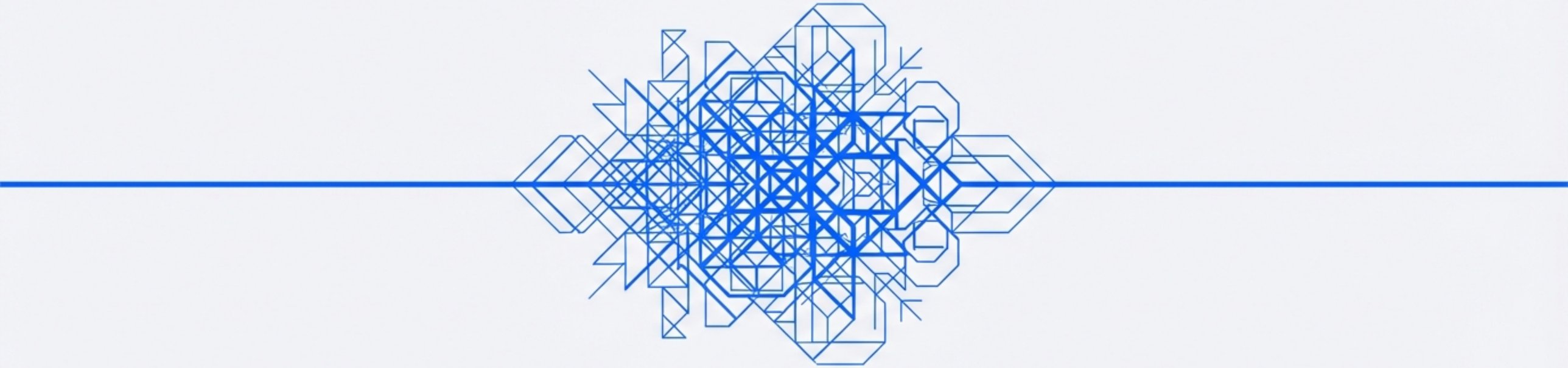
Gemini 3.1 Pro & Deep Think: 推論時間計算量スケーリングの時代へ



次世代推論モデルの全容、競合比較、そしてAGIへの到達点

2026年2月：AI開発のパラダイムシフト

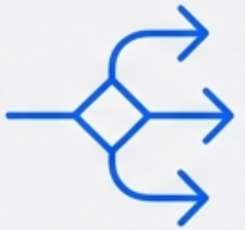
スケールから推論へ

学習時のパラメータ拡大競争は終了した。価値は「推論時間計算量 (Inference-Time Compute)」— すなわち、長く深く考える能力へと移行している。

流動性知能の獲得

ARC-AGI-2: 77.1%

従来の「記憶」に頼るモデルでは突破不可能だった壁を、Gemini 3.1 Pro が推論レイヤーの最適化により破壊。Claude Opus 4.6 (68.8%) をを凌駕。



