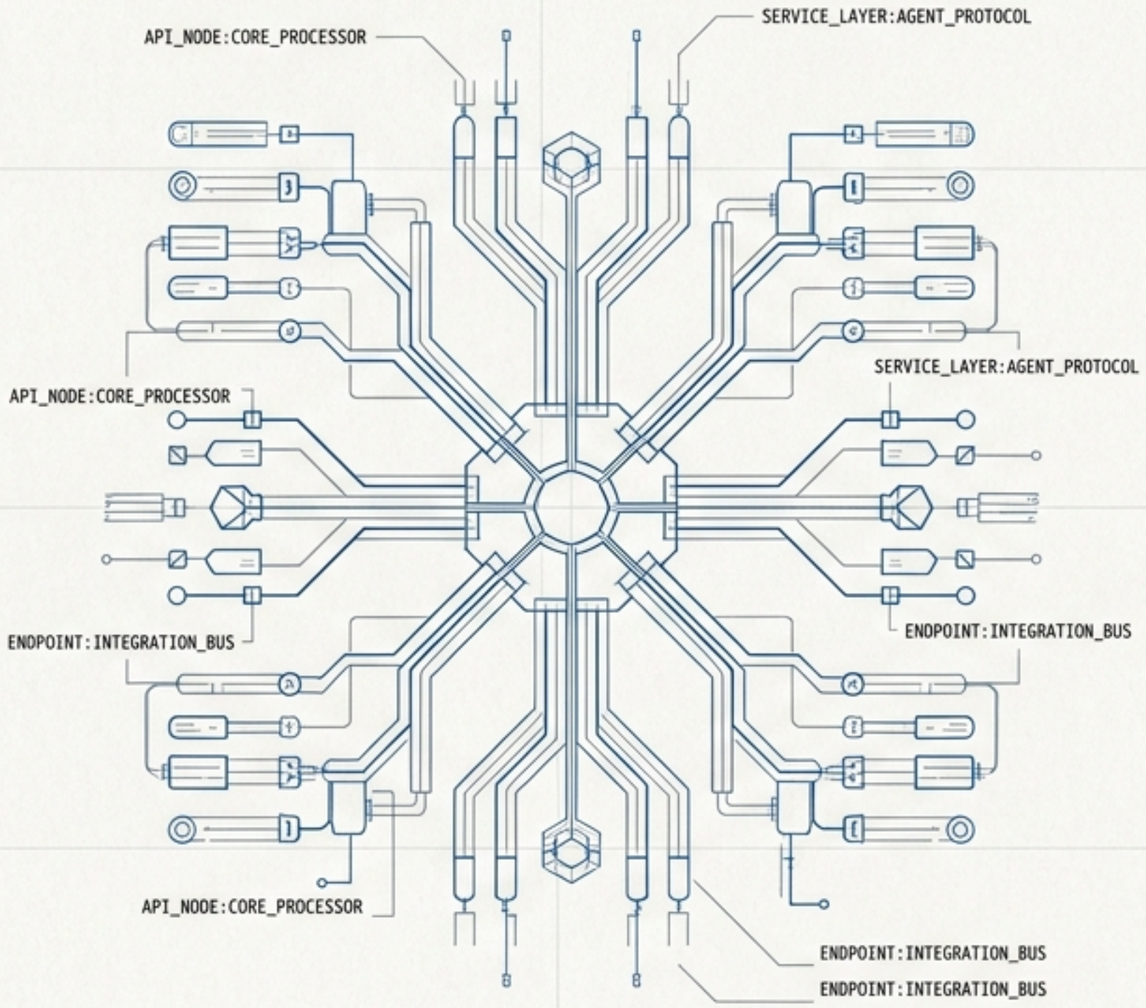
