ChatGPT

# Qwen3.7-Max 調査報告

## エグゼクティブサマリー

Qwen3.7-Max は、Alibaba/Qwen が 2026年5月に公開した「agent era」向けの最新プロプライエタリモデルで、単なる対話品質よりも、長時間の自律実行、外部ツール利用、コーディング、MCP を介した業務自動化を前面に出しています。公式説明では、Claude Code、OpenClaw、Qwen Code など異なるエージェント・ハーネス間で一般化すること、35時間の完全自律カーネル最適化、1,000回超のツール呼び出し、さらに実測では 1,158 回のツール呼び出しと 10.0x の幾何平均高速化を達成したことが、最大の差別化要因として提示されています。 [1]

ただし、導入判断で重要なのは、**派手な公式ベンチマークと、独立評価の厚みは別問題**だという点です。独立ソースでは、Artificial Analysis の Intelligence Index で Qwen3.7-Max は 57、1M コンテキスト、112 tok/s と高い水準にありますが、同時に 97M 出力トークンという非常に強い冗長性も観測されています。BenchLM でも上位帯に入りますが、公開カバレッジはまだ限定的です。公開の Terminal-Bench / SWE-bench の可視ページからは、現時点で Qwen3.7-Max の独立再現結果を私は確認できませんでした。したがって、**「強い可能性は高いが、法律・知財用途では未検証部分が残る」**というのが実務的な評価です。 [2]

知財実務との相性は、単独判断モデルとしてより、**検索・OCR・PDF 解析・表計算・コード実行を束ねるオーケストレーション層**として使うときに高いです。Model Studio は OpenAI 互換 Responses API で Web search、Web extractor、Code interpreter、Knowledge base search、Function calling を提供し、Qwen-OCR はスキャン文書や表の抽出も担えます。したがって、先行技術調査、パテントランドスケープ、請求項要素分解、契約レビュー、期限抽出のような「収集→構造化→比較→要約」の流れには強く、逆に最終的な請求項確定、FTO 結論、発明者認定、秘匿性の高い未公開案件の丸投げには向きません。 [3]

## 公式仕様とリリース

Qwen3.7-Max の公式ポジショニングはかなり明確で、Qwen チーム自身が「agent era 向けの最新 proprietary model」と呼び、コーディング、オフィス自動化、長周期の自律実行、ハーネス横断一般化を中核能力としています。公開時期は、Qwen 公式ブログが 2026年5月19日、Alibaba Cloud Community の英語記事が 2026年5月21日で、報道各社はその間の Alibaba Cloud Summit での発表として扱っています。 [4]

| 公式ページ | 何が分かるか | 導入判断での意味 |
|---|---|---|
| Qwen 公式「Qwen3.7: The Agent Frontier」 | Qwen3.7-Max の位置づけ、長時間自律実行、総合ベンチマーク主張。 [5] | 最初に読むべき一次情報。 |
| Alibaba Cloud Community 版「Qwen3.7: The Agent Frontier」 | 35時間・1,158 ツールコール・10x 高速化、ベンチマーク条件、Cross-Harness Generalization の説明。 [1] | 技術的な読み解きに最重要。 |
| Model Studio Models / API | サポートモデル、OpenAI 互換 Responses API、ツール呼び出し、地域配備の考え方。 [6] | 実装面の現実解。 |
| Model Studio Pricing | `qwen3-max` 系のトークン課金、バッチ割引、キャッシュ割引。 [7] | 公式価格の基準。 |

| 公式ページ | 何が分かるか | 導入判断での意味 |
|---|---|---|
| Web search / Web extractor / Code interpreter / OCR / KB docs | エージェントに必要な検索・抽出・実行・RAGの構成要素。 [8] | 実務ワークフローを組めるかを決める。 |

## 仕様の要点

公式記事ベースで見ると、Qwen3.7-Max の本質は「高性能チャットモデル」ではなく、**長時間動き続けるツール指向エージェント基盤**です。Alibaba Cloud Community 版では、数百〜数千ステップの自律実行、MCP とマルチエージェントによるオフィス生産性支援、クロス・スキャフォールド一般化が強調され、35時間の自律カーネル最適化では 432 回のカーネル評価と 1,158 回のツール呼び出しを行ったとされています。 [1]

一方で、**Qwen3.7-Max 固有のパラメータ数は、私が確認できた公式 3.7 発表資料では明示されていません。**近い一次情報としては、2025年9月公開の Qwen3-Max が「1 兆超パラメータ」「36 兆トークンの事前学習」「Qwen3 系設計 + global-batch load balancing loss」を採用すると説明されており、Qwen3.7-Max はそこからさらにエージェント学習と長周期 RL を強化した系譜と読むのが妥当です。シリーズ全体の訓練データ開示文書は、Qwen/Wan 系がテキスト・画像・動画・音声を跨ぐ「trillions of tokens」規模で、公開 Web データ、契約パートナーデータ、ラベル付き/委託データ、合成データを用いたと述べるに留まり、Qwen3.7-Max 固有のコーパス量は開示していません。 [9]

