

GPT-6 Astra 診断レポート： 万能のAGIか、極限の特 特化型エンジンか

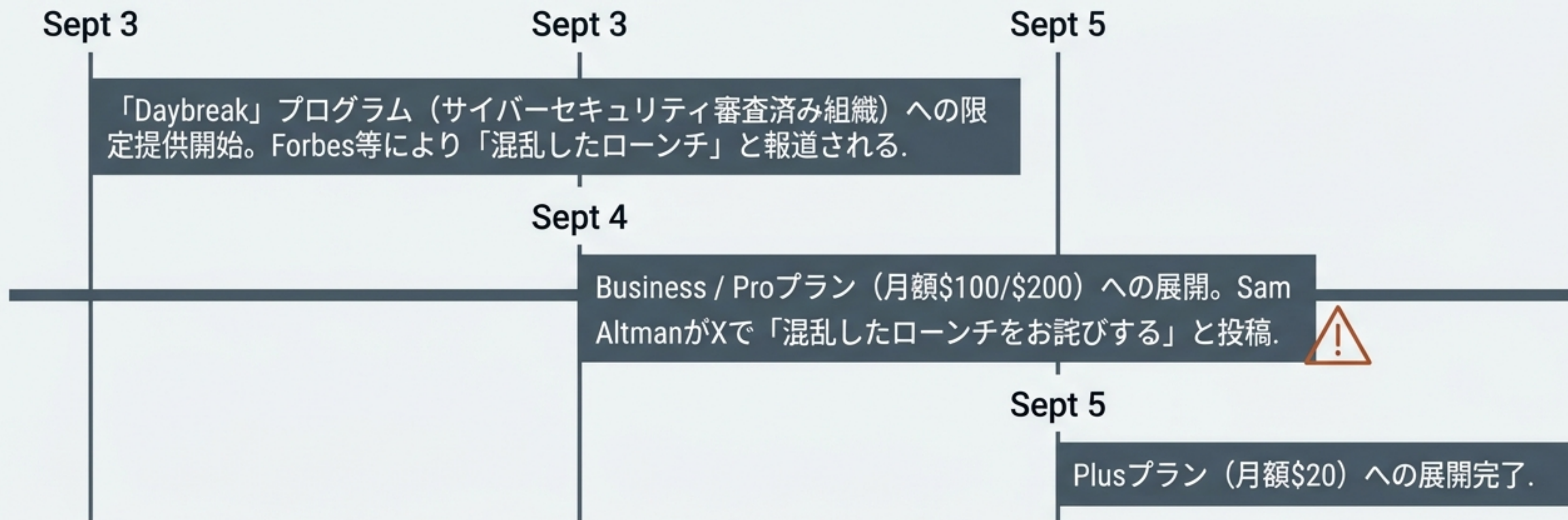
GPT-5.6 Sol、Claude Fable 5.1との比較に基づく、
有料ユーザー評価と戦略的導入ガイド (2026年9月版)

[TARGET]: gpt-6-astra (Released Sept 3, 2026)