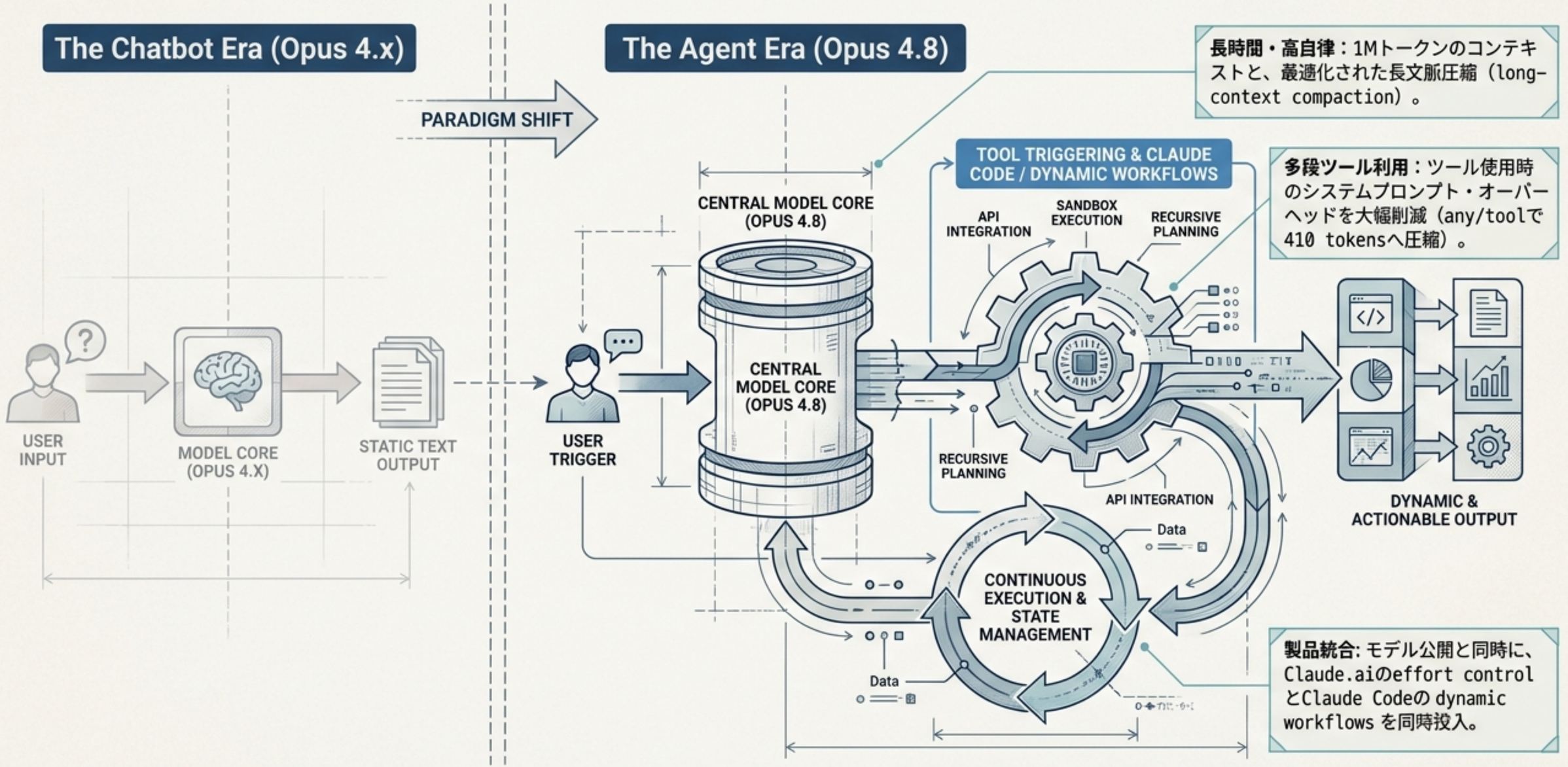