## 旧 Qwen 系から見たアーキテクチャ変化

| 世代 | 公開されているサイズ情報 | 主要な設計・学習上の変化 | Qwen3.7-Max にどうつながるか |
|---|---|---|---|
| Qwen2.5 | 7B〜72B、最大 18T トークンの事前学習。 [10] | 知識量とコーディング/数学強化。 [10] | まだ「長時間エージェント」より基礎性能中心。 |
| Qwen3 | 0.6B〜235B の dense/MoE 系、約 36T トークン、119 言語・方言、MCP と function calling を強化。 [11] | ハイブリッド reasoning と agent 能力を公開系に持ち込んだ。 [12] | 3.x 系の共通基盤。 |
| Qwen3-Max | 1T 超パラメータ、36T トークン。 [13] | MoE、global-batch load balancing loss、1M context training、Qwen2.5-Max 比 30% MFU 改善。 [13] | 3.7-Max の「大規模 proprietary backbone」の直系。 |
| Qwen3.5 | 397B-A17B を open-weight 化。 [14] | より高疎な MoE、Gated DeltaNet + Gated Attention、multi-token prediction、201 言語/方言、早期 text-vision fusion。 [15] | 効率化・多言語・マルチモーダル化。 |
| Qwen3.6-Plus | サイズ非公開。 [16] | 長周期 planning と tool-calling で首位級、real-world agents への寄せ。 [16] | 3.7 の前段となる「現実世界エージェント」化。 |

| 世代 | 公開されているサイズ情報 | 主要な設計・学習上の変化 | Qwen3.7-Max にどうつながるか |
|---|---|---|---|
| Qwen3.7-Max | サイズ非公開。[1] | Environment scaling、Task/Harness/Verifier 分離、cross-harness RL、長時間自律最適化、reward hacking 監視。[17] | 3.x 系を「実運用エージェント」へ押し切った段階。 |

補足すると、Qwen3.7-Max 自体は公開ウェイトではなく、少なくとも公式発表では proprietary 扱いです。独立評価サイトでも text input / text output の reasoning モデルとして整理されており、図面やスキャン PDF をそのまま読む用途では、Qwen-OCR や Qwen3.7-Plus のような別モデルとの分業が前提になります。[18]

## 第三者評価とベンチマーク

独立評価を見ると、Qwen3.7-Max は「かなり強い」が、**検証レイヤーはまだ薄い**です。最も使いやすい独立ソースは Artificial Analysis で、同社の Intelligence Index v4.0 は GDPval-AA、$\tau^2$-Bench Telecom、Terminal-Bench Hard、SciCode、AA-LCR、AA-Omniscience、IFBench、Humanity's Last Exam、GPQA Diamond、CritPt を含む英語・テキスト中心の総合指数です。その基準で Qwen3.7-Max は 57 を記録し、同価格帯平均を大きく上回ります。[19]

### 独立比較の見取り図

| モデル | Artificial Analysis Intelligence Index | コンテキスト | 補足 |
|---|---|---|---|
| Qwen3.7-Max | 57。[20] | 1M。[20] | 112 tok/s、ただし 97M 出力で非常に冗長。[20] |
| GPT-4.1 | 26。[21] | 1M。[21] | 画像入力対応の現行比較対象。[21] |
| GPT-4o | 19。[22] | 130k。[22] | 非 reasoning。[22] |
| GPT-4o mini | 13。[23] | 128k。[23] | 低価格帯。[23] |
| Llama 4 Maverick | 18。[24] | 公開スニペットでは未確認。 | Qwen3.7-Max より明確に下。[24] |
| Llama 3.3 70B | 14。[25] | 公開スニペットでは未確認。 | open-weight とのギャップが大きい。[25] |

現行の第三者比較では、**GPT-4 そのものより GPT-4.1 / GPT-4o 系が主比較軸**になっており、GPT-4 の新しい同条件比較は確認しにくい状態です。したがって、2026年時点の購入判断では GPT-4 ではなく GPT-4.1 と GPT-4o を対照群に置く方が実務的です。[26]

### 公開ベンチマークで確認できる範囲

| 出典 | 何を見るか | Qwen3.7-Max の数値 | 読み方 |
|---|---|---|---|
| Artificial Analysis | 独立総合指数 | 57。[20] | 強いが、冗長性が高くコスト増を招きやすい。[20] |
| Arena.ai Text Arena | 人手比較 ELO | `qwen3.7-max-preview` が 1474±10、暫定順位 17。[27] | **Preview** の数値であり、製品版そのものではない点に注意。[27] |
| BenchLM | 集約型比較 | Overall 91/100、暫定 #5、verified #3。公開 51/247 ベンチ。[28] | 上位帯だが、公開カバレッジはまだ部分的。[28] |