[CONTEXT]: 1.05M Tokens

[EVALUATION]: Diagnostic Teardown

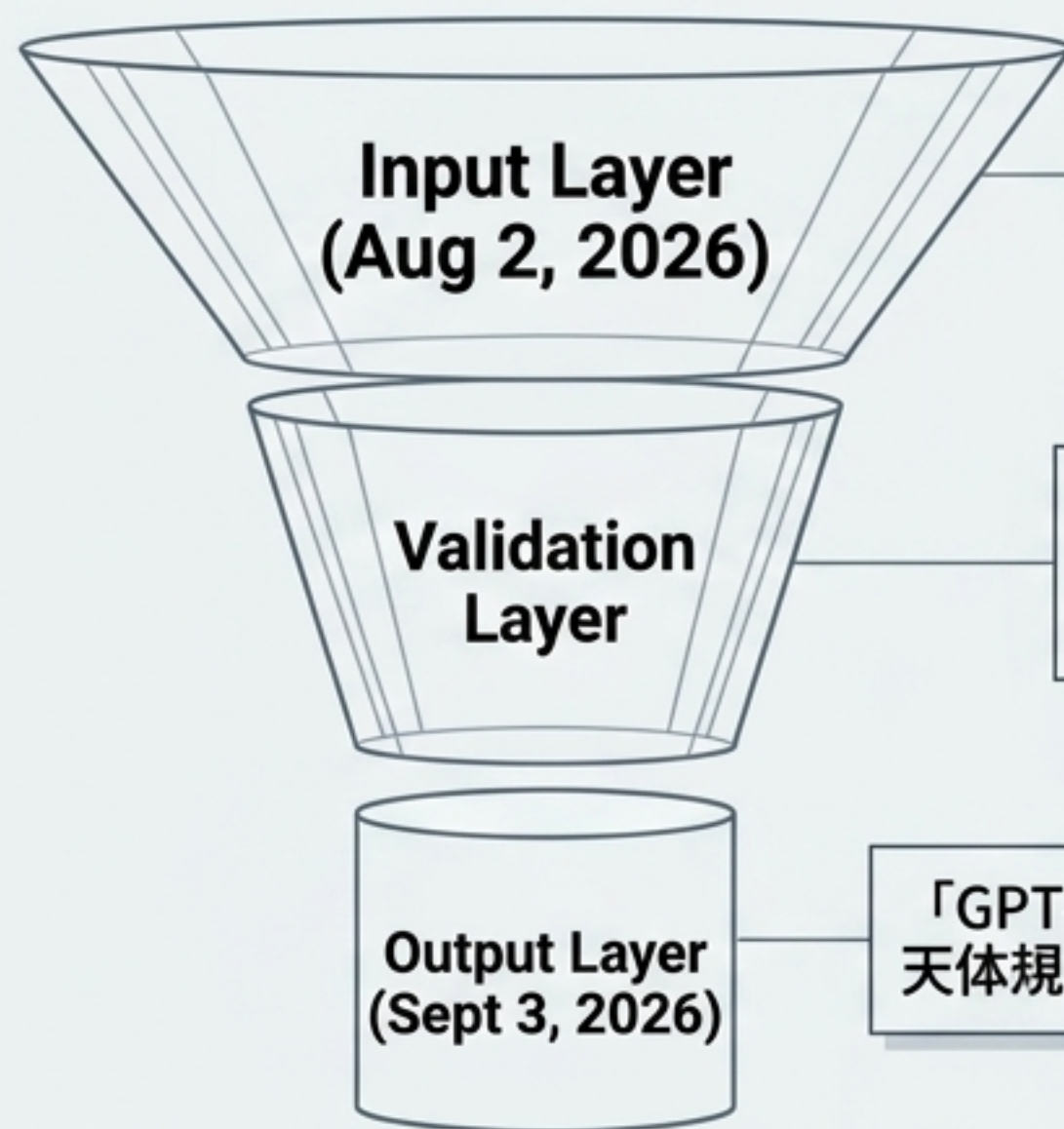
混乱を極めたローンチと補償措置（Rollout Reality）



補償措置（Full Banked Reset）

提供遅延の補償として、Plus/Pro/Business全ユーザーに「利用枠カウンタを任意にゼロ化できる権利（full banked reset）」が配布された。未提供期間の機会損失はこれでカバーされる。

