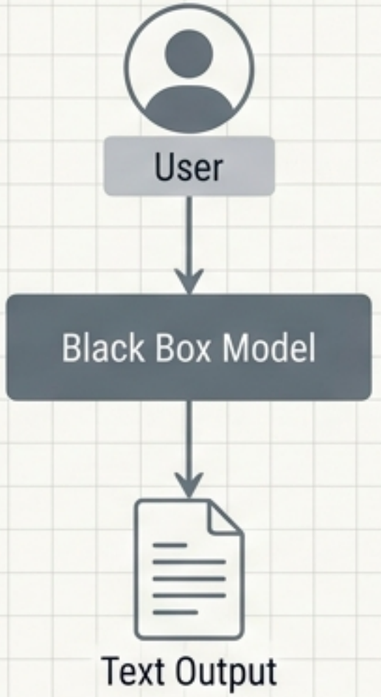
Qwen3.7-Max 徹底解剖：知財・法務実務における「自律型エージェント」の可能性と実装要件

誇大広告を排除した、専門家向けアーキテクチャ評価と導入プレイブック

[CONFIDENTIAL / FOR INTERNAL STRATEGY REVIEW ONLY]

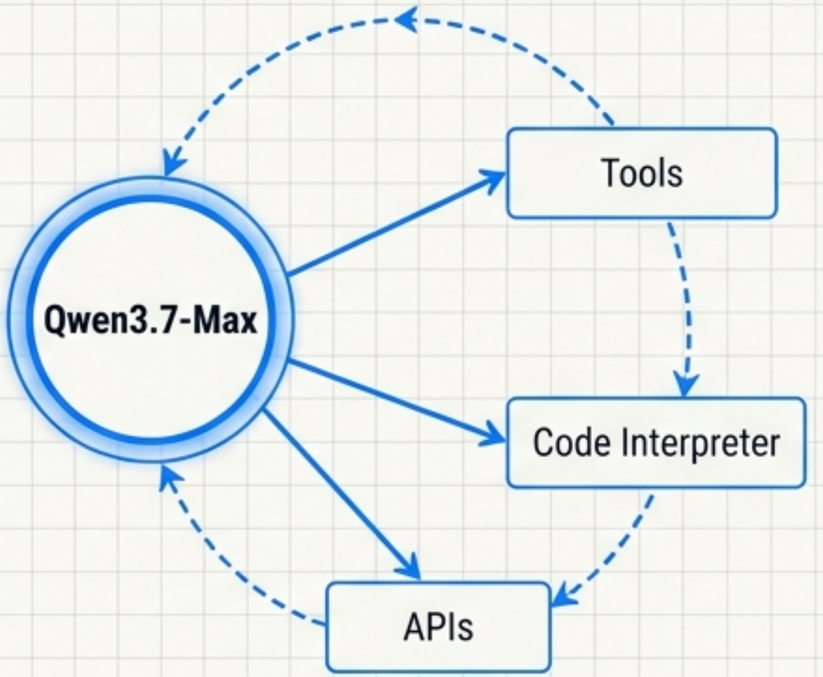
THE ORCHESTRATION CANVAS: MODEL COMPARISON

従来のLLM（対話型）



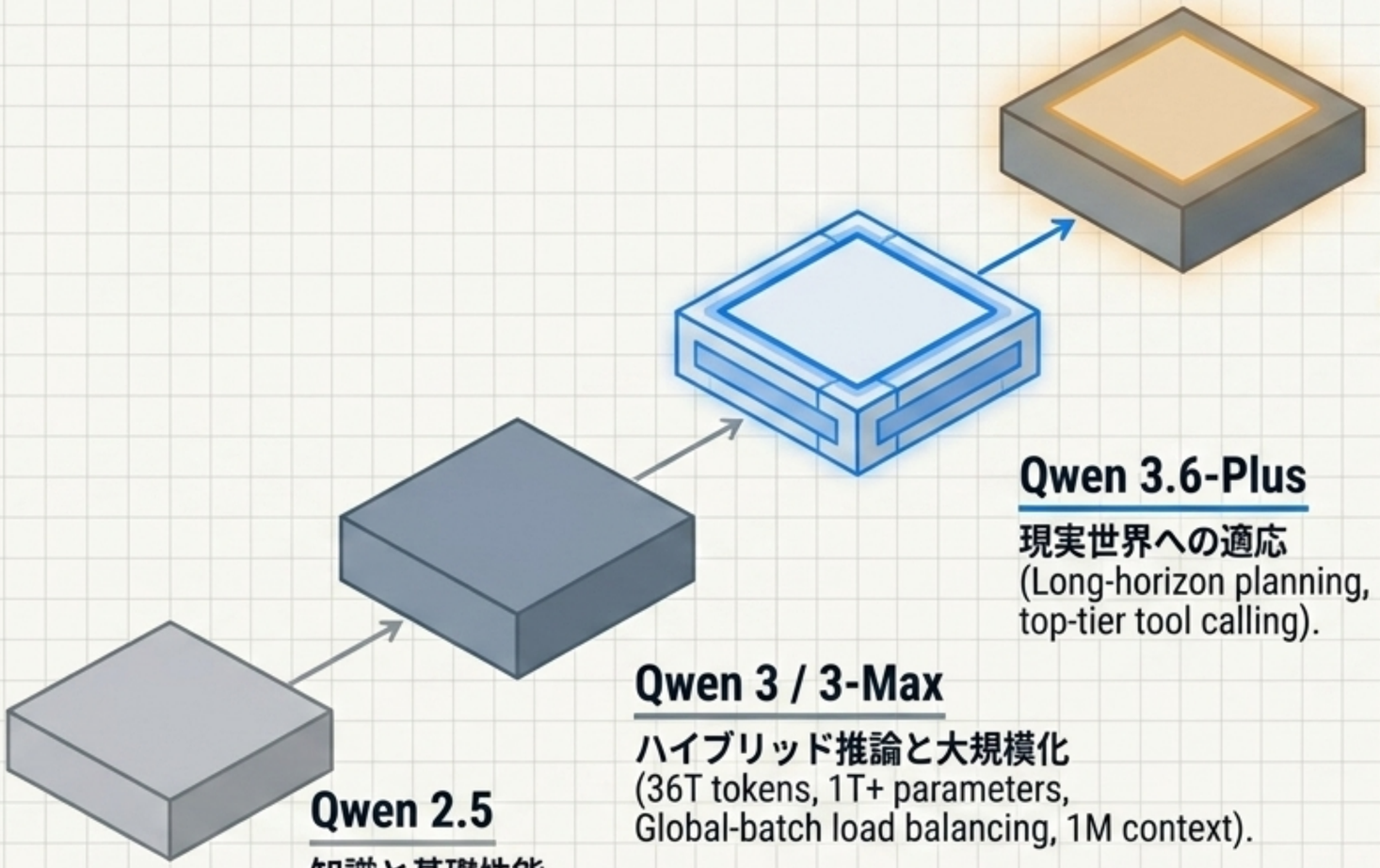
- 単一ターンの対話に特化
- モデル内部の重み（知識）に完全に依存
- 人間がプロンプトで実行を主導する

Qwen3.7-Max（オーケストレーション層）



- **自律実行: 35時間の完全自律カーネル最適化を実証。**
- **ツール連携: 1,158回の自律的な外部ツール呼び出し (10倍の高速化)。**
- **適応力: Claude CodeやOpenClawなど異なる環境を跨いで汎化 (Cross-Harness Generalization)。**

THE ORCHESTRATION CANVAS: MODEL EVOLUTION TIMELINE



Qwen 2.5

知識と基礎性能
(18T tokens, 72B).

Qwen 3 / 3-Max

ハイブリッド推論と大規模化
(36T tokens, 1T+ parameters,
Global-batch load balancing, 1M context).

