Diagnostic Report

Grok 4.7 導入診断・ROI評価レポート

GPT-6 Astra / Claude Fable 5.1 対比における真の性能、
隠れたコスト、知財・法務領域への適性

調査基準日: 2026年9月22日 |
対象: Tech Executives, AI Strategists,
Legal Consultants

誇大広告と現実の境界線：Grok 4.7の3つの本質

The Claim

「長時間の推論」に特化した特化型ワーカー。最大50万トークンの文脈処理と、難度に応じた推論量量調整機能（xHighなど）を搭載。

The Reality

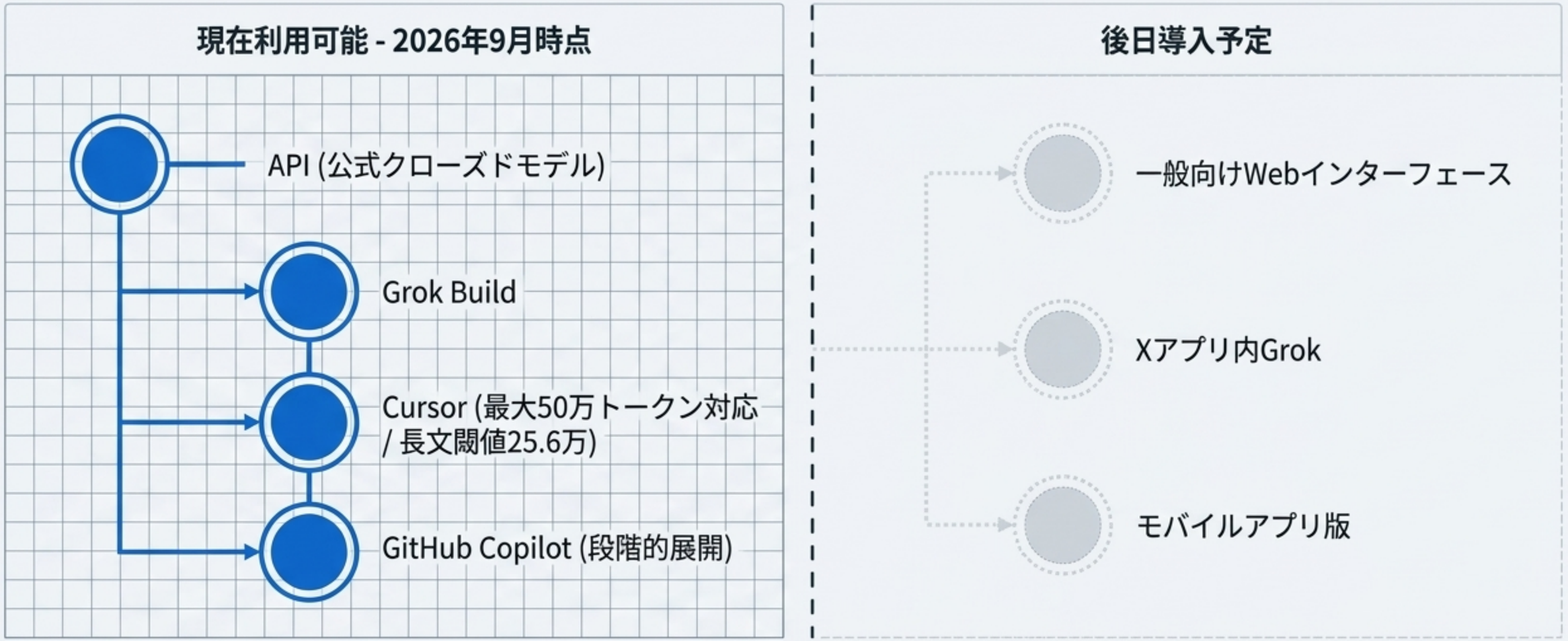
独立評価による実力は「総合4位」。知識業務の分析力は向上するが、GPT-6 Astraのような「万能なトップモデル」ではない。

The Hidden Cost

見かけのAPI単価は据え置きだが、xHigh設定時のトークン出力が約2.1倍に膨張し、実質的なコストは上昇する。

[Official Model Card / Release Data]

