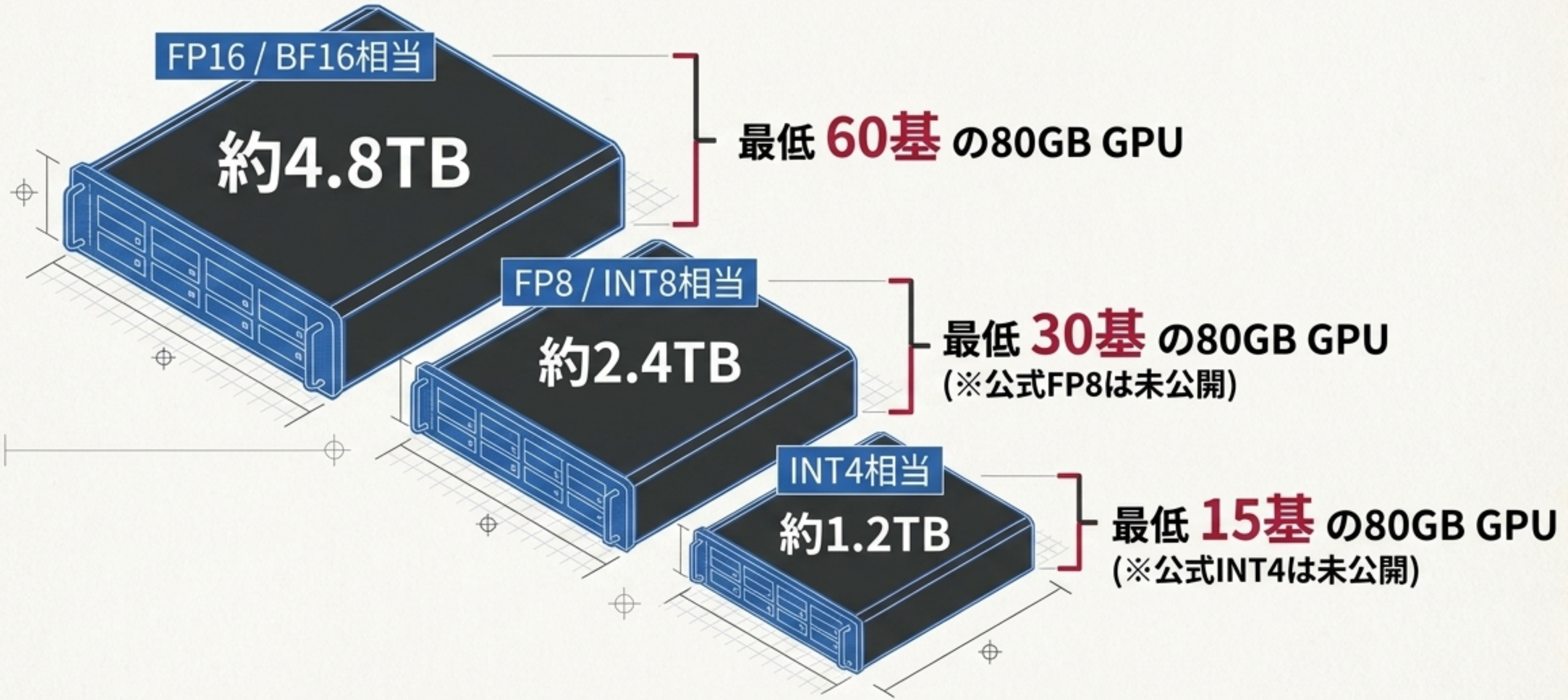
22回のゲートウェイ要求、44回のツール呼び出しで失敗ゼロ。10点中9点。

ハルシネーションの質的変化

354件の引用のうち、7つの主張群は「コードから証明できる範囲を超えていた」。

TAKEAWAY: 引用が存在することと、それが主張を論理的に立証していることは別である。二次判定モデルによる検証が必要。

オープンウェイトの錯覚：2.4兆パラメータが要求する物理的制約



Strategic Implication: 「誰でもローカル実行できる民主化」ではない。95Bのアクティブパラメータだけをメモリに置けば動くわけではなく、ルーターが選択し得る全専門家の重み（1.2TB〜）を保持する必要がある。通常の単一8GPUノードでのLoRAは困難。