### 公式自己申告の headline benchmark

以下は **Qwen 公式の自己申告**ですが、比較対象とセットで見れば相対位置は分かります。

| ベンチマーク | Qwen3.7-Max | 比較対象 | コメント |
|---|---|---|---|
| Terminal-Bench 2.0-Terminus | 69.7。[1] | DS-V4-Pro Max 67.9。[1] | agentic coding/ terminal 操作の主張点。 |
| SWE-Verified | 80.4。[1] | Opus-4.6 Max 80.8、DS-V4-Pro Max 80.6。[1] | 最上位帯だが、差は僅差。 |
| GPQA Diamond | 92.4。[1] | Opus-4.6 91.3。[1] | 高難度 STEM reasoning を強く訴求。 |
| HLE | 41.4。[1] | Opus-4.6 40.0。[1] | 難問総合系。 |
| Kernel Bench L3 | 1.98x median speedup / 96% win rate。[1] | Opus-4.6 98%、GLM 5.1 78%、Kimi K2.6 80%、DeepSeek V4 Pro 54%。[1] | 長周期最適化の具体例としては説得力がある。 |

ただし、ここでのベンチマーク体系には QwenWebDev、CoWorkBench、QwenWorldBench、SpreadSheetBench-v1 など**内部評価**もかなり混ざっています。CoWorkBench は法務・医療・金融などの生産性ドメインを含むと説明されていますが、法務・知財の公開再現ベンチマークではありません。さらに、Qwen 公式もスコアが多様な agent scaffold から来ていることを強調しており、単一ハーネス最適化ではない点は前向きですが、逆に言えば再現条件が複雑です。[1]

実際、現在アクセスできる公開 Terminal-Bench ページの可視結果には Qwen3.6-35B-A3B や Qwen 3 Coder は見えますが、Qwen3.7-Max は見当たりませんでした。また、公開 SWE-bench ページの可視テキストからも Qwen3.7-Max を私は確認できませんでした。したがって、**Terminal/SWE 系 headline score は 2026年6月時点では vendor-claim 扱いが安全**です。[29]

## コミュニティと業界での評価

業界報道のトーンは概ね好意的で、特に日本メディアは「エージェント基盤」「35時間の自律作業」「1000回超のツール呼び出し」「オフィスワークフロー自動化」を見出しレベルで取り上げています。GIGAZINE、Impress AI Watch、CodeZine はいずれも、Qwen3.7-Max を“チャット AI”というより“働くエージェント”と

して紹介しており、VentureBeat も Claude Code のような外部ハーネスと組み合わせられることを注目点にしています。[30]

コミュニティの温度感は、**能力には高評価、運用面には慎重**です。中国圏の V2EX では、国内レイテンシと API 安定性、ツール呼び出しとの相性、前世代からの大きなコーディング改善を評価する声が目立ちます。その一方で、「複雑プロジェクトの上限性能は Claude + Claude Code がまだ上」「明示した変更要求から逸脱して“自分の案”を入れがち」「チーム Token Plan の消費が速い」といった実務寄りの不満も繰り返し出ています。[31]

Reddit / Qwen_AI 系では、Artificial Analysis の高得点を歓迎する声がある一方、「hallucinates too much」という直球の不満、トークン消費の大きさ、open-weight への期待も目立ちます。独立評価でも Artificial Analysis は Qwen3.7-Max を「high intelligence but very verbose」と評しており、この“高性能だが冗長”という印象は媒体横断で一致しています。[32]

GitHub の実装者コミュニティでは、能力評価よりも**接続と認証の摩擦**が目立ちます。例として、OpenCode/Claude Code 連携での 401、OpenAI auth type で `qwen3.7-max` が選べない、モデル一覧に出ない、プロバイダ互換で 404/401 が出る、といった issue が複数あります。これはモデル本体の能力というより、提供面の成熟度の問題です。知財のような業務システム連携前提の現場では、この点は無視できません。[33]

| 繰り返し出る評価 | 観測された内容 | 実務上の意味 |
|---|---|---|
| 強み | コーディング/エージェント性能の伸び、国内低遅延、業務オーケストレーション適性。[34] | 知財調査・表計算・文書処理に向く。 |
| 懸念 | 冗長でコストが上がりやすい。[35] | 長い調査・ドラフトでは予算管理が必要。 |
| 懸念 | 幻覚、過剰編集、tool-call まわりの不安定さ。[36] | 「根拠付き出力」と JSON 制約が必須。 |
| 懸念 | open-weight 不在、オンプレ不可、統合摩擦。[37] | 機密案件では代替アーキテクチャ検討が必要。 |

参考までに、OpenCode が観測する最近 2か月の利用データでは、Qwen3.7-Max は最近の token usage で #18、観測量シェア 0.3%、完了セッション 251,280 とされており、「話題先行で終わっていないが、まだ圧倒的覇権でもない」位置です。[38]