導入経路の現実：Grok 4.7が「今」利用できる場所



⚠ **注意：**普段使用している一般向けGrok画面で選べるモデルとは、現時点で区別して検証する必要がある。

公式ベンチマーク：特定領域における圧倒と、明確な劣後

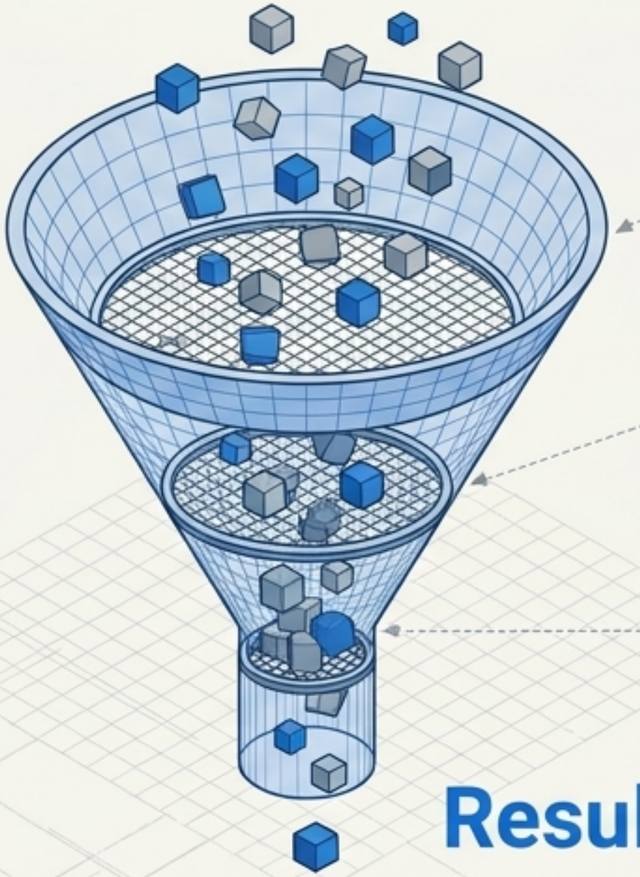
Verification Tag
[Official Benchmarks / SpaceXAI Docs]

評価対象	Grok 4.7 (xHigh)	GPT-6 Astra	Claude Fable 5.1
電気工学 (EEBench)	64.0%	39.4%	56.4%
ソフトウェア開発 (DeepSWE v1.1)	71.0%	72.7%	70.0%
端末操作作業 (Terminal-Bench 4.0)	38.0%	37.3%	57.9%
複数時間の事務作業 (AA-Briefcase)	1,657	1,487	1,678

端末操作や自律的エージェント業務においては、Claude Fable 5.1に大きく水をあけられている。

Verification Tag
[Vals AI Evaluation Criteria]

Harvey Legal Benchmark「19.6%」の正しい解釈



課題に対する部分的な正答
(高確率で通過)

関連判例・条文の正確な引用
(一部脱落)