コスト破壊：フロンティア級性能をコモディティ価格で提供

Qwen3.8-Max (QwenCloud) - 100万トークンあたり

入力 $2.00
出力 $6.00

(参考) Claude Opus 4.6	入力 $5.00 / 出力 $25.00
(参考) Fable 5	入力 $10.00 / 出力 $50.00

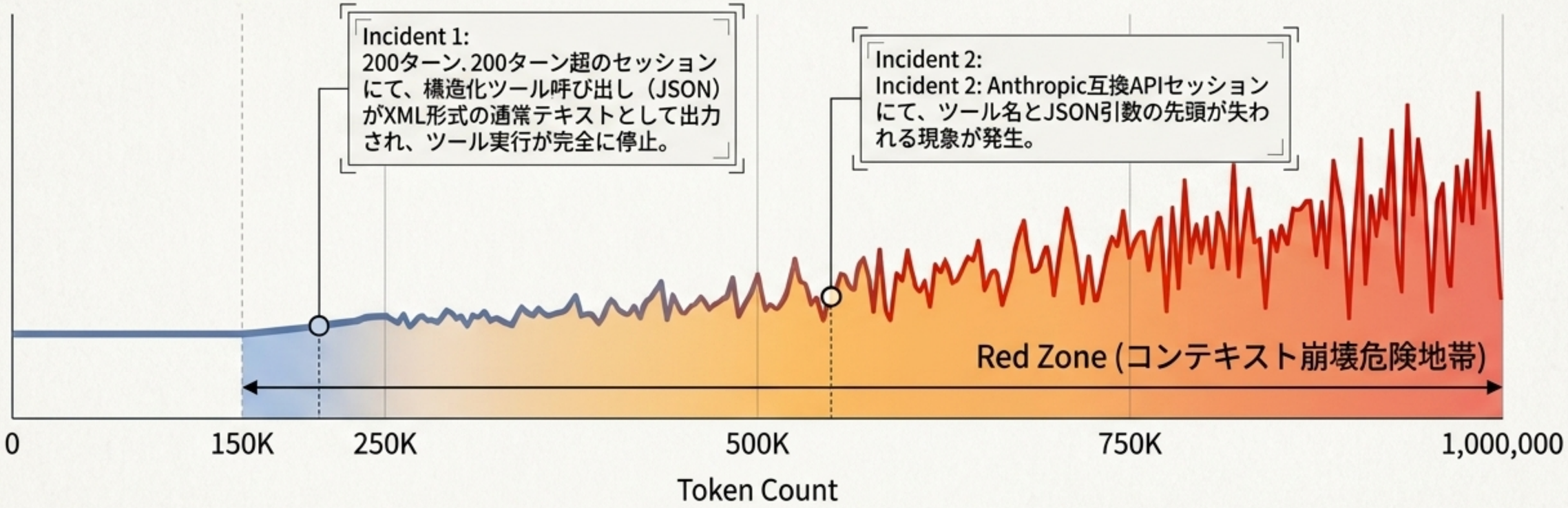
Tokyo Region Pricing (Alibaba Cloud)

入力: 12 RMB / 出力: 36 RMB
キャッシュヒット入力: 1.5 RMB

※100万トークン入力＋20万トークン出力の場合、概算で 19.20 RMB ($3.20)。

Takeaway: 性能差が小さい業務において、ClaudeやGPTの代替（第2モデル）やルーティング先として採用することで、システム全体の運用コストを劇的に引き下げることが可能。

既知の脆弱性：長時間エージェント運用における「文脈の崩壊」



Risk Mitigation (本番環境への教訓): 100万トークンの最大長に依存した反復送信はコストと遅延を急増させる。本番環境では「会話履歴の外部状態管理」「定期的な再要約」「コンテキストキャッシュの利用」が不可欠。

