

DeepSeek-V4-Pro-0813

Technical Due Diligence

ステルスリリースの検証と、エンタープライズ導入に向けた客観的評価

企業の意思決定者・AIエンジニア・セキュリティ責任者向けテクニカル・ブリーフィング

結論：APIは即時PoCへ移行、セルフホストは公式証明を待機せよ



【GO】API利用

Action: 既存APIでの即時PoC・検証

Why: モデルID (deepseek-v4-pro) の裏側は既に0813へ更新済み。コード変更なしで劇的なAgent性能向上と圧倒的な低コスト（出力 \$0.87/1M）の恩恵を享受可能。

Caveat: 本番導入にはカナリアリリースと自社Agentによる回帰試験（pass@1の検証）が必須。



【HOLD】セルフホスト

Action: 0813固有のセルフホスト運用

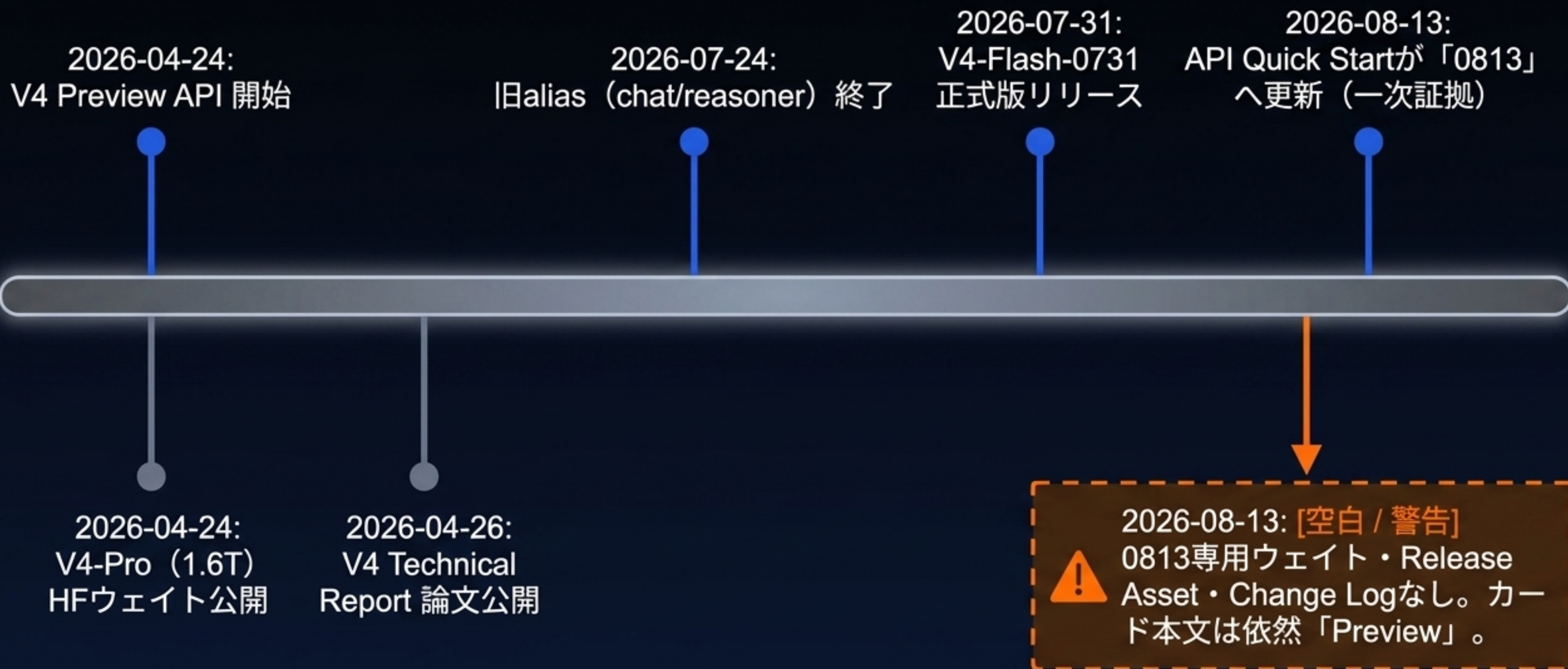
Why: 0813固有のリリースノート、ウェイト、ハッシュが未公開。既存のV4-Proウェイトが0813 APIと完全に同一（bit-identical）であるという公式の証明が存在しない。