## 知財実務への適用設計

知財実務で Qwen3.7-Max を使う場合、正しい設計は「モデル単体に判断させる」ことではなく、**特許 DB、契約 DB、社内ナレッジ、OCR、表計算、検索を組み合わせた多段パイプライン**に置くことです。Model Studio は OpenAI 互換 Responses API 経由で web search、web extractor、code interpreter、knowledge base search、function calling を持ち、Qwen-OCR はスキャン文書・表・数式抽出に対応します。J-PlatPat、Espacenet、USPTO Patent Public Search、PATENTSCOPE は公式の検索基盤として使えます。Qwen3.7-Max は text-only モデルなので、図面、オフィスアクション PDF、契約書スキャンは先に OCR/VL で文字化してから渡す設計が基本になります。[39]

| ユースケース | 推奨ワークフロー | 使う機能 | 例示プロンプト | 人手ゲート |
|---|---|---|---|---|
| 特許ドラフティング | 発明メモを JSON 化 → 関連先行技術取得 → 請求項骨子 → 明細書ドラフト → サポート要件/用語統一チェック | KB、web/extractor、code interpreter、Patent DB | `役割: JP/US/PCT 実務の特許ドラフター。入力: 発明開示、検索結果、実施例。出力: 独立請求項3案、従属請求項、サポート不足箇所、追加ヒアリング質問。制約: 取得済み根拠以外は推測しない。各請求項要素に根拠IDを付ける。` | 弁理士/特許代理人が必ず最終起案。 |
| 先行技術調査 | 多言語キーワード展開 → IPC/CPC 候補生成 → J-PlatPat/Espacenet/PATENTSCOPE/USPTO 横断 → Top-N マトリクス化 | web search、web extractor、function calling、code interpreter、Patent DB | `対象発明について JP/EN/CN の検索式を作成し、CPC/IPC 候補、検索ブロック、除外語、上位20件の要約、請求項要素との対応、残る新規性論点を一覧化せよ。` | 調査報告の結論は人が確定。 |
| クレーム解析 / FTO 一次評価 | 対象製品仕様・BOM・フローチャートを構造化 → 請求項要素分解 → literal / possible DOE / not met / unknown を付与 | OCR、KB、code interpreter、function calling | `各クレーム要素を表形式で分解し、対象製品との対応を literal / possible DOE / not met / unknown で評価。証拠ドキュメントID、頁、理由、追加で必要な証拠を併記せよ。結論は暫定と明示。` | FTO 結論は必ず弁護士/弁理士レビュー。 |
| パテントランドスケープ | データ収集 → assignee 正規化 → CPC/時系列/国別集計 → クラスタリング → 白地図/ヒートマップ | Patent DB、KB、code interpreter | `過去5年の出願を assignee, CPC, jurisdiction, legal status で集計し、上位出願人、技術クラスター、急伸テーマ、空白領域、買収候補/提携候補の仮説を作れ。` | 経営判断資料は出典監査後に利用。 |
| 契約レビュー | OCR/テキスト抽出 → 条項抽出 → プレイブック比較 → リスク採点 → 修正文案 | OCR、KB、structured JSON、code interpreter | `NDA/JDA/ライセンス契約を条項別に抽出し、機密保持、成果帰属、改良発明、監査、輸出管理、準拠法、紛争解決をプレイブックと比較。差分、頁番号、修正文案を示せ。` | 法務が承認するまで採用禁止。 |
| ドケッティング自動化 | OA/メール/PDF から期限・案件番号・国・イベント抽出 → JSON → docket system 反映前レビュー | OCR、structured JSON、function calling | `オフィスアクション/メールから期限、管轄、案件番号、依頼者、必要アクションを抽出し、confidence を付けた JSON で出力。送信・登録は行わず、人の確認待ちにする。` | 自動登録ではなく承認ワークフロー必須。 |

実装上のコツは、**根拠のある出力形式を先に決める**ことです。特に知財実務では、自由文より「document_id / page / paragraph / evidence_quote / confidence」を必須項目にした JSON を先に出さ

せ、その後に narrative を生成させる方が事故が少ないです。Model Studio には structured JSON 出力ガイドと function calling があるため、ドケッティング、請求項マトリクス、契約差分抽出は相性が良いです。[40]

### 日本・EU・米国・中国での最低限のコンプライアンス観点

| 法域 | 主な論点 | 実務上の含意 |
|---|---|---|
| 日本 | PPC は 2023年に生成AI利用時の個人情報取扱いへの注意喚起を公表し、APPI では外国第三者提供に関する規律も強化されています。[41] | 未公開出願、発明者履歴書、従業員メール、相手方契約書の個人情報は原則マスキングし、越境移転の法的根拠と記録を整備すべきです。 |
| EU | GDPR は個人データ処理全般に適用され、EU AI Act は AI に関する調和的ルールを定めています。[42] | EU 居住者データを扱う場合は、EU 域内地域を優先し、DPA・役割分担・人の監督を明記した設計が必要です。 |
| 米国 | USPTO は AI-assisted inventions について「人の significant contribution」があれば特許性を否定しないとし、USPTO 倫理規則は実務家の confidentiality を要求しています。[43] | 発明者認定は AI 使用の有無ではなく人間の貢献記録が重要で、弁護士・弁理士特権情報の外部モデル送信には契約面を含め慎重さが必要です。 |
| 中国 | CAC の「生成式人工智能服务管理暂行办法」と、2025年施行の network data security 規則は、生成AI と個人情報/越境移転を明確に規制します。[44] | 中国本土起点の個人情報や重要データを扱う案件では、北京/中国本土側処理や越境移転審査の要否を個別確認すべきです。 |