Claude Opus 4.8 診断報告：自律型エージェントへの資渡を解剖する過渡期を解剖する

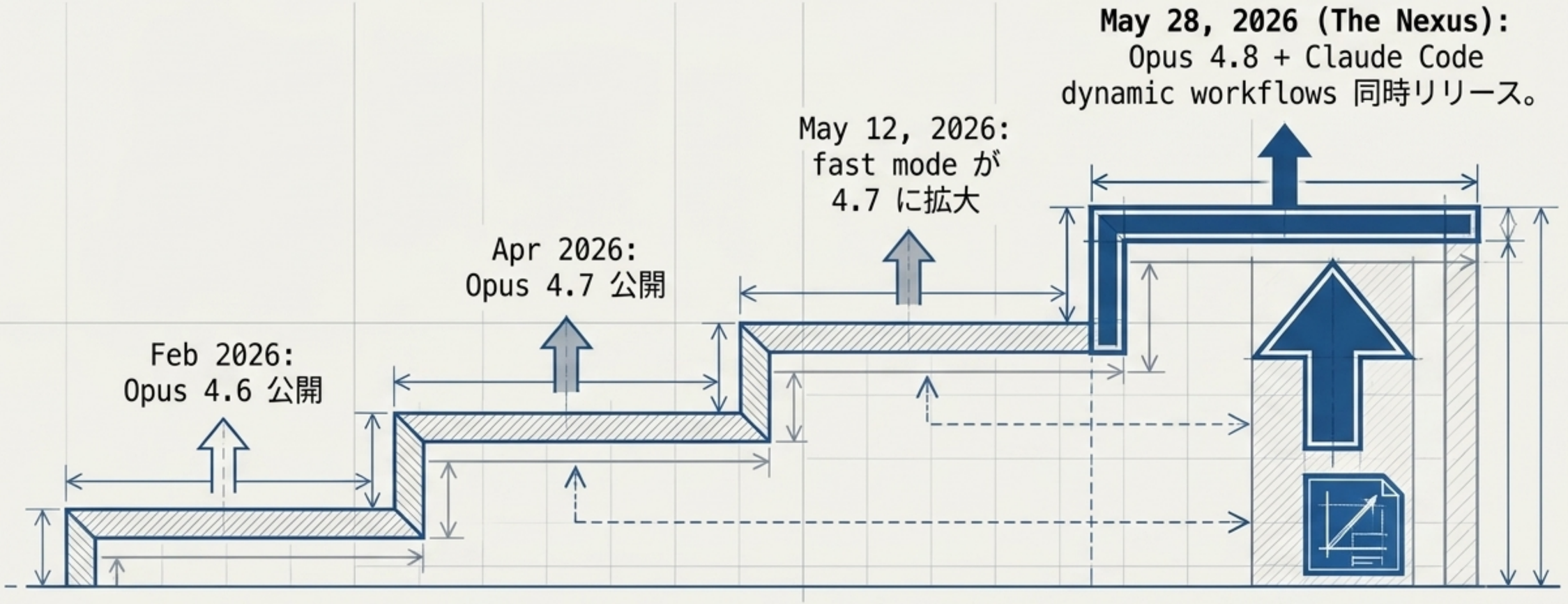
モデル知能の進化と、エージェントスタックの未熟さが交錯する最前線の分析

TARGET_AUDIENCE: [ENGINEERING_LEADERSHIP, PRODUCT_MANAGERS, STRATEGY] // STATUS: CONFIDENTIAL_DIAGNOSTIC

パラダイムシフト：単なる「モデル更新」ではなく「エージェント運用を前提とした製品レイヤーの更新」



Opus 4.x シリーズの高速反復：統合点としての「4.8」



Opus 4.8は突然の飛躍ではなく、2026年前半を通じて連続的に束ね直された
「モデル本体・fast mode・Claude Code・エンタープライズ制御」の最終統合点である。

システム仕様と経済性：実働パラメータの診断

System Console

Context & Output

CONTEXT_WINDOW: 1M tokens
(API/Bedrock/Vertex)

MAX_OUTPUT: 128k
(同期 Messages API) /
300k (Batch API Beta)

Reasoning Control

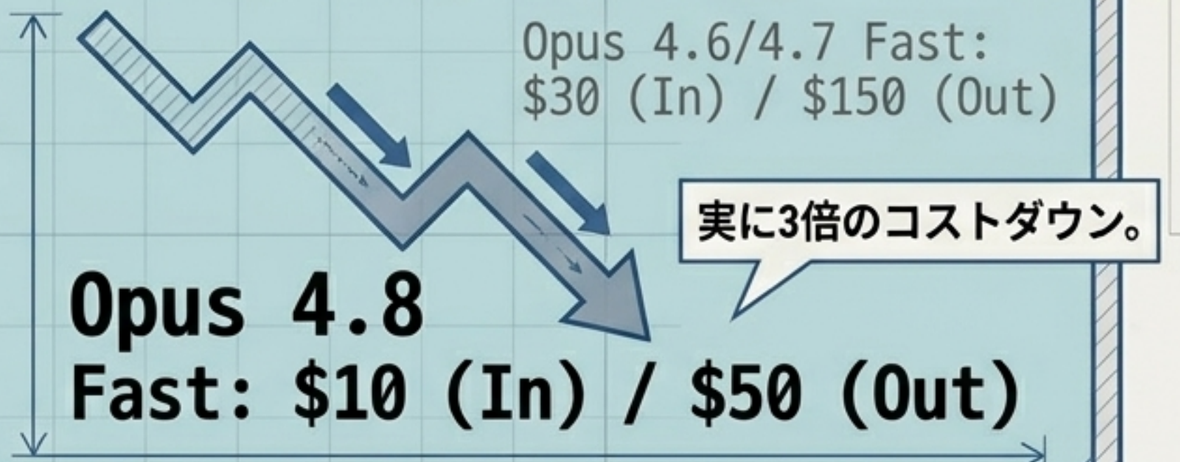
THINKING_MODE: adaptive のみ
(手動 budget_tokens は非対応)