Caveat: 機密データ処理のためセルフホストが必須の場合は、公式の0813アーティファクト公開を待つべき。

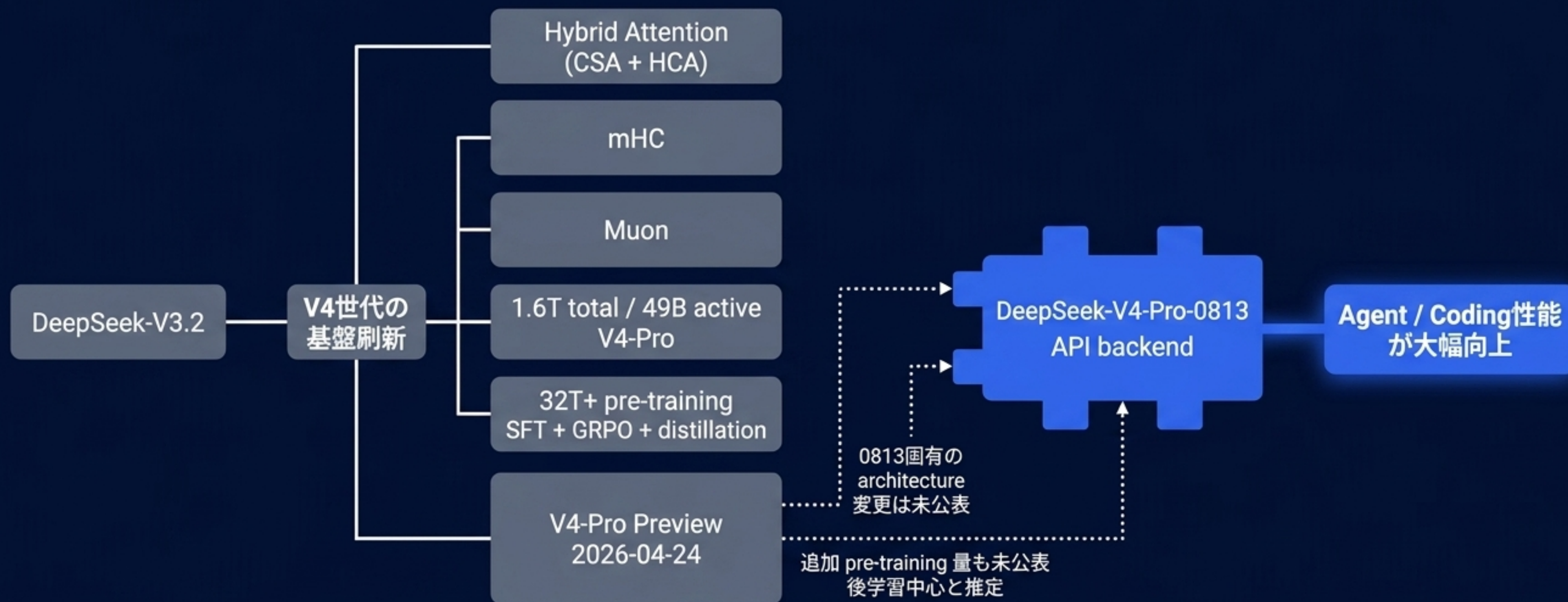
ファクトチェック・マトリクス：検証済み事実と未確認情報の分離

判定項目	ステータス	詳細
DeepSeek-V4-Pro-0813 APIの稼働	● 確認済み	公式APIドキュメントおよび料金ページで明示。
Agent性能の大幅向上	● 確認済み	コミュニティ報告および初期検証で強い裏付けあり。
「Previewから正式版へ移行」の意図	■ 強く支持	公式明言はないが、機能追加 (Responses API等) から推認。
1.6T / 49B / 1M 基本仕様	■ V4として継続	アーキテクチャ変更の公表なし。
0813専用公式リリースノート	⚠ 未公開	最終更新は0731Flashに留まる。
0813固有のHFウェイト / 実行バイナリ	⚠ 未確認	既存の公開ウェイトとの同一性は未証明。
新しい事前学習データ	⚠ 公表なし	追加の事前学習に関する言及なし。