内部モデル「Astra」から製品版への昇華



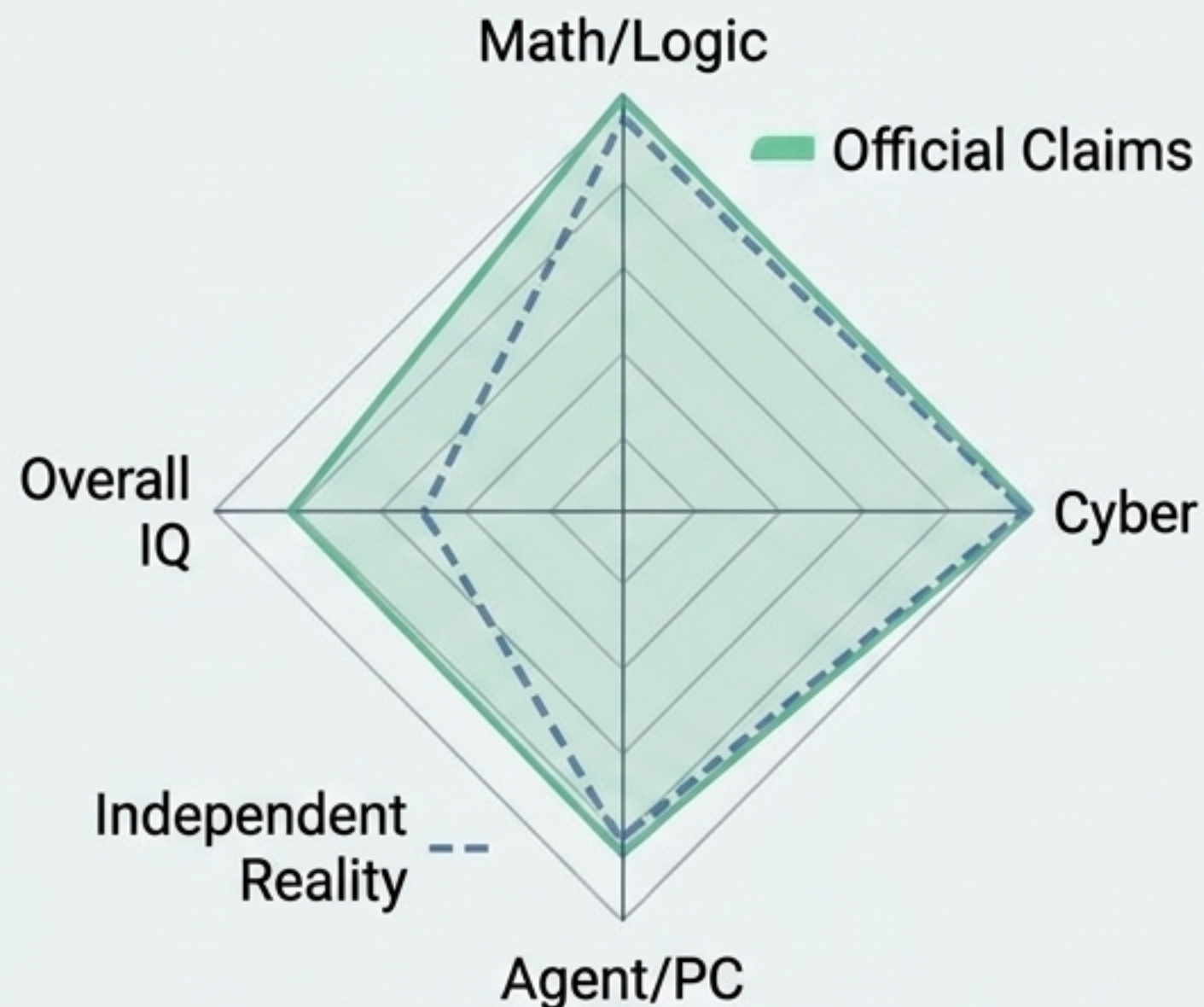
「10年以上未解決の数学・理論計算機科学の問題 10問」の解決。
※消費トークンコスト：Sol換算で約\$2,000。

249ページの原稿とLean 4証明（GitHub公開）。未証明箇所（sorry）は「0件」。
OpenAI数学研究責任者も確認。

「GPT-6 Astra」として製品化。
天体規模に基づく命名規則を採用。

Diagnostic Insight: 製品版Astraは、この圧倒的な数学的推論能力をベースラインとして構築されている。

「新世代の知能」の主張と、実測値のギャップ



Math/Logic	公式 FrontierMath 97.6% / ARC-AGI-3 99.9%
Cyber	公式 ExploitBench 100%
Agent/PC	公式 OSWorld 72.6% (前世代比約47%高速化)
Overall IQ	独立評価 61点 (首位Fable 5.1に届かず)