DEFAULT_EFFORT: high

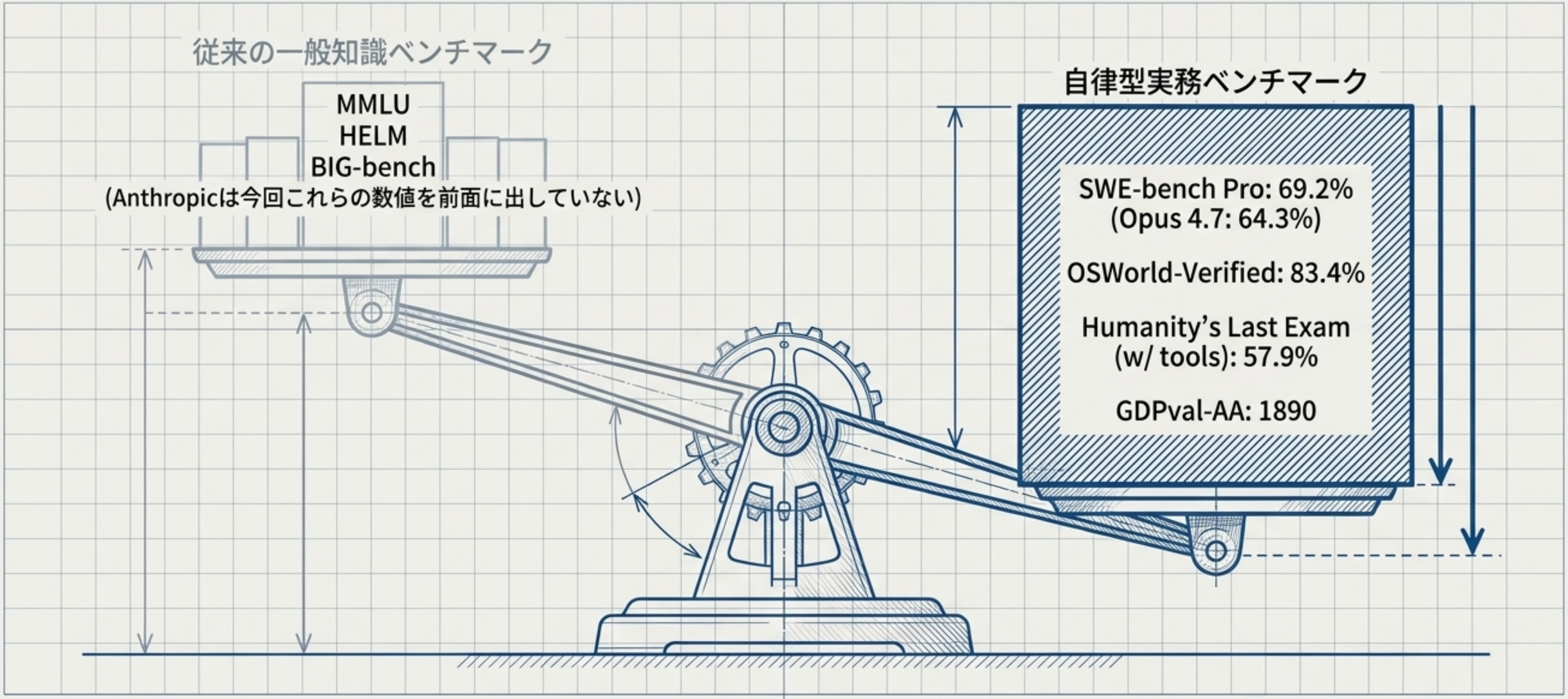
Optimization

PROMPT_CACHE_MIN: 1,024 tokens
(4.7より短文で適用可能)