Qwen 3.6-Plus

現実世界への適応
(Long-horizon planning,
top-tier tool calling).

Qwen 3.7-Max

実運用エージェントへの到達
(Environment scaling,
Cross-harness RL,
継続的な自律最適化).

Key Takeaway: モデルの進化は「回答の賢さ」から「業務の完遂能力」へと明確にシフトしている。

独立評価による現在地：GPT-4.1 / GPT-4o を凌駕する推論力

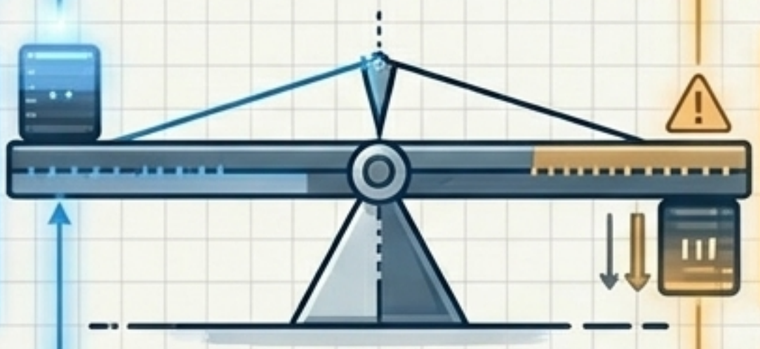
モデル	Artificial Analysis Index	Context	Speed / Notes
Qwen3.7-Max	**57**	**1M**	**↗ 112 tok/s**
GPT-4.1	26	1M	現行の主要比較対象（画像対応）
GPT-4o	19	130k	標準的なエンタープライズ基準
Llama 3.3 70B	14	128k	オープンウェイトとの顕著なギャップ

⚠ **Analyst Note:**
内部ベンチマーク（Terminal-Bench 69.7, SWE-Verified 80.4）では圧倒的なスコアを主張するが、独立機関での完全な再現検証は途上。実務導入の比較対象は旧来のGPT-4ではなく、GPT-4.1との直接比較が妥当。

トレードオフ：「高度な知能」と「極端な冗長性」のパラドックス

構造的強み (Strengths)

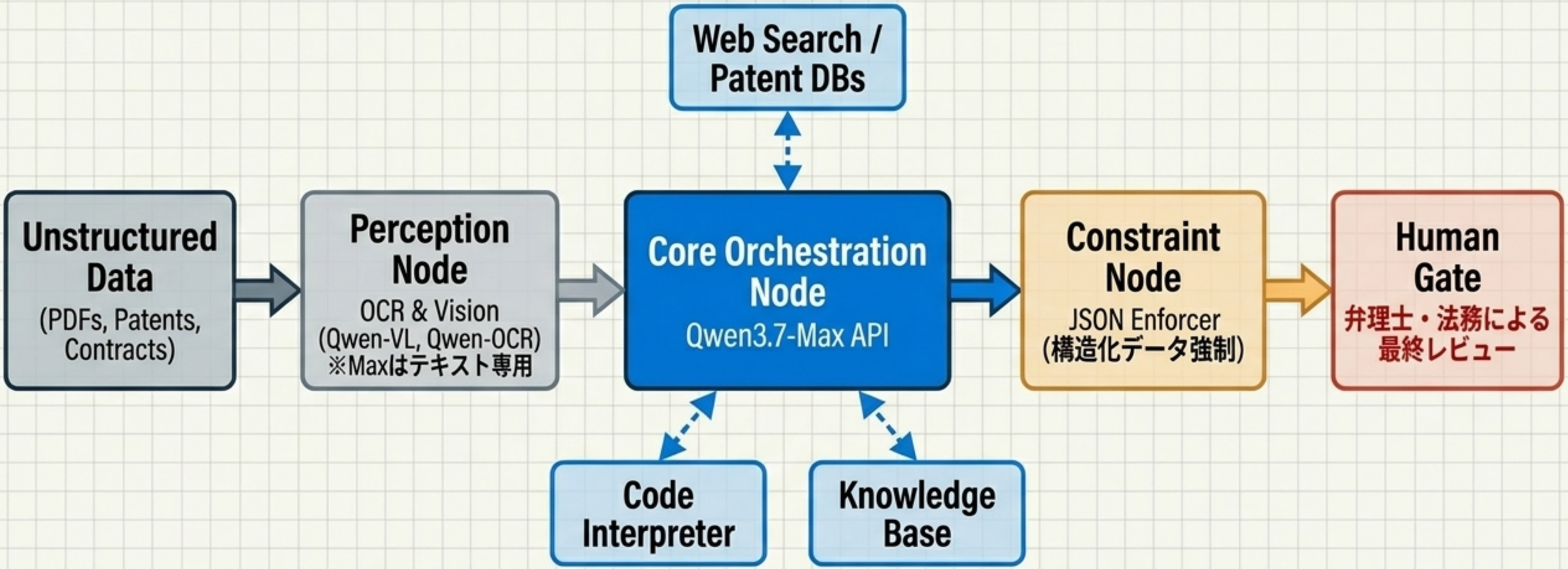
- 卓越した推論力 (Index 57)。
- 自律的なツール実行とコード生成。
- 国内（中国/アジア）における低遅延なAPI応答。