加えて、Alibaba は Model Studio について「データは学習に使わない」「入力・出力は Member Content であり、処理は顧客の目的のため・顧客に代わって行う」と説明していますが、**法令順守責任そのものは顧客側に残る**と規定しています。つまり、ベンダーの“not used for training”だけでは足りず、案件側のデータ分類・転送制御・保管期間・権限管理まで自社責任で設計すべきです。[45]

## 限界とリスク

最大の限界は、**Qwen3.7-Max の headline performance が強くても、法律・知財実務に直結する独立検証はまだ乏しい**ことです。公式記事には law を含む internal productivity benchmark が出てきますが、外部再現可能な patent/FTO/legal benchmark は見当たりません。さらに、公開された強いスコアの一部は内部 benchmark や複雑な scaffold 条件に依存しています。[1]

二つ目の限界は、**幻覚と冗長性**です。Artificial Analysis は Qwen3.7-Max の Intelligence Index 評価で 97M 出力トークンを観測し、非常に verbose と評しました。コミュニティでも hallucination、過剰編集、想定外の修正提案が繰り返し指摘されています。Qwen 系全般では tool-call hallucination や invalid tool response の issue も見られます。知財実務では、これは「実在しない引用文献」「実在しないクレーム対応関係」「契約条項の取り違え」に直結し得ます。[46]

三つ目は、**長時間エージェント固有の攻撃面**です。これはベンダーが明示的に認めているわけではありませんが、Model Studio が web search、web extractor、knowledge base search、function calling を自然に統合できる以上、外部 Web/PDF/ナレッジに混入した命令文や汚染データが、推論やツール実行に影響する classic prompt-injection / retrieval poisoning の面を持つ、と推論するのが自然です。したがって、知財文書や Web ページは「信頼できるソース」「OCR 正規化」「HTML/PDF サニタイズ」「ツール呼び出しは

allow-list」「system prompt と取得コンテンツの分離」で扱うべきです。これはツール構成から導かれる実装上のリスク評価です。 [47]

| 失敗モード | 典型症状 | 抑止策 |
|---|---|---|
| 先行技術・法令・契約条項の幻覚 | 実在しない引用、頁番号のずれ、条文名の混同 | “根拠ID付き JSON → narrative” の二段出力、引用原文の再検証。 |
| Tool-call 不良 | 途中で malformed JSON、誤引数、存在しない操作 | strict schema、retries、tool_choice 制御、実行前バリデーション。 |
| 長時間 drift | 初期目的から逸脱、過剰編集、止まりどころ不明 | verifier、非改善ターン停止、diff 承認、予算と時間の上限。 |
| Prompt injection / poisoned retrieval | 取得文書中の命令でモデル方針が崩れる | retrieval content を untrusted 扱いし、ツール実行権限を最小化。 |
| 連携面の不整合 | 401/404、モデル一覧不一致、provider format 差 | 本番前に接続検証、fallback provider、SLA とサポート窓口確認。 |

権利帰属については、Alibaba の国際 Model Studio Terms が比較的明快です。入力と出力は Member Content に含まれ、Alibaba は Output の知的財産権を主張しない一方、Alibaba Base IPR は Alibaba 側に残り、出力利用に伴う紛争責任は顧客側が負います。また、Model Studio やその出力を競合サービスの訓練・開発に使うことは禁止されています。したがって、「出力を使ってよい」ことと、「その出力が第三者権利を侵害しない」ことは全く別問題です。特に請求項や契約条項の転用では、元文書との類似性監査を残すべきです。 [48]

## 導入形態とコスト運用

Qwen3.7-Max そのものは、少なくとも現時点の公式説明では **proprietary / managed service 前提**と見るべきです。Qwen3.7 公式記事は Model Studio 経由の API 提供を示していますが、Alibaba の一般 API ドキュメントや pricing ドキュメントはなお `qwen3-max` / `qwen3-max-2026-01-23` を中心に書かれており、3.7 系の一般公開ドキュメント整備はやや遅れ気味です。つまり、**実モデルの提供は進んでいるが、ドキュメント整備は追随中**という状態です。 [49]

| 導入形態 | 位置づけ | 強み | 弱み | 知財用途での向き |
|---|---|---|---|---|
| Alibaba Cloud 国際/グローバル managed API | 実質本命。国際配備では endpoint/data は Singapore、計算資源は中国本土を除くグローバル動的配置。 [50] | 最新 proprietary をそのまま使える。検索・抽出・コード実行も統合可能。 [51] | 越境データ・外部委託・ログ管理の審査が必要。 | 日本企業の一般案件向き。 |
| EU managed API | endpoint/data が Frankfurt、計算資源も EU 限定。 [52] | GDPR 対応を組みやすい。 | 価格・利用可能モデルの確認が必要。 | EU 個人データ混在案件向き。 |

| 導入形態 | 位置づけ | 強み | 弱み | 知財用途での向き |
|---|---|---|---|---|
| 中国本土 managed API | endpoint/data/storage は Beijing、計算資源も中国本土。 [53] | 中国案件のデータローカライゼーションと親和性。 | 越境ワークフローは別審査。 | 中国起点案件向き。 |
| open-weight 代替の self-host | Qwen3.5-397B-A17B や Qwen3-235B-A22B など“近い代替”のみ。 [54] | オンプレ/閉域を作れる。 | **Qwen3.7-Max 本体ではない**。能力差と運用負荷が大きい。 | 機密最優先時の代替案。 |