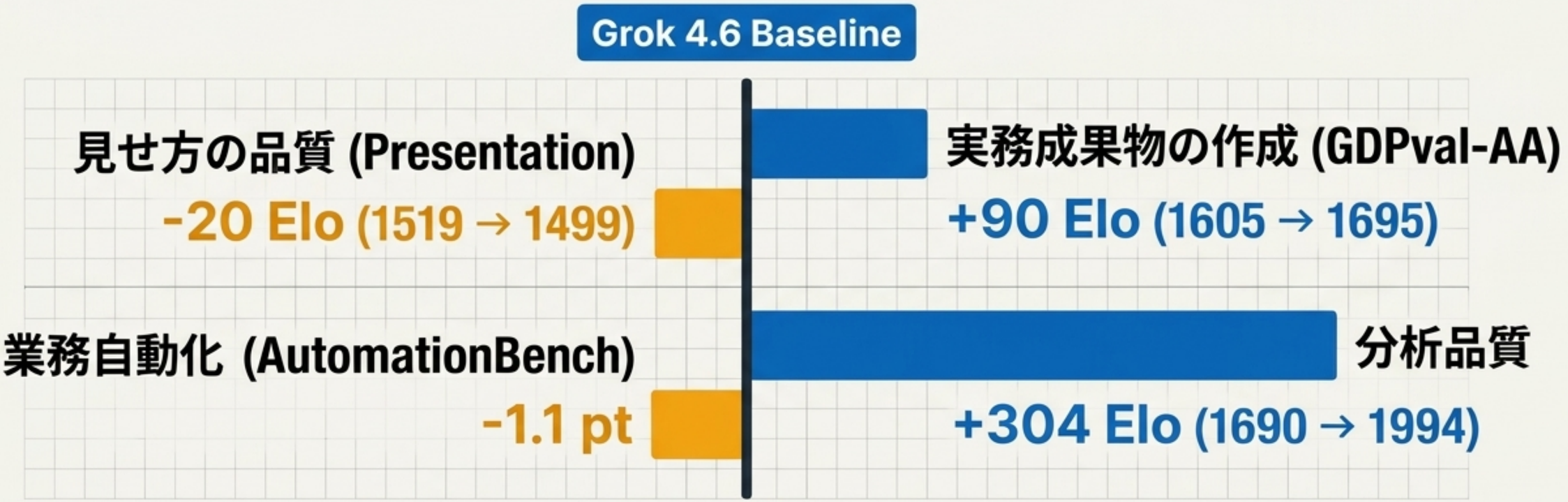
すべての評価条件の完全な網羅
(All-or-Nothing採点)

Resulting Output: 19.6% 合格

「法律問題の正答率が19.6%」ではない。重要な条件を1つでも欠けば「不合格」となる、極めて厳格な専門業務の実用性テストである。部分的な精度は高い。

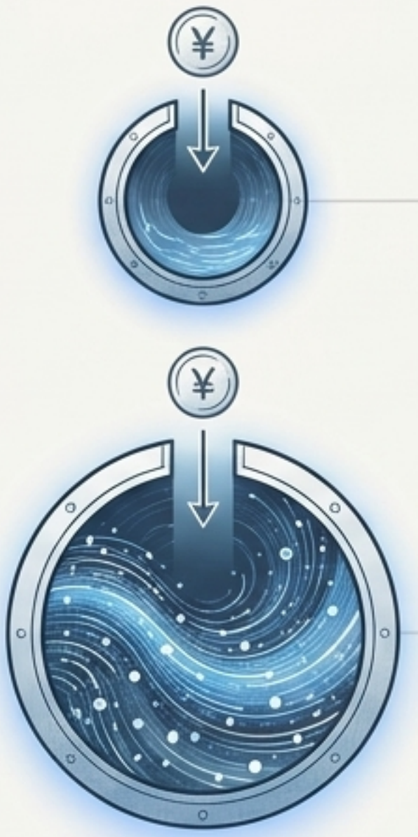
Verification Tag
[Artificial Analysis - Independent Data]

独立評価機関の現実：分析力と「見せ方」のトレードオフ



資料の「深い分析」には強いが、「読みやすい完成品」への仕上げ能力は低下している。出力後の人間による清書プロセスが必須。

トークン消費の錯覚：xHigh設定がもたらす隠れたコスト増



Grok 4.6 High設定 1課題あたりの平均出力: 約3.8万トークン

Grok 4.7 xHigh設定 | 1課題あたりの平均出力: 約8.1万トークン (2.1倍)

宣伝文句の「単価据え置き」は事実。しかし、xHigh設定での推論プロセスの膨張により、出力トークン量が約2.1倍に増加。結果として、出力部分の請求額は自動的に跳ね上がる。

Verification Tag [API Pricing & Real-World Cost Data]