ガバナンス・レーダー：多極化するAI規制とコンプライアンス要件

EU（AI Act）

2026年8月2日よりGPAI義務本格化。下流システム提供者としての透明性、人間監督、重大インシデント報告が要求される。

中国（生成式人工智能服务管理暂行办法）

中国国内で一般向けサービスを提供する場合、合法性、コンテンツ管理の順守が必要（API利用自体とは別の法的問題）。

日本（AI事業者ガイドライン）

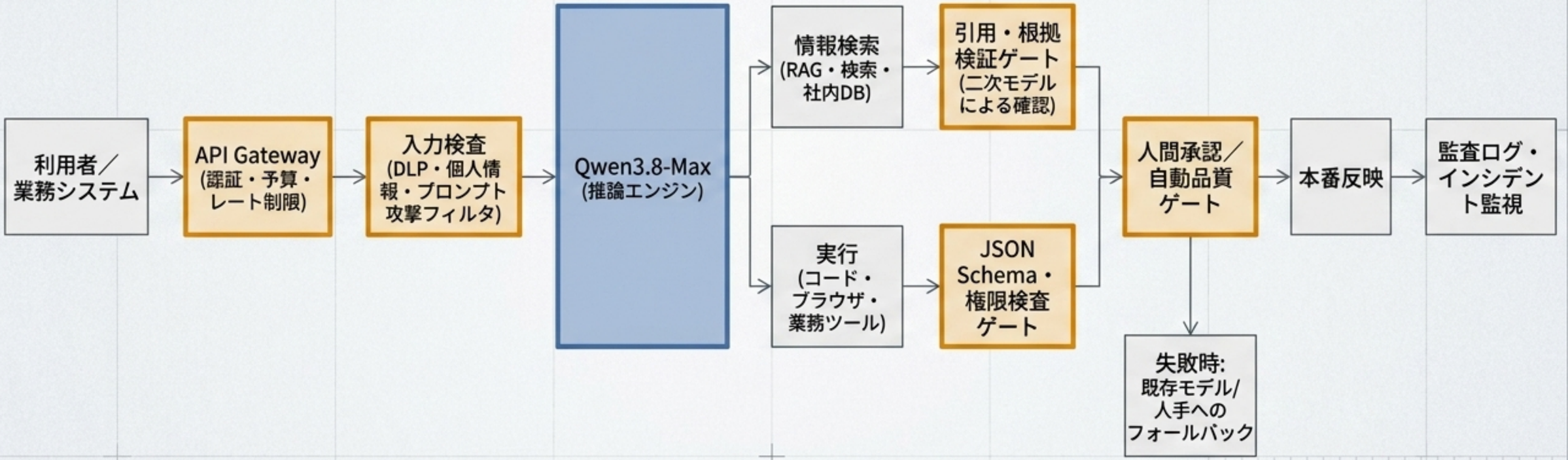
経済産業省ガイドラインに基づく、リスクベース評価、セキュリティ、継続監視の自社ガバナンスへの組み込み。

米国（BIS 輸出管理）

先端計算品目に対する輸出管理。中国系モデルの訓練・ホスティング能力やGPU調達コストに影響し得る中長期な構造的供給網リスク。

Architect's Dossier

統合アーキテクチャ・ブループリント：安全性を担保する保護設計



ユースケース適否判断：機械的検証が可能なタスクへの特化

推奨（優先検証領域）

- 既存コードベースの調査・設計レビュー
- フロントエンド試作とテスト生成
- 長文書・PDF・動画の分析（RAG連携）
- 定型オフィス業務・GUI操作の検証

理由: 成果物をテスト、静的解析、スキーマ検証で「機械的に確認できる」ため。

非推奨（現段階での単独利用回避）

- 無監督の本番インフラ変更
- 医療診断、法的最終判断、大額決済
- 自動採用・人事評価、安全制御（工場・インフラ）

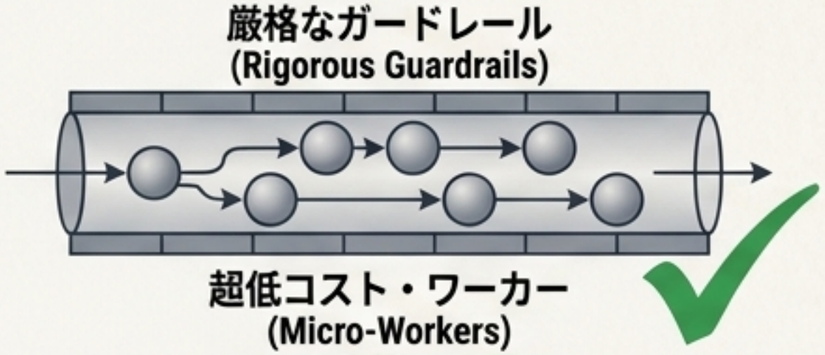
理由: モデル能力の不足ではなく、安全性評価・モデルカードの欠落と、長期間ツール利用の信頼性データが不足しているため。

マイグレーション戦略：段階的なカナリアテストと評価基準

From GPT-4o / OpenAI API	OpenAI互換エンドポイントを利用。ただし、ストリーミング、JSON Schema、エラー処理の挙動差分を再試験すること。
From Claude	Anthropic互換APIを利用可能。必須テストとして「長時間ツールセッションでの構造崩壊（XML化）」を検証。
From Qwen3.7-Max	同一評価セットで3.7と3.8を並列実行し、純粋な品質向上分とトークン消費増加分を比較測定。
From RAG（長文処理）	最初から100万トークンを丸ごと投入せず、検索・再ランキング・引用確認（二次判定）と組み合わせた設計を維持する。

Synthesis：自律型メインエージェントではなく、「厳格なガードレール内の超低コスト・ワーカー」

非推奨 (Not Recommended)



推奨 (Recommended)

1. Qwen3.8-Maxの真の価値は、人間に代わって16日間自律稼働することではない。
2. 安全性データが未知数で、長時間のコンテキスト保持に不安定さを残す現状では、「単一の巨大な頭脳」としてシステムの中央に据えるべきではない。
3. 最適解は、徹底した権限管理、JSON Schema検証、既存モデルへのフォールバックを備えたアーキテクチャの末端に、「使い捨て可能な超低コストの専門タスク・ワーカー」として大量配置することである。

Closing Note: 重み、ライセンス、モデルカードが公開される「次週」以降、ローカルホスティングの評価フェーズへと移行する。短期的にはAPIによる数週間の本番類似テストを開始せよ。