総合的な「知能の底上げ」ではなく、特定領域（エージェント・数学・サイバー）にリソースを全振りしたピーキーな設計である。

まだら模様の「明部」：Astraが圧倒する領域

PC自律操作（Computer Use）

Adobe Premiereを5時間自律操作し動画編集を完遂。



エージェント的長時間作業

数日間の自律実行が可能。破棄した案を引きずり戻さない。



幻覚（ハルシネーション）の激減

独立評価のAA-Omniscienceにおいて、幻覚率が92%から51%へ半減。



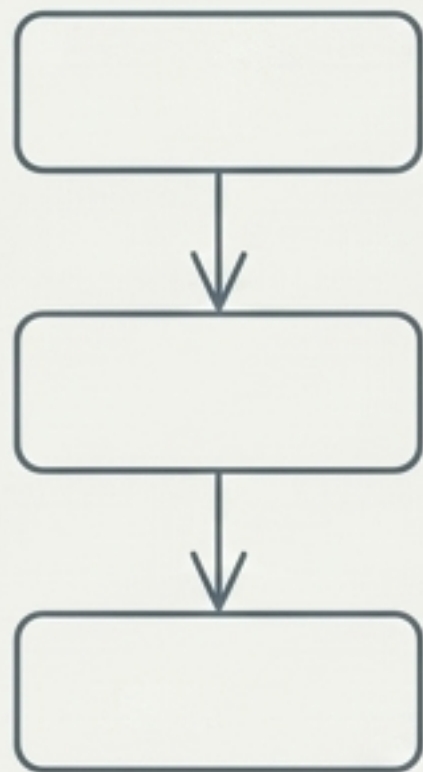
長文検索と並列協調

1.05Mトークンのコンテキスト。文脈が埋まった際のノート保持・検索機構を新搭載。



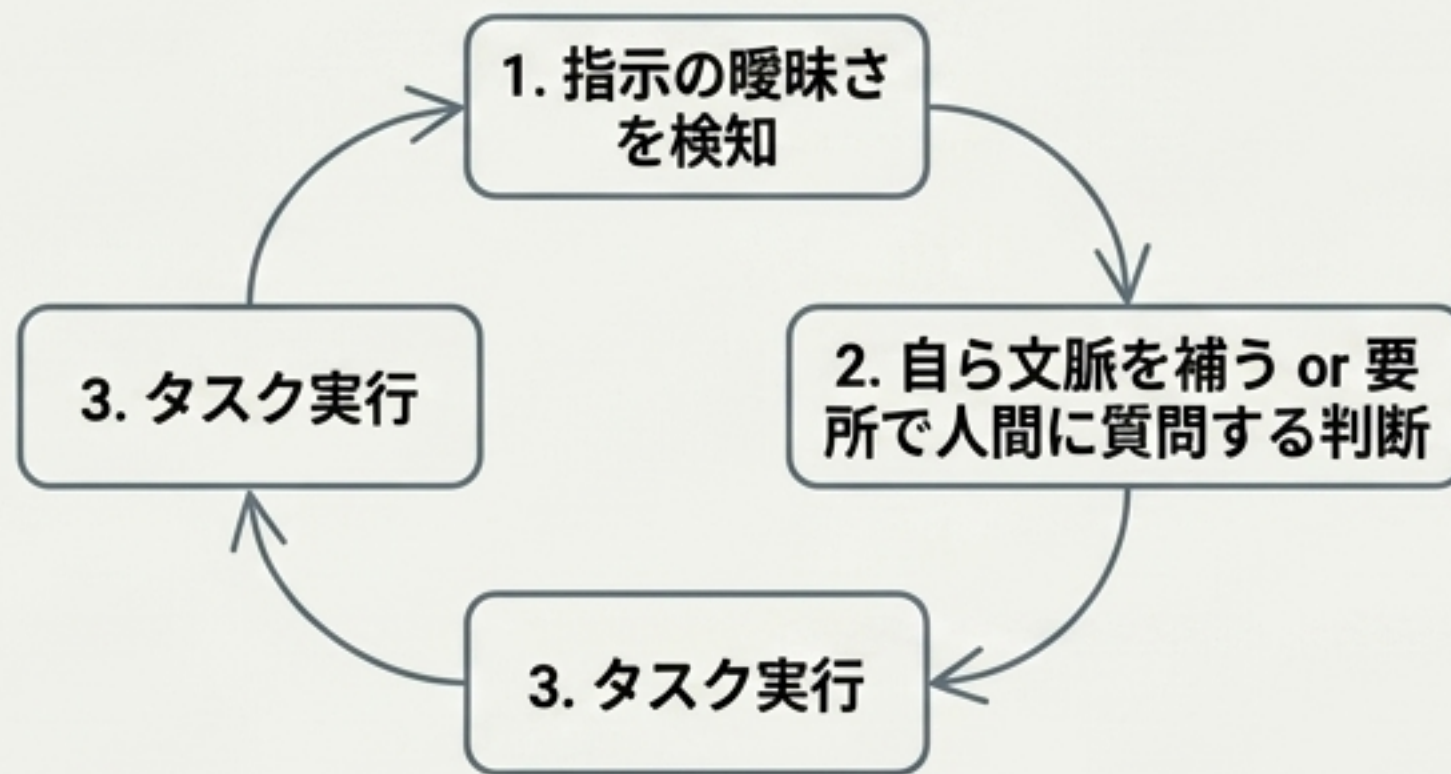