Fast Modeの破壊的経済性



評価軸の重心移動：「静的知識」から「動的エージェント行動」へ



Insight: Opus 4.8の主戦場は、単発のクイズではなく「リポジトリ横断・複数段階・ツール併用のエージェント仕事」に完全に移行している。

同世代モデル比較マトリクス：「得意領域」の分水嶺

Benchmark	Opus 4.8	GPT-5.5	Gemini 3.1 Pro
Group A: リポジトリ横断 / 長文脈エージェント的仕事			
SWE-bench Pro	**69.2%**	58.6%	54.2%
GDPval-AA	**1890**	1769	1314
Group B: 端末中心の自律コーディング			
Terminal-Bench 2.1	74.6%	**78.2%**	70.3%

Opus 4.8は「全方位最強」ではない。端末自動化（Terminal）においては依然としてOpenAI系に優位性が残るが、長文脈の判断・知識労働・ツール連携の一貫性において明確な首位を獲得している。

誠実さ（Honesty）のメカニズム：誤答率低下のパラドックス

内部アラインメントの強化
(Mythos Preview級のprosocial traits)

→

不確実性の感知
(Unsupported claimを出さない)

→

能動的な停止機能 (Abstain)
(わからない時に適当に答えない)

→

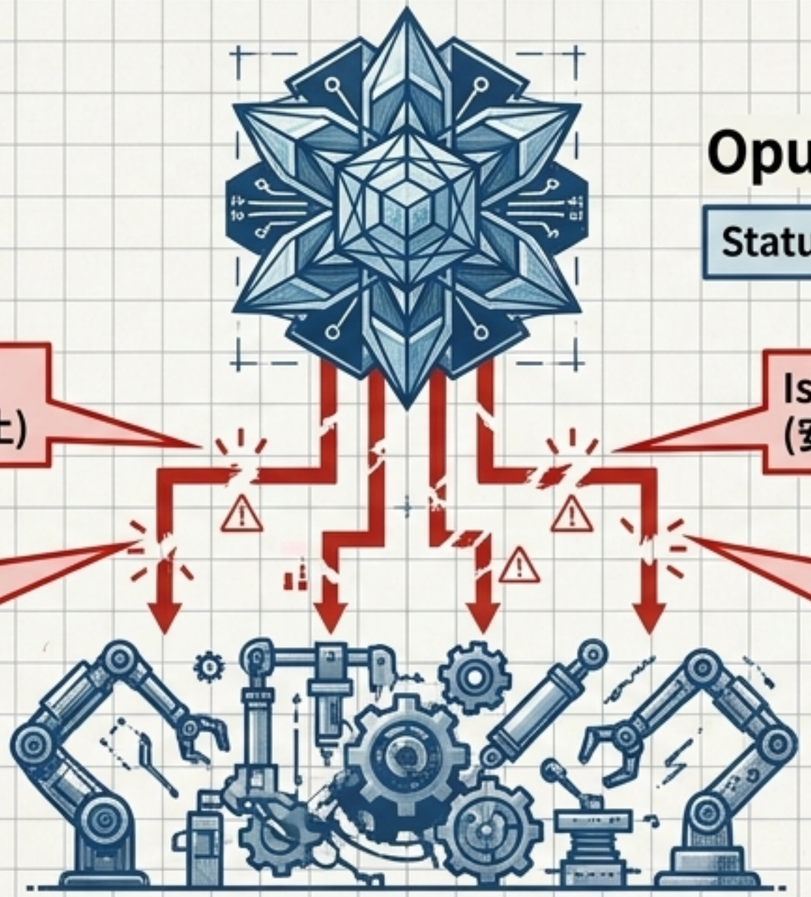
事実ハルシネーション（誤答率）の劇的な低下

インパクト:

自分のコードの欠陥を見逃して黙って通してしまう確率が、前世代の約4分の1に減少

安全性はもはや「有害な回答の拒否」だけではない。プロ用途（法務・コーディング）においては、**「進捗を誇張しない」「証拠のない成功宣言をしない」**という**実務的信頼性**が安全概念の核となっている。

統合インサイト：進化する「脳」と、バグを抱える「神経伝達」



Opus 4.8（推論・モデル）

Status: Highly aligned, Honest, Intelligent.