非対称リリース：進行するAPI、停滞するアーティファクト



アーキテクチャ・デルタ：基礎基盤の継続と「後学習」の推論



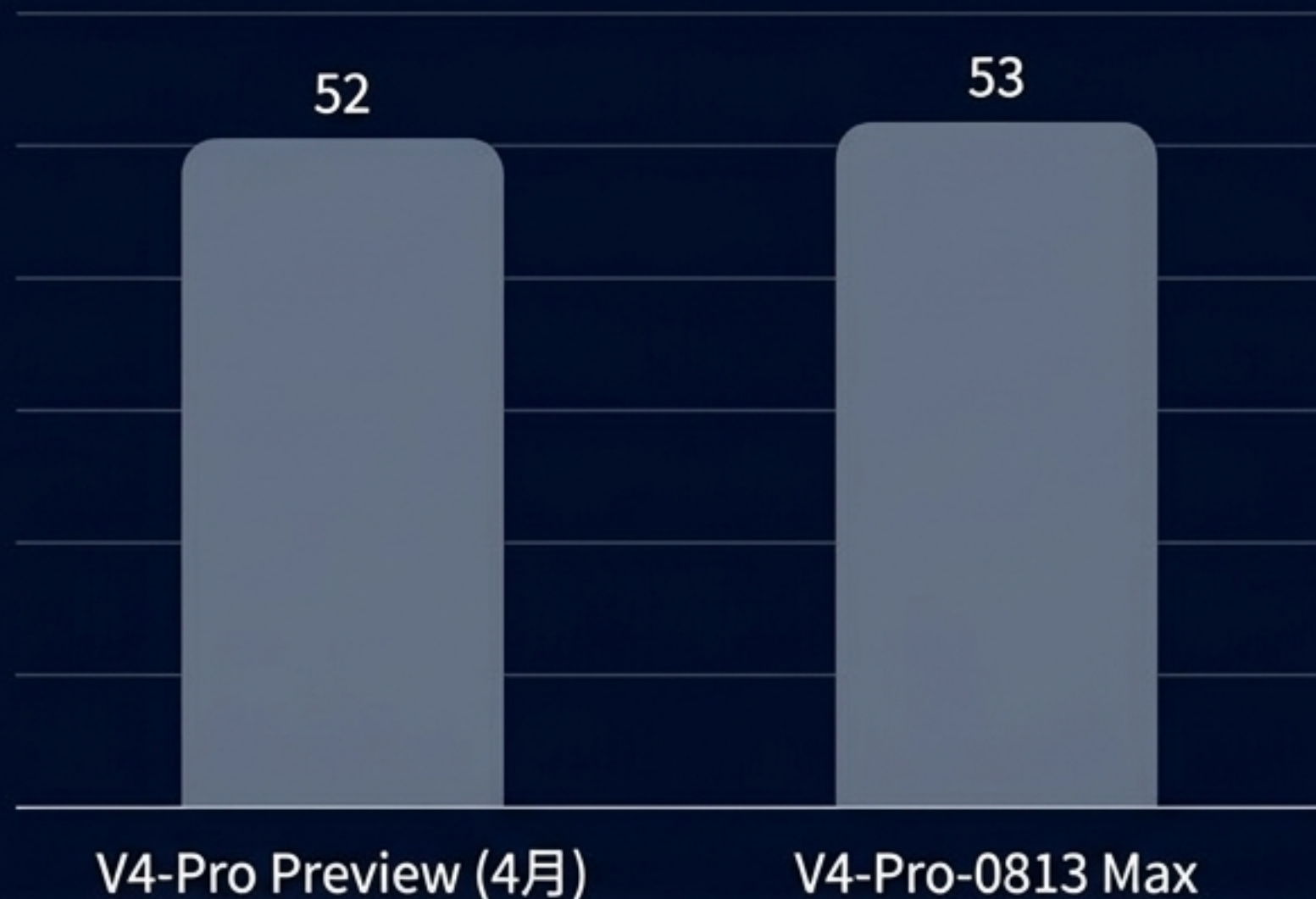
Analyst Insight: Flash-0731が「only re-post-trained」であったこと、および基本仕様の変更公表がないことから、0813は新アーキテクチャではなく「Agent特化の再後学習（Re-post-trained）モデル」とであると強く推定される。