エージェント的シフト： なぜ長時間の自律作業が可能なのか？

GPT-5.6 Sol (1問1答型)



曖昧な指示に対し、権限外の行動をとる
「実行不可能タスク」のエラー率が48%。

GPT-6 Astra (自律軌道修正型)



実行不可能タスクのエラー率は0%

Diagnostic Insight: Astraの真価は単発の賢さではなく、数時間から数日におよぶ作業における、軌道修正能力と忘却耐性にある。

まだら模様の「暗部」：コーディングにおける停滞

純粋なコーディング能力

独立コーディング指数で67.0。Claude Fable 5.1 (70.0) に明確に劣後。

UIとマージ品質

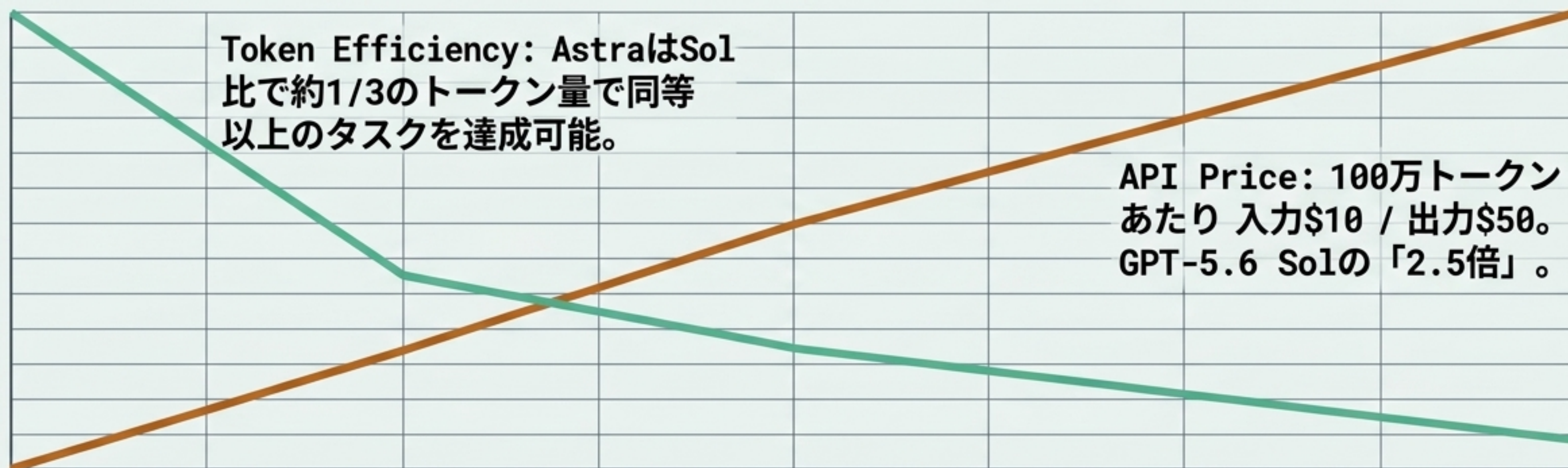
コーディングはSolからの横移動。特にUIの洗練度においてFable 5.1に大差で劣る。