## コストの見方

公式英語 pricing ページで直に確認できるのはまだ `qwen3-max` 系で、中国本土では入力 0–32K が $0.359 / 1M、出力 $1.434 / 1M、128K 超では入力 $1.004、出力 $4.014 まで上がります。バッチは 50% オフ、context cache 割引もあります。したがって、**長いプロンプトを何度も回す設計ではキャッシュ戦略がかなり重要**です。 [7]

一方、Qwen3.7-Max 自体の価格は、2026年6月時点では英語公式価格表での露出がまだ一貫しておらず、独立ソースでは $2.50 / 1M input、$7.50 / 1M output と観測されています。OpenCode の実運用観測では平均セッションコストが $1.40、平均 2M tokens/session、98% cache ratio という値も出ています。したがって、**公式 3.7 専用価格が未整理の間は、「AA/OpenCode の実測を暫定目安」「正式稟議は契約価格確認後」**が妥当です。 [55]

また、エージェント運用ではモデルトークン料金だけでなく、**検索系ツールの別料金**も見落とせません。Model Studio の web search ドキュメントでは search strategy fee が別建てで、国際配備では agent strategy が 1,000 call あたり $10.00 とされています。検索を多用する prior-art search や competitive intelligence では、この部分が無視できません。 [56]

## インフラ見積もり

Qwen3.7-Max 本体は self-host 不可と考えるのが現実的です。もしオンプレが必須なら、代替候補は open-weight の Qwen3.5-397B-A17B や Qwen3-235B-A22B になります。単純な重み保存下限で計算すると、397B は BF16 で約 794GB、INT8 で約 397GB、235B は BF16 で約 470GB、INT8 で約 235GB で、さらに KV cache・並列化・ランタイム overhead を足す必要があります。したがって、**“1 台の汎用 GPU ワークステーションで代替する”発想は非現実的**で、少なくともマルチ GPU / マルチノード構成が前提です。これは Qwen3.7-Max と同等という意味ではなく、あくまで閉域代替の下限感です。 [54]

## 長時間エージェント運用の実務論

長時間の知財エージェントでは、Responses API の `previous_response_id` による会話履歴の簡略管理、explicit cache による固定プロンプト/案件履歴の再利用、外部 state store による証拠 ID 管理、そして telemetry/logging が重要です。Alibaba は explicit cache について「キャッシュ作成は標準入力価格の 125%、ヒットは 10%」としており、同一 prefix 再利用でコスト・レイテンシ削減を狙えます。また Model telemetry は 30 日保持、推論ログを SLS に流せます。IP 実務では、このログを「誰が何の案件でどの外部文献を参照し、どの期限抽出をしたか」の監査証跡に使う設計が有効です。 [57]

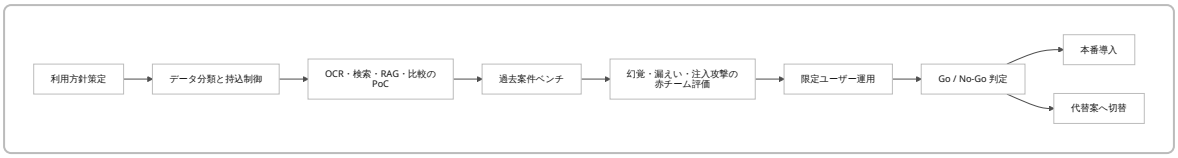
# 推奨事項とパイロット計画

結論から言うと、Qwen3.7-Max は **知財実務の“収集・比較・構造化・下書き”には有望、最終判断の代替には不適**です。特に日本語・中国語・英語を跨ぐ先行技術調査、請求項マトリクス作成、landscape 分析、契約差分レビュー、期限抽出のような作業では、Qwen3.7-Max の「長い文脈 + ツール連携 + コード実行 + 多段オーケストレーション」が効きます。逆に、FTO 最終意見、発明者認定、最終請求項確定、訴訟ホールド下文書の丸投げは避けるべきです。 [58]

| タスク | 適性 | 推奨度 | 条件 |
|---|---|---|---|
| 先行技術調査の検索式生成・初期スクリーニング | 高い | 採用推奨 | Patent DB と人手レビューを必須化。 |
| パテントランドスケープ | 高い | 採用推奨 | code interpreter で集計・可視化、出典監査を残す。 |
| クレームチャート一次作成 | 中〜高 | 条件付き採用 | 根拠 ID と“不明”ラベルを強制。 |
| FTO 一次整理 | 中 | 条件付き採用 | 結論ではなく論点整理用途に限定。 |
| 特許ドラフティング下書き | 中〜高 | 条件付き採用 | 最終請求項は弁理士/代理人が起案。 |
| 契約レビュー一次抽出 | 中〜高 | 採用推奨 | プレイブック差分＋頁番号付きに限定。 |
| ドケッティング自動化 | 中〜高 | 採用推奨 | 自動登録ではなく承認ワークフロー前提。 |