実行コストの真実：カタログ価格 vs. 実働単価

カタログ上のAPI料金 / 100万トークン

入力（<20万トークン）: $2	出力: $6
長文入力 (>20万トークン): $4	出力: $12

注意: 長文閾値を超えると、超過分だけでなくリクエスト全体に長文料金が適用される。

1タスクあたりの実費 - Artificial Analysis基準

High設定: $2.73 / 課題

xHigh設定: $3.74 / 課題（コスト約37%増）

常にxHighを使うべきではない。標準業務は「High」で運用し、複雑な推論が必要な課題のみ「xHigh」にエスカレーションする運用設計が必須。

日本語環境における初期評価：実務投入へのハードル

現時点で、日本語の特許実務や専門文書における大規模な独立評価は存在しない。初期の利用者報告に基づくリスク評価が必要。

1. 日本語の硬さ

生成される日本語文章に不自然さや硬さが残る報告が多数。顧客向け最終成果物には不向き。

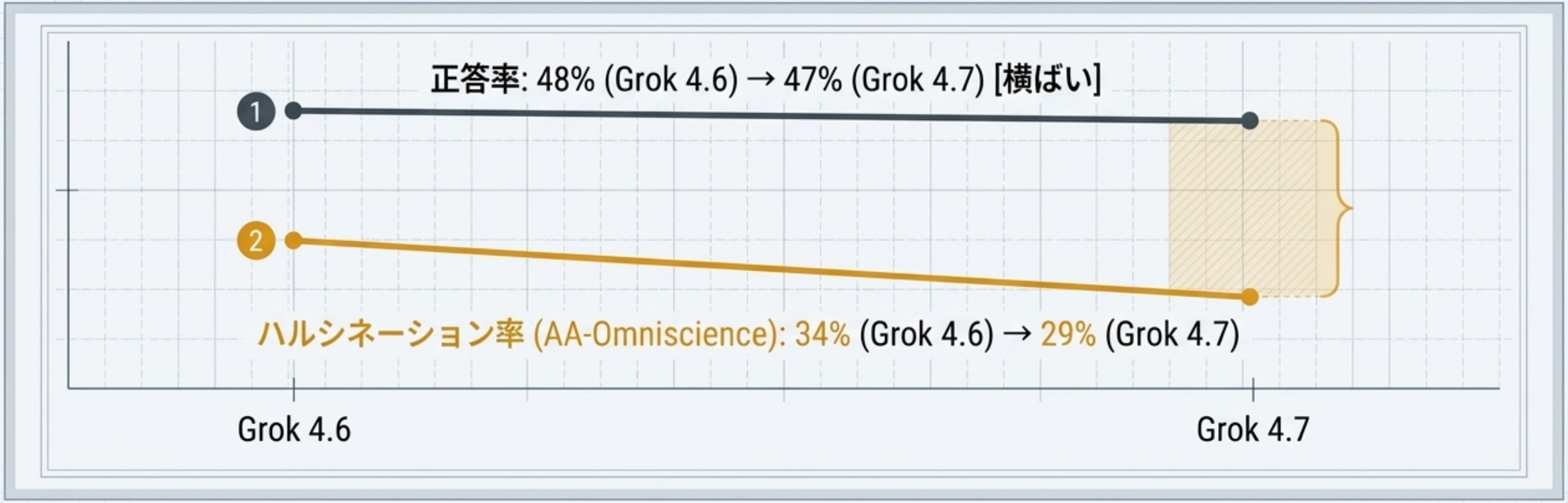
2. 大量文脈での理解低下

大量のコンテキスト（50万トークン等）を入力した際、日本語特有の文脈理解が低下する兆候あり。

3. 仕上げの劣後

初稿作成（HTML化など）は有望だが、**アニメーションやUIの細かな仕上げにおいてはGPT-6 Astra**に支持が集まる。

信頼性の構造：「ハルシネーション低下＝正答率向上」ではない

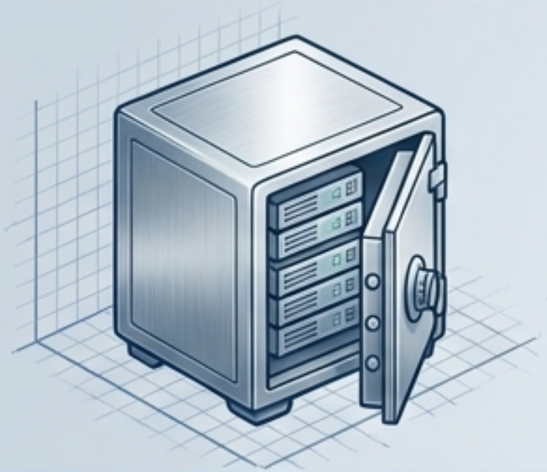


Insight Box

「29%」は全回答中の誤り率ではない。知識不足に直面した際の「知ったかぶり」が減り、「回答保留」を選択する能力が向上したことを意味する。根拠確認 (ファクトチェック) の必要性は依然として変わらない。

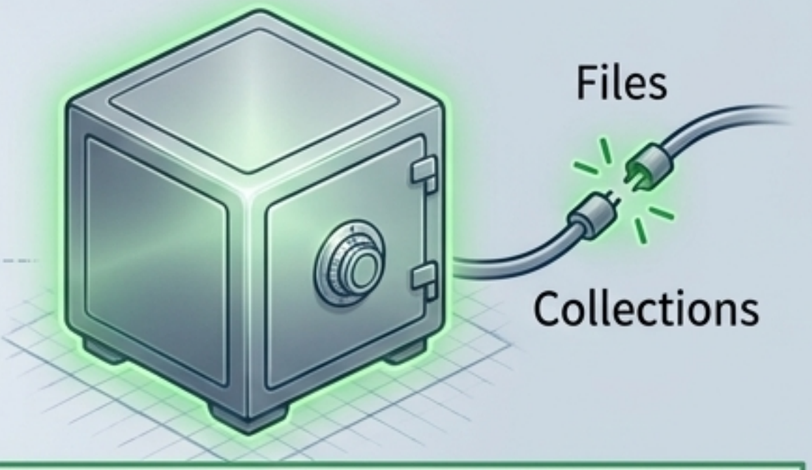
法務・知財領域のトラップ：データ保持とAPI仕様

標準設定



- API入力・出力は学習に利用されない（明示的許可なしの場合）。