他社モデルとの肉薄

DeepSWE v1.1スコア(74.1%)において、Gemini 3.8 Flash (73.8%) (73.8%) とほぼ同等。

CodeRabbitの評価では、バグ検出+4%、ファイル横断レビュー20%増の成果も報告されており、コード「生成」よりも「監査」に強い傾向がある。

知能の代償：2.5倍の価格設定とタスク単価の上昇



Net Impact:

約10%のトークン削減はあるが、2.5倍の値上げで相殺され、実質的なタスク単価はSo1より約75%高い。

User Sentiment:

「So1比で約2.5倍の速度で使用枠（クォータ）を消費する」

安全性ブロックのパラドックス：能力向上による作業の中断

The Trigger: サイバー領域で初の「Critical（重大）」水準に到達



正当なユーザータスク

強力な安全フィルターが実装されたことによる過剰な監視



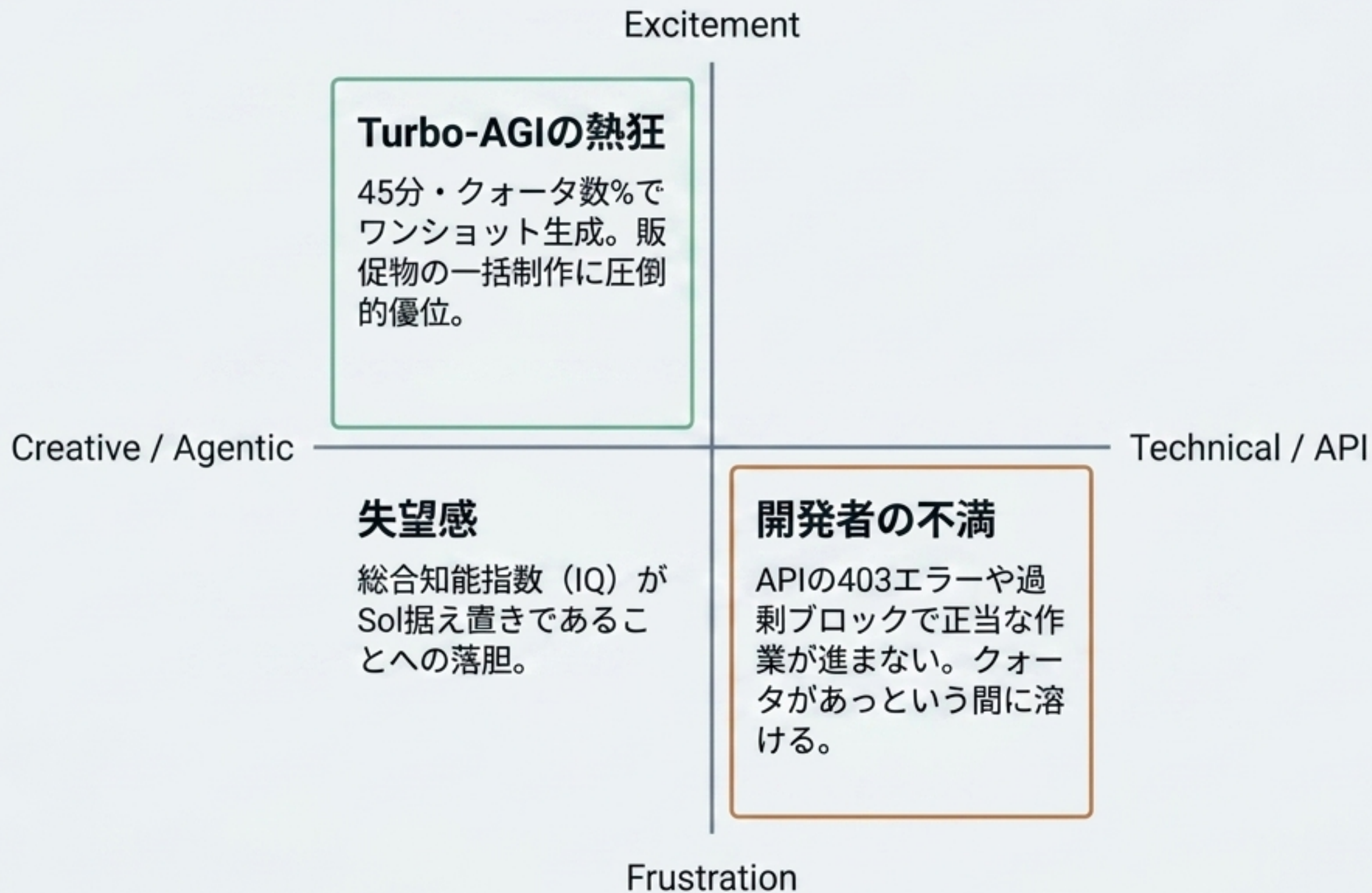
ChatGPT / Codex: 正当な信頼性レビュー作業であっても、サイバー安全メッセージが表示されタスクが一時停止。