### パイロットの最小構成

| 期間 | マイルストーン | 成果物 | 成功指標 | 主担当 |
|---|---|---|---|---|
| 週初 | ガバナンス確定 | データ分類、案件持込ルール、利用禁止データ定義 | 機密区分マトリクス承認 | 情報セキュリティ、法務、知財部 |
| 前半 | 技術プロトタイプ | OCR→検索→RAG→比較→JSON 出力の一連フロー | 主要 6 ユースケースが再現 | AI/LLM エンジニア、知財アナリスト |
| 中盤 | 過去案件でのベンチ試験 | 既存調査票・請求項表・契約レビューとの比較 | 先行技術調査時間 30–40%短縮、期限抽出 recall 99% 以上、根拠付き出力率 95% 以上 | 知財実務担当、PM |
| 後半 | リスク・赤チーム評価 | prompt injection / data leakage / hallucination 試験報告 | 重大事故ゼロ、回避策定義済み | セキュリティ、法務、AI エンジニア |
| 終盤 | 限定本番 | 1–2 チームでの実案件運用 | ユーザー満足度、コスト/案件、再作業率、誤抽出率 | 部門責任者、運用担当 |




## 調達と法務レビューのチェックリスト

| 項目 | 確認すべきこと | 受入基準 |
|---|---|---|
| データ学習利用 | 顧客データを学習に使わないか | 契約・DPA で“not used for training”を確認。 45 |
| 出力の権利と責任 | Output の権利帰属、紛争責任、競合訓練禁止 | Output 不主張、ただし顧客責任と Alibaba Base IPR carve-out を理解。 48 |
| 地域拘束 | SG/EU/北京のどこに endpoint / data / compute が置かれるか | EU 案件は Frankfurt、中国案件は北京など案件別に地域固定。 59 |
| ログ保持 | telemetry/log retention、SLS 出力、削除ポリシー | 30日保持・SLS 連携・削除運用を文書化。 60 |
| 監査資料 | Trust Center レポート、サブプロセッサー一覧 | NDA 後に報告書と sub-processor list を取得。 61 |
| レート制限 | account/workspace/API key 跨ぎの制限 | 高負荷試験で上限確認、failover 設計。 62 |
| ツール課金 | search/extractor/tool calling の別料金 | prior-art など検索多用案件で別立て見積り。 63 |
| OCR 分業 | スキャン文書を Max に直接渡すか、OCR/VL を前段に置くか | 図面・PDF は OCR/VL 前処理を標準化。 64 |
| 長時間セッション | キャッシュ、状態管理、停止条件 | explicit cache と verifier 停止条件を必須化。 65 |
| 法域別要件 | 日本/EU/US/中国の個人情報・発明者認定・越境移転 | 各案件類型ごとに法務テンプレートを作成。 66 |

総合評価としては、**Qwen3.7-Max は「知財チームの仕事を自動化し切るモデル」ではなく、「知財チームが使う多言語エージェント基盤としてかなり有力」**です。導入するなら、いきなり全面展開ではなく、まずは先行技術調査・landscaping・契約一次レビュー・期限抽出の 4 領域で限定導入し、根拠付き出力率、再作業率、案件単価、漏えい事故ゼロを指標に 6〜8 週間で判定するのが最も堅実です。公式主張は強く、独立評価も好調ですが、知財の最終結論を任せるには、まだ「人が責任を持つ前提」で使うべき段階です。 67

---

1 17 18 37 49 58 67 https://www.alibabacloud.com/blog/qwen3-7-the-agent-frontier_603154
https://www.alibabacloud.com/blog/qwen3-7-the-agent-frontier_603154

2 20 35 46 55 https://artificialanalysis.ai/models/qwen3-7-max
https://artificialanalysis.ai/models/qwen3-7-max

3 https://www.alibabacloud.com/help/en/model-studio/qwen-code-interpreter
https://www.alibabacloud.com/help/en/model-studio/qwen-code-interpreter

4 5 https://qwen.ai/blog?id=qwen3.7
https://qwen.ai/blog?id=qwen3.7

6 52 https://www.alibabacloud.com/help/en/model-studio/models
https://www.alibabacloud.com/help/en/model-studio/models

7 50 53 59 https://www.alibabacloud.com/help/en/model-studio/model-pricing
https://www.alibabacloud.com/help/en/model-studio/model-pricing

8 56 63 https://www.alibabacloud.com/help/en/model-studio/web-search
https://www.alibabacloud.com/help/en/model-studio/web-search

9 13 https://www.alibabacloud.com/blog/qwen3-max-just-scale-it_602621
https://www.alibabacloud.com/blog/qwen3-max-just-scale-it_602621

10 https://www.alibabacloud.com/blog/601786
https://www.alibabacloud.com/blog/601786

11 https://arxiv.org/abs/2505.09388
https://arxiv.org/abs/2505.09388

12 https://www.alibabacloud.com/blog/alibaba-introduces-qwen3-setting-new-benchmark-in-open-source-ai-with-hybrid-reasoning_602192
https://www.alibabacloud.com/blog/alibaba-introduces-qwen3-setting-new-benchmark-in-open-source-ai-with-hybrid-reasoning_602192