診断テーブル：Preview vs 0813 のAPI仕様差分

機能 / 仕様	V4-Pro Preview	V4-Pro-0813
Context Length	1M	1M [継続]
Max Output	384K	384K [現行仕様明示]
Anthropic API	対応	対応 [継続]
FIM (Fill-in-the-Middle)	—	non-thinkingのみ対応 [明示]
Thinking Mode	対応	thinking / non-thinking選択可 (thinkingがデフォルト) [仕様明確化]
Responses API	未確認/Flashのみ	対応 (✓) [0813での実用上の大きな進化]

性能評価レーダー：汎用知能の微増とAgent能力の爆発的進化

汎用知能 - Artificial Analysis Index



基礎的な知能水準の「桁違いのジャンプ」は起きていない。（※TTFT 1.71sの高速性）

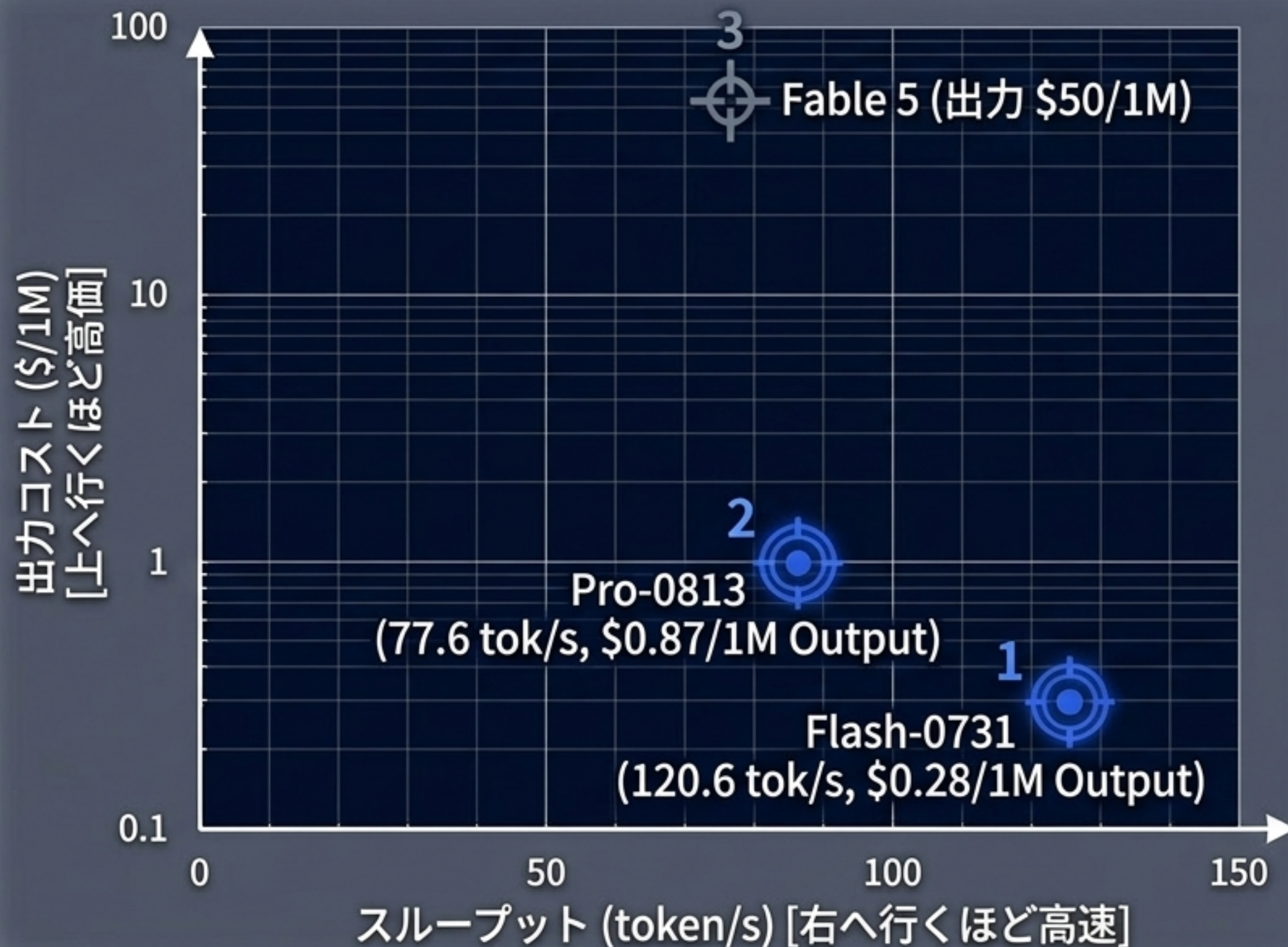
Agent/Coding能力 - ベンチマークデルタ



単純なチャット品質の調整では説明できない、ソフトウェアエンジニアリングと自律タスク実行能力への「全振り」最適化。

※独立検証済みの絶対値ではなく報告値として扱う必要あり。

ユニットエコノミクス：破格のTCOと最適ルーティング



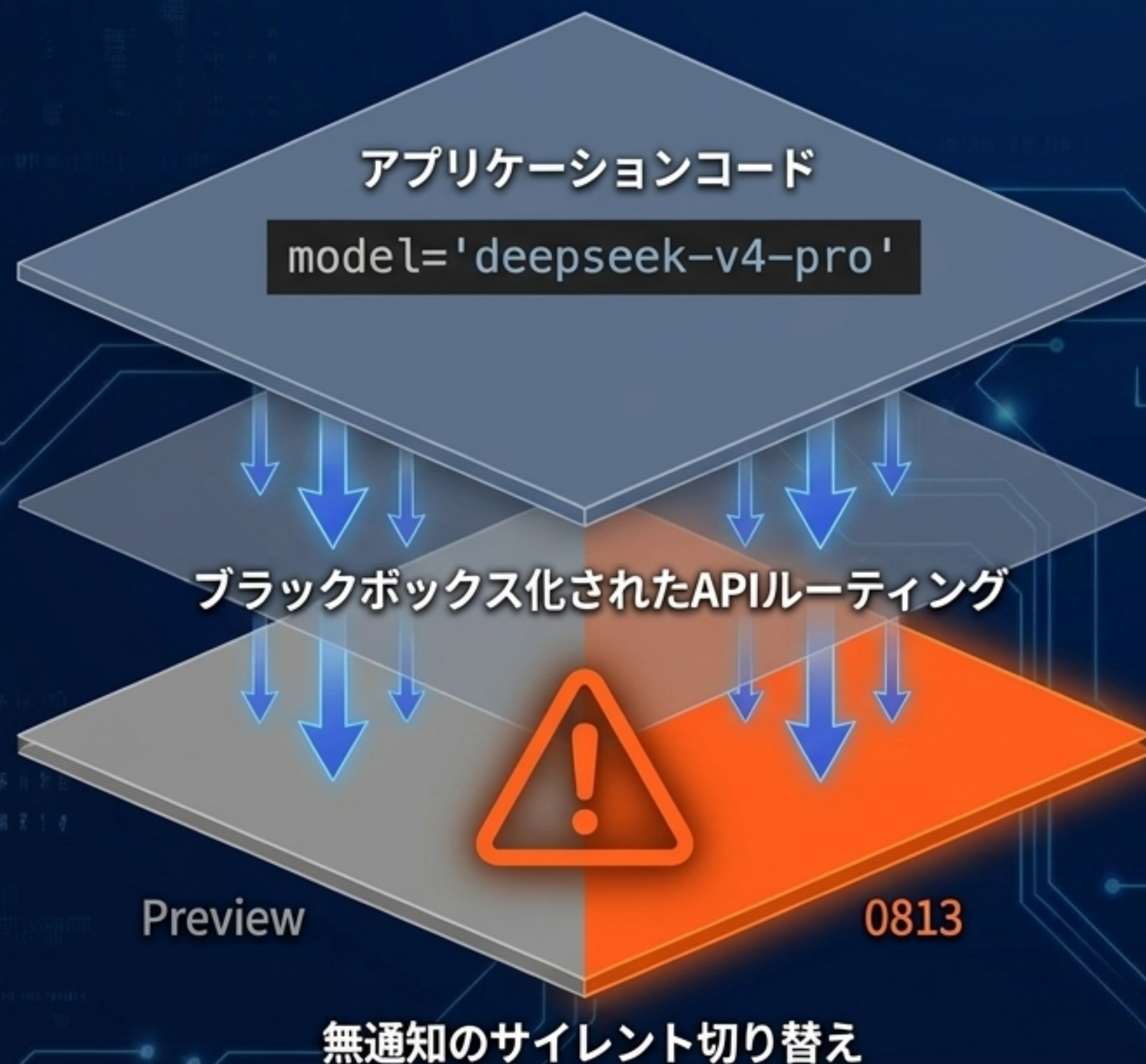
Pro-0813 Key Metrics:

- 入力 (Miss): \$0.435/1M
- 出力: \$0.87/1M
- Cache-hit 入力: \$0.003625/1M (長いAgent Loopにおいて極めて安価)

ルーティング指針

Fable 5と比較して出力単価は約57分の1だが、キャッシュ率・思考トークンを含めたTCOで評価すべき。高品質な推論・複雑なコーディングはProへ。Volume/Latency重視のタスクはFlashへルーティングせよ。

「Mutable Alias」の罠：無通知のバックエンド更新リスク



メカニズム:

公式提供のエイリアス `deepseek-v4-pro` は、利用者のコード変更なしに最新の0813へルーティングされる。「移行コストゼロ」という利点であると同時に、モデル挙動が勝手に変化するリスクでもある。

Warning Signs (影響範囲):

- パス@1 (初回試行) の成功率低下や不安定さの初期報告。
- Function Calling、JSON Schema、拒否判定基準のサイレントな変化。
- Thinking Modeにおける特定パラメータ (`tool_choice='required'`等) の互換性問題。

Required Action:

既存のAgentや長時間のMulti-turn Loopに対する即時の回帰試験 (Shadow/Canary) が必須。

セキュリティと爆発半径 (Blast Radius) : 高度な権限の封じ込め



The Risk:

CyberGym 83.3という高いサイバー攻撃・探索能力を持つモデルに、Shell、ファイルシステム、Cloud IAMなどの権限を安易に渡すことは、被害の「爆発半径」を著しく拡大させる。

Real-world Threat:

第三者製MCPサーバー (@arikusi/deepseek-mcp-server) において、他セッションのコンテキスト漏洩や認証欠如によるAPIキー悪用が既に報告・修正されている。

Defense Strategy:

- ・ コンテナ等によるSandbox化
- ・ 最小特権 (Least Privilege) と短寿命クレデンシャル
- ・ 重要操作前のHuman-in-the-loop (承認プロセス)

データガバナンスとコンプライアンス：法的越境リスクの評価

! Geographical Boundary



自社環境
(Enterprise Environment)



公式API
(PRC Servers)