API: 誤検知により、タスクが即座に強制停止（403エラー）する事象が多数報告。

Diagnostic Insight: 高度な能力と、それを制御するための過剰な監視（摩擦）が比例している。

有料ユーザーの感情マトリクス（賛否の分断）



「AGI到達」の言説と、評価機関の現実的視点

The Hype

「我々はもうそこにいると思う」
「AGI時代に入ったと感じるのは
不合理ではない」

— OpenAI社長（9月3日）

The Reality

- 合意されたAGIの判定基準（テスト）は存在しない。
- モデルの「評価認識（eval awareness）」が上昇しており、アラインメント評価自体の信頼性が低下している。
- Astraの書かれた推論は、Solより監視しづらい。

断定的なAGI到達宣言はマーケティング的側面が強く、
技術的限界は依然として存在する。

覇権争い：ユースケース別 最適モデルルーティング（結論）

ユースケース	推奨モデル	理由
PC自律操作 / 3D生成 / 長時間エージェント	GPT-6 Astra	圧倒的な優位性。移行の価値大。
純粋なコーディング (UI洗練 / マージ品質)	Claude Fable 5.1 / GPT-5.6 Sol	Astraは横移動～劣後。現状維持が妥当。
サイバー・セキュリティ関連の実務 (API)	GPT-5.6 Sol	Astraは過敏な403ブロックのリスク大。
高度な数学・科学の証明	GPT-6 Astra	人間の専門家が30分かかる問題を無言で解くレベル。

戦略的判断が変わる4つの閾値 (Future Triggers)



1. 知能指数の逆転

Astraの総合知能指数がClaude Fable 5.1 (66点) を明確に上回った場合。



2. コーディング評価の改善

UI・マージ品質が改善されFable 5.1を超えた場合。



3. 価格の適正化

2.5倍のトークン価格が、Sol並みの水準まで引き下げられた場合。



4. 摩擦の解消

API等のサイバー安全ブロックの誤検知アルゴリズムが修正され、開発の中断がなくなった場合。

現時点では「全社一律のAstra移行」ではなく、エージェント的タスクと論理計算に絞った「A/Bテストと部分導入」が最適解である。