- 監査目的で「30日間保持」される。未公開発明データには不適格。

Zero Data Retention設定



- データ保持ゼロ。機密情報の入力が可能。
- トレードオフ: FilesやCollectionsなどの機能に制限が生じる。

このAPIの条件を、一般向けGrokや他社経由のサービス（Cursor等）にそのまま当てはめてはならない。経路ごとのNDA・約款確認が必須。

[Consultant Synthesis & Evaluation] ✓

知財実務における適性評価：何に使い、何を避けるべきか

✓ High Fit（適性が高い作業）

- ✓ 複数資料からの比較表や報告書の構造化
- ✓ 特許データ（大量のテキスト）を加工・集計するコードの作成
- ✓ クレーム構成要件の機械的な抽出（抽出漏れの確認用）

✗ Caution (検証が必要な作業)

- ✗ 引用文献の文脈と、対比内容の高度な意味的照合（FTO分析）
- ✗ クレーム解釈や進歩性の最終判断
- ✗ 日本語での流暢な拒絶理由通知応答の起案（日本語の修正時間がかかる）

【Synthesized Action Framework】

導入マトリクス：Grok 4.7 活用ポートフォリオ

[Final Consultant Verdict]

Verdict & 2026 Integration Strategy

1

設定の二重化 (Tiered Reasoning)

デフォルトは「High」設定に固定し、コスト増を防ぐ。複雑なコーディングや複数資料の推論のみ「xHigh」に切り替えるワークフローを構築する。

2

「清書」からの分離 (Decouple Presentation)

Grok 4.7は「思考エンジン」として扱い、最終的な日本語の流暢さやドキュメントの仕上げにはGPT-6 Astra等の別モデル（または人間）を組み合わせる。

3

知財領域のデータ隔離 (Isolate IP Data)

API利用時は必ず「Zero Data Retention」を申請。一般UIからの未公開情報入力は厳禁とし、利用経路ごとのガバナンスを徹底する。