Issue #63604: Malformed JSON
(tool_useのJSON破損による会話停止)

Issue #63819: Safety Classifier Down
(安全判定レイヤーの一時不通による実行停止)

Issue #63884: Premature Output
(並列ツール結果が返る前に、数値を捏造して先走り出力)

Issue #63523: Boundary Violation
(current worktree の外側を書き換える越権行為)

Claude Code / Tools（実行レイヤー）

モデル内部（脳）は誠実になったが、ツールオーケストレーション（手足の神経）には重大なfailure modeが残存している。これが公開直後のユーザー報告が荒れた根本原因である。

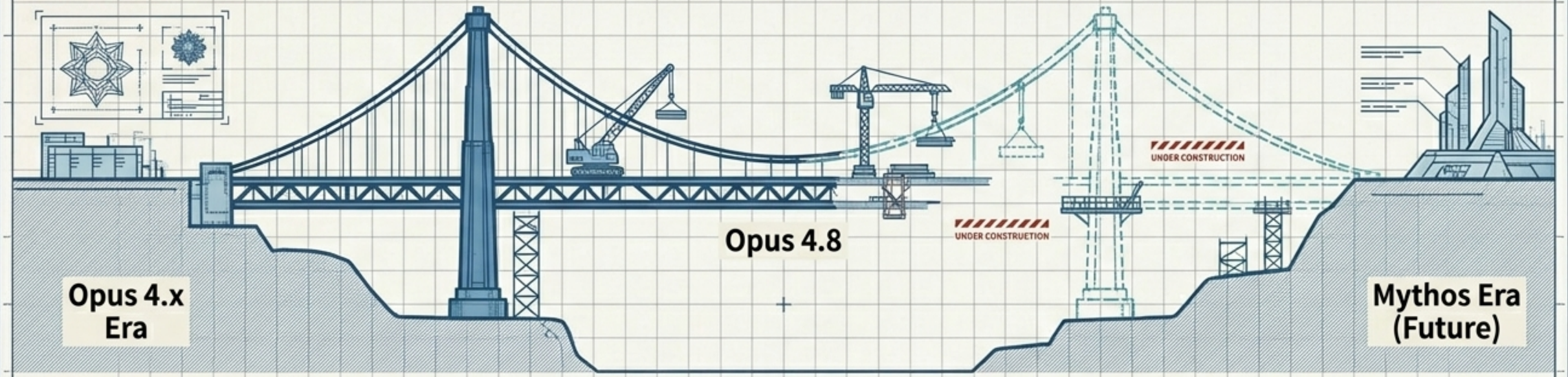
プロバイダー別提供状況：環境による機能の非対称性

Deployment Status			
Provider	**Context Limit**	**Fast Mode**	**Notes**
Claude API (Anthropic)	1M	✔ (Research Preview)	備考: 新機能の中心。
Amazon Bedrock	1M	✘	備考: Global / Regional endpoint 完備。
Vertex AI	1M	✘	備考: Multi-region / Regional 対応。
Microsoft Foundry	⚠ 200k	✘	備考: コンテキスト長に厳しい制限あり。

独立測定 (Artificial Analysis) による診断

TTFT (Time-to-first-token) は Google (7.18s) や Amazon (8.88s) が Anthropic直結 (17.95s) より高速なケースがあり、配備先で体感性能が大きく変わる点に留意。

戦略的結論：次世代への「実戦的だが未完成な橋渡し」



1. 全能の神ではない	2. Human-in-the-loop の絶対要請	3. Mythosへの布石
派手な世代交代ではなく、実務エージェントとして深く使い込込むむためのアーキテクチャ更新である。	金融・法務・リサーチ・本番環境コード修正などの高リスク領域では、モデルが優秀でも「エージェントスタックの粗さ」をカバーするため、結果検証と逐次ゲーティングを外してはならない。	「答えないことで正確さを稼ぐ」誠実さのアプローチと、劇的に安価になったFast modeは、Anthropicが目指す安全で堅牢な高自律プラットフォームの完成（Mythos級）へ向けた明確なマイルストーンである。