摩擦とコスト (Friction & Costs)

- 極端な冗長性 (Verbosity)：評価時に97Mの出力トークンを消費。長大な調査案件では予算枯渇のリスク。
- 幻覚と過剰編集 (Hallucination)：指示を逸脱して「独自の修正案」を加えようとする傾向。
- 統合の摩擦：サードパーティ連携時のAPIエラー（401/404）やプロバイダ互換性の未成熟さ。

知財実務における最適アーキテクチャ：「回答者」ではなく「指揮者」として配置する



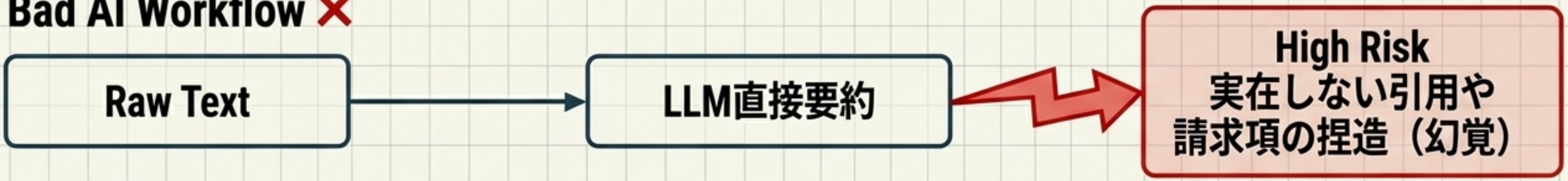
決してモデル単独に最終判断（FTO結論や発明者認定）を委ねない。
情報を「収集→構造化→比較」するパイプラインの中核として機能させる。