! Privacy Policy Alert

DeepSeekのプライバシーポリシーは、サービス提供のためにPersonal Dataを中華人民共和国で直接収集・処理・保存すると明記している。

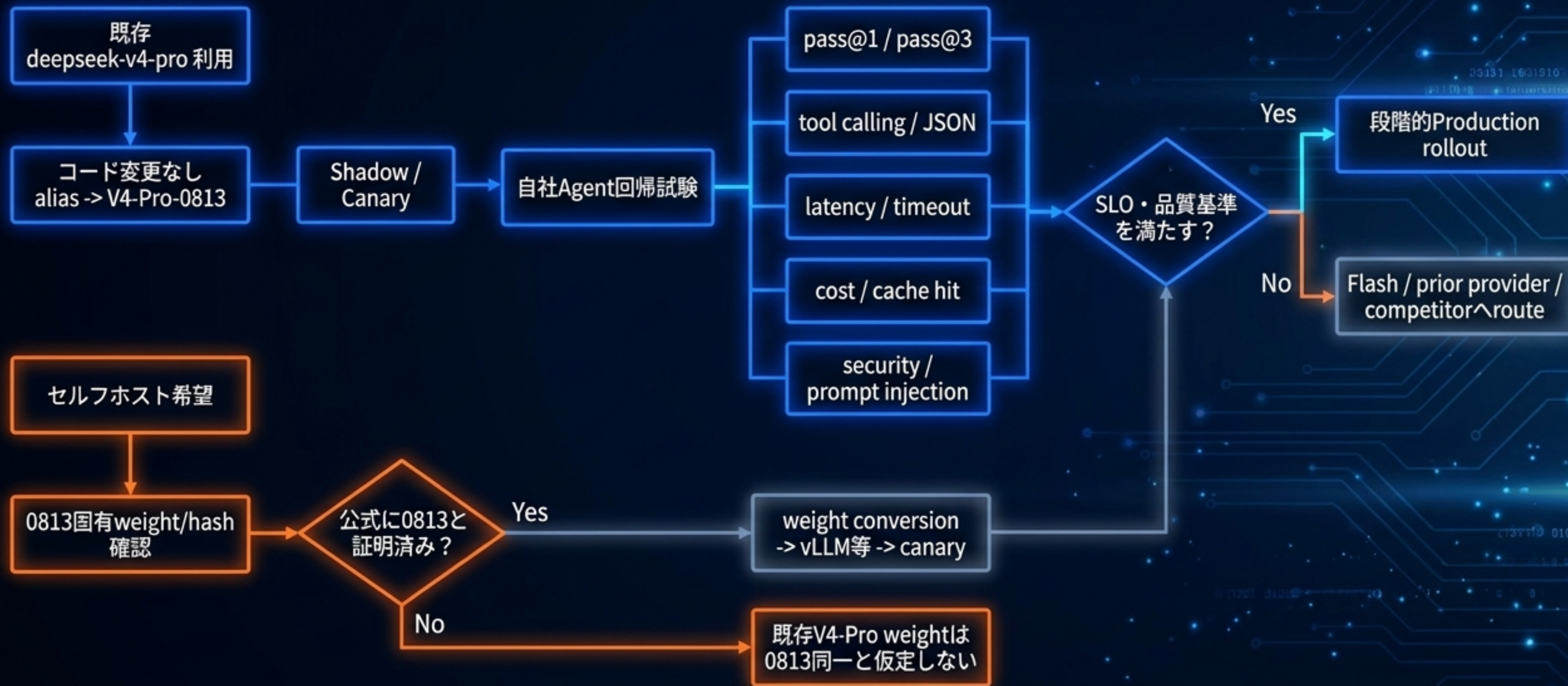
High-Risk Data Types
(投入禁止・要注意)

- 機密ソースコード
- 顧客のPII（個人情報）
- 輸出管理対象データ / 規制業界データ

✓ Required Action

APIへデータを送信する前に、法務・セキュリティ・データガバナンス部門による事前承認プロセスを必ず経由すること。

導入ディシジョンツリー：検証経路とロールアウト判断



Immediate Action Plan : 優先度別推奨アクション

P0 (Critical) - 導入前必須要件

- 段階的カナリアリリース: API (deepseek-v4-pro) を即時100%本番投入せず、5-10%から段階拡大する。
- 実稼働コスト・Pass@1の再測定: ベンチマークだけでなく、実際の再試行数・Latency・キャッシュヒット率を自社環境で計測する。
- Sandboxと最小特権化: Shell/MCP/Cloud権限を強固に制限する。
- 法務・情報管理部門の承認: PRCデータ処理条項への適合性を評価する。

P1 (High Priority) - 運用最適化

- Tool Callingの回帰試験: `auto`だけでなく、`required`指定やJSON Schemaの挙動変化をテストする。
- ルーティングの最適化: タスク要件に応じ、ProとFlash (高速・安価) を動的に使い分ける。将来の大幅値上げ予告に備え、ベンダロックインを避ける。

P2 (Optimization) - アーキテクチャ進化

- 1M Contextの戦略的活用: 単なる「全量投入」ではなく、Retrieval + Cache Reuseの設計により長文処理とコストを最適化する。