14 15 54 https://www.alibabacloud.com/blog/602894
https://www.alibabacloud.com/blog/602894

16 https://www.alibabacloud.com/blog/alibaba-unveils-qwen3-6-plus-to-accelerate-agentic-ai-deployment-for-enterprises-and-alibaba%E2%80%99s-ai-applications_603005
https://www.alibabacloud.com/blog/alibaba-unveils-qwen3-6-plus-to-accelerate-agentic-ai-deployment-for-enterprises-and-alibaba%E2%80%99s-ai-applications_603005

19 https://artificialanalysis.ai/
https://artificialanalysis.ai/

21 26 https://artificialanalysis.ai/models/gpt-4-1
https://artificialanalysis.ai/models/gpt-4-1

22 https://artificialanalysis.ai/models/gpt-4o-chatgpt-03-25
https://artificialanalysis.ai/models/gpt-4o-chatgpt-03-25

23 https://artificialanalysis.ai/models/gpt-4o-mini
https://artificialanalysis.ai/models/gpt-4o-mini

24 https://artificialanalysis.ai/models/llama-4-maverick
https://artificialanalysis.ai/models/llama-4-maverick

25 https://artificialanalysis.ai/models/llama-3-3-instruct-70b
https://artificialanalysis.ai/models/llama-3-3-instruct-70b

27 https://arena.ai/leaderboard/text
https://arena.ai/leaderboard/text

28 https://benchlm.ai/models/qwen3-7-max
https://benchlm.ai/models/qwen3-7-max

29 https://www.tbench.ai/leaderboard/terminal-bench/2.0
https://www.tbench.ai/leaderboard/terminal-bench/2.0

30 https://gigazine.net/news/20260521-qwen-3-7/
https://gigazine.net/news/20260521-qwen-3-7/

31 34 https://www.v2ex.com/t/1214878
https://www.v2ex.com/t/1214878

32 https://www.reddit.com/r/LocalLLaMA/comments/1tie6gy/qwen37_max_scored_by_artificial_analysis_27b35b/
https://www.reddit.com/r/LocalLLaMA/comments/1tie6gy/qwen37_max_scored_by_artificial_analysis_27b35b/

33 https://github.com/anomalyco/opencode/issues/29558
https://github.com/anomalyco/opencode/issues/29558

36 https://www.reddit.com/r/Qwen_AI/comments/1tmj4s4/qwen37_max_is_superb_but_hallucinates_too_much/
https://www.reddit.com/r/Qwen_AI/comments/1tmj4s4/qwen37_max_is_superb_but_hallucinates_too_much/

38 https://opencode.ai/data/qwen/qwen3-7-max
https://opencode.ai/data/qwen/qwen3-7-max

39 51 https://www.alibabacloud.com/help/en/model-studio/qwen-api-via-openai-responses
https://www.alibabacloud.com/help/en/model-studio/qwen-api-via-openai-responses

40 https://www.alibabacloud.com/help/en/model-studio/qwen-structured-output
https://www.alibabacloud.com/help/en/model-studio/qwen-structured-output

41 66 https://www.ppc.go.jp/files/pdf/230602_kouhou_houdou.pdf
https://www.ppc.go.jp/files/pdf/230602_kouhou_houdou.pdf

42 https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng
https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

43 https://www.uspto.gov/subscription-center/2024/uspto-issues-inventorship-guidance-and-examples-ai-assisted-inventions
https://www.uspto.gov/subscription-center/2024/uspto-issues-inventorship-guidance-and-examples-ai-assisted-inventions

44 https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

45 https://www.alibabacloud.com/help/en/model-studio/what-is-model-studio
https://www.alibabacloud.com/help/en/model-studio/what-is-model-studio

47 https://www.alibabacloud.com/help/en/model-studio/rag-knowledge-base
https://www.alibabacloud.com/help/en/model-studio/rag-knowledge-base

48 https://www.alibabacloud.com/help/en/legal/latest/alibaba-cloud-international-website-product-terms-of-service-v-3-8-0
https://www.alibabacloud.com/help/en/legal/latest/alibaba-cloud-international-website-product-terms-of-service-v-3-8-0

57 https://www.alibabacloud.com/help/en/model-studio/compatibility-with-openai-responses-api
https://www.alibabacloud.com/help/en/model-studio/compatibility-with-openai-responses-api

60 https://www.alibabacloud.com/help/en/model-studio/model-telemetry/
https://www.alibabacloud.com/help/en/model-studio/model-telemetry/

61 https://www.alibabacloud.com/en/trust-center/security-compliance-practice?_p_lc=1
https://www.alibabacloud.com/en/trust-center/security-compliance-practice?_p_lc=1

62 https://www.alibabacloud.com/help/en/model-studio/rate-limit
https://www.alibabacloud.com/help/en/model-studio/rate-limit

64 https://www.alibabacloud.com/help/en/model-studio/qwen-vl-ocr
https://www.alibabacloud.com/help/en/model-studio/qwen-vl-ocr

65 https://www.alibabacloud.com/help/en/model-studio/explicit-cache-best-practice
https://www.alibabacloud.com/help/en/model-studio/explicit-cache-best-practice