知財ユースケース診断マトリクス

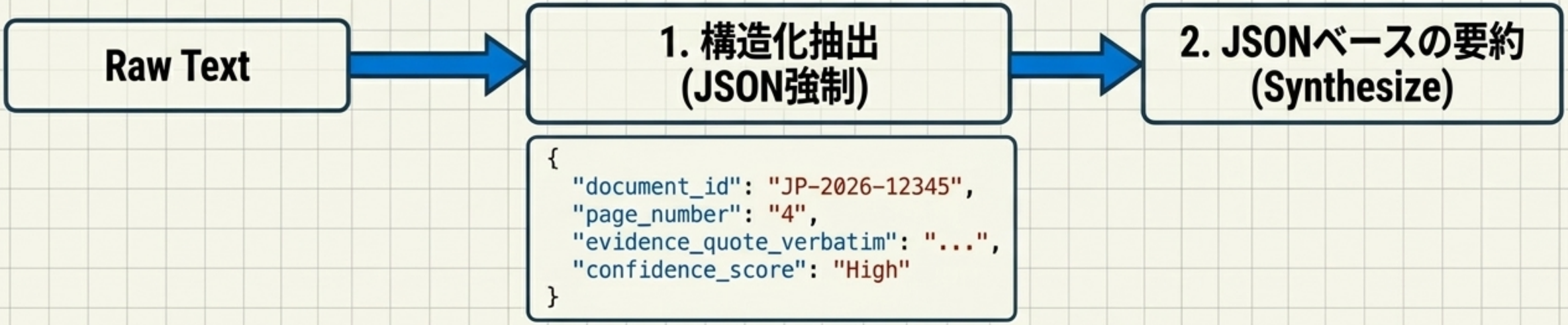
ユースケース	推奨機能	プロンプト制約	ヒューマンゲート
先行技術調査 (Prior Art)	Web search, Patent DB API	CPC/IPC候補とTop-Nマトリクス化を強制。	人間が最終的な新規性を判定。
クレーム解析 / FTO	OCR, Code Interpreter	請求項を要素分解。文言侵害/均等侵害を判定し、必ず根拠IDと頁番号を付与。	必須の弁護士・弁理士レビュー。
パテントランドスケープ	Code Interpreter, KB	出願人を正規化し、過去5年分のCPC/管轄国別クラスタリングを実行。	経営報告前にデータ出典の監査。
契約レビュー	OCR, Structured JSON	自社プレイブックと条項を比較。差分テキストと該当ページ番号を抽出。	法務部門の承認が必須。

実装の絶対法則：「物語 (Narrative)」の前に「構造化データ (JSON)」を強制する

Bad AI Workflow ✕



Golden Rule Workflow ✓



**Model StudioのStructured JSONとFunction Callingを活用し、
実在しない特許や条項の「捏造」をアーキテクチャレベルで封じ込める。**

グローバル・コンプライアンス要件：データ主権と地域別配備

日本 (Japan - APPI/PPC)

越境移転の法的根拠が必要。
未公開出願や発明者履歴書は
原則マスキング必須。

欧州 (EU - GDPR/AI Act)

Frankfurtエンドポイントの使
用を推奨。
DPA（データ処理契約）と
「人間の監督」設計が必須。

米国 (US - USPTO)

AI使用自体は特許性を否定し
ないが、「人間の著しい貢献
（Significant Contribution）」
の記録が必要。

中国 (China - CAC)

Beijingエンドポイント。ネッ
トワークデータセキュリティ
規則に準拠。重要データの
越境移転審査に注意。

WARNING: ベンダーの「学習に使用しない (Not used for training)」という規約だけでは不十分。
法令遵守とデータ保護の最終責任は顧客（導入企業）側に残る。

脅威モデリング：自律型エージェント特有のリスクと防御策

Failure Mode (障害モード)	Symptom (発生症状)	Engineering Defense (技術的防御策)
幻覚（Hallucination）	実在しない文献の引用、誤ったクレーム対応付け。	二段階出力アーキテクチャ（JSON ID抽出 → 要約生成）と原文の一致検証アルゴリズムの実装。
長時間のドリフト（Tool-Call Drift）	エージェントが初期目的から逸脱し、無限に編集を繰り返す、または無効なAPIを叩き続ける。	厳格なスキーマ検証、反復回数の上限設定、および予算/時間キャップのハードコード。
プロンプト・インジェクション（Retrieval Poisoning）	外部のPDFやWebデータに潜む隠し命令によってエージェントが乗っ取られる。	検索取得コンテンツを「Untrusted」として扱い、OCR/HTMLのサニタイズを実行。ツール実行のAllow-list化。

デプロイメント戦略と運用エコノミクス

Managed API (Global/EU/China)

- 主要な導入経路。Qwen3.7-MaxとModel Studioツールへのフルアクセス。
- コスト警告: モデル入出力費（$2.50/$7.50 per 1M）に加え、検索戦略費（1,000エージェントコールあたり$10.00）が別途発生。

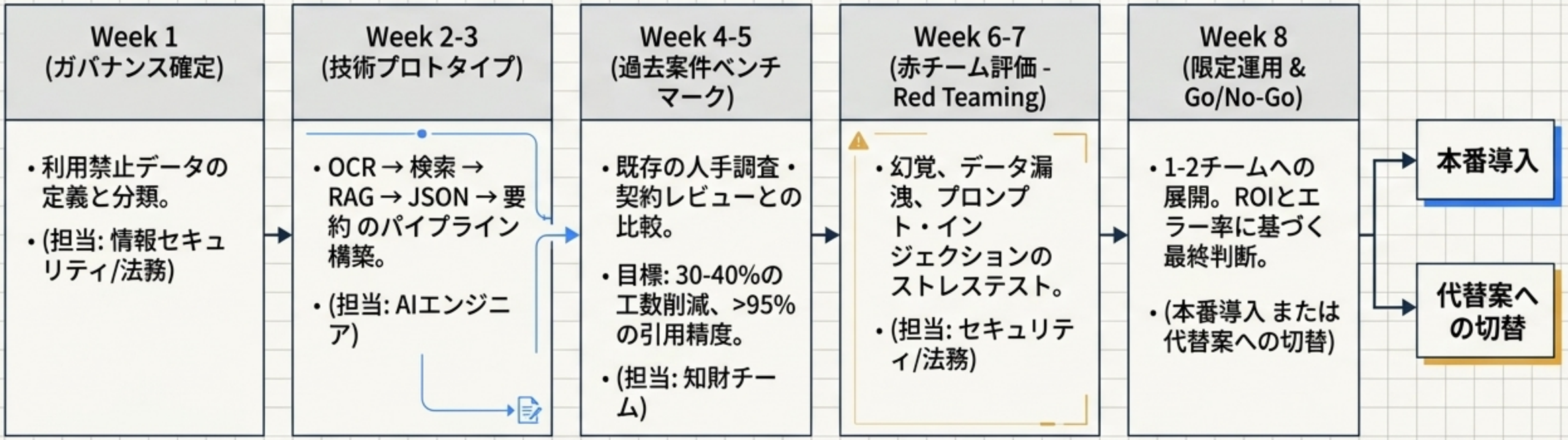
Context Caching Optimization

- 長大な知財プロンプトや過去案件履歴の入力には明示的キャッシュ（作成125%、ヒット10%のコスト比率）が必須。
- 予算枯渇を防ぐ生命線。

The Self-Host Reality

- Qwen3.7-Maxはプロプライエタリ。オンプレミス展開はオープンウェイト代替（Qwen3.5-397B等）に限られる。
- 397BモデルにはBF16で約800GBのVRAMが必要。安価で容易なフォールバック手段ではない。

最短稼働に向けた8週間パイロット・ロードマップ



最終見解：Qwen3.7-Maxに対する戦略的スタンス

01 業務適合性

「収集・比較・構造化・下書き」の自動化基盤として極めて有望。複雑な多言語先行技術調査やランドスケープ分析において、長文脈とツール連携の強みが活きる。

02 限界とガバナンス ⚠

FTOの最終意見、発明者認定、最終的な請求項の確定といった「責任を伴う最終判断」をモデルに委ねてはならない。

03 導入の条件

導入の成否は、モデルの性能以上に、企業側の「厳格なJSON制約の実装」「データ処理地域の統制」「キャッシュを活用したコスト管理」の設計力にかかっている。

「知財チームを代替する魔法の箱」ではなく、「知財チームの能力を拡張する高度なオーケストレーション・エンジン」として実装せよ。