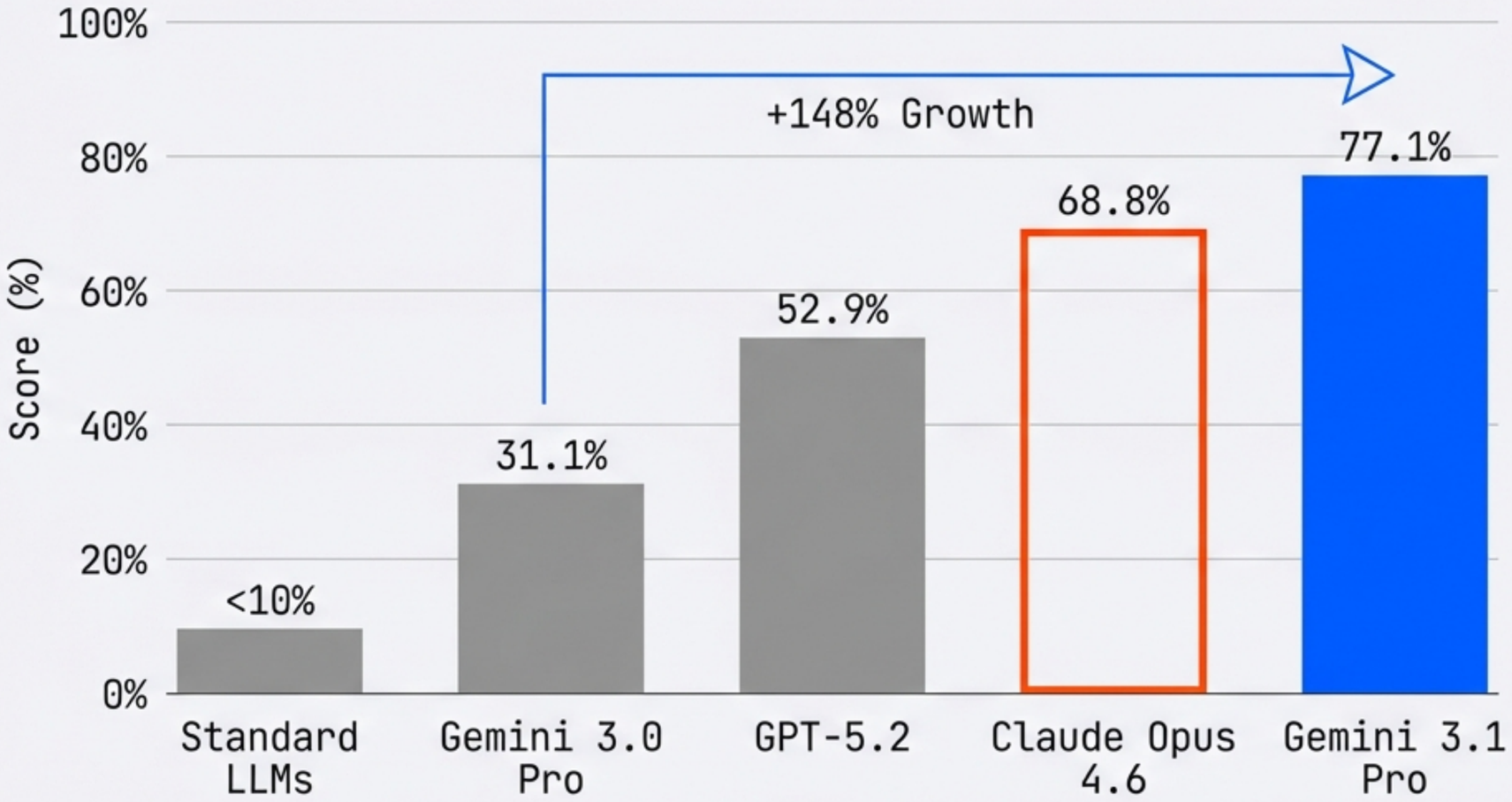
インテリジェント・ルーティング

Cost: 1/7.5 (vs Opus)

圧倒的なコスト優位性 ($2.00/1M tokens) により、タスクの90%をGeminiに、特化型タスクのみを他社モデルに振り分ける経済圏が確立。

「スケールだけでは足りない」：流動性知能の壁

ARC-AGI-2ベンチマーク推移



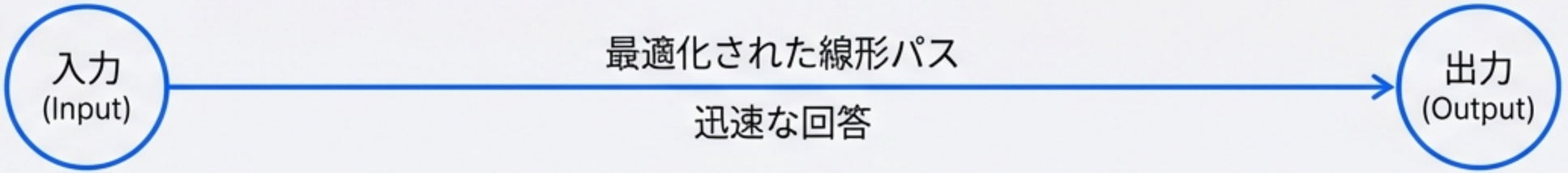
Key Insight

ARC-AGI-2は、未知の変換ルールをその場で発見する「適応力」を測定する。

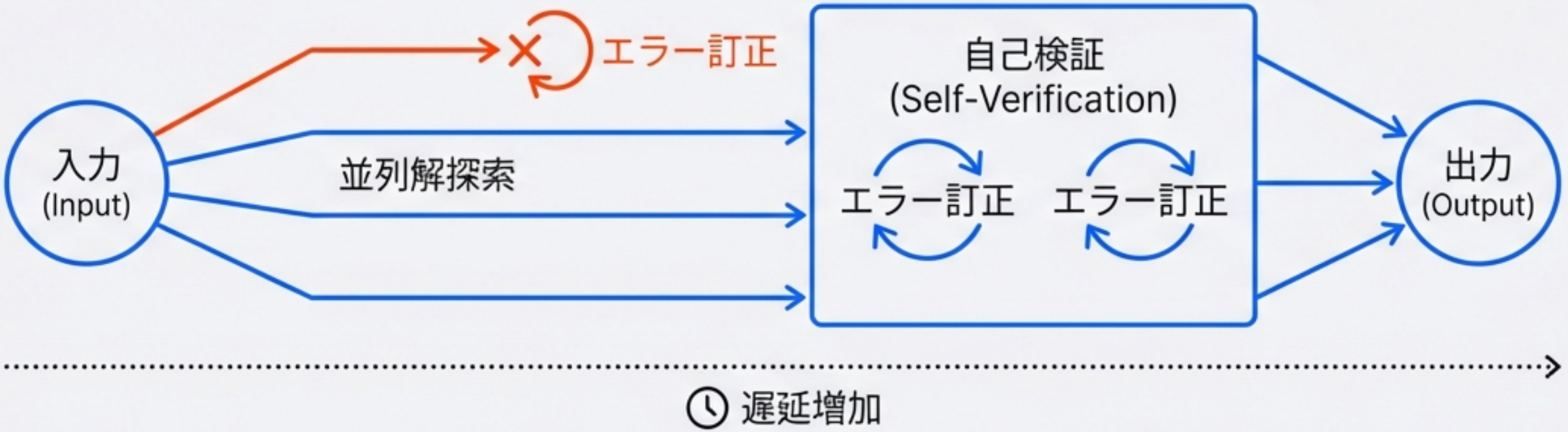
- Old World: 過去のパターンの記憶と再生
- New World: 未知の論理体系への適応

思考の構造：Deep Think アーキテクチャ

標準推論 (Standard Inference) - Gemini 3.1 Pro



Deep Think モード (Deep Think Mode) - Inference-Time Compute



Takeaway: 精度と引き換えに計算コストと時間を支払う「System 2」的思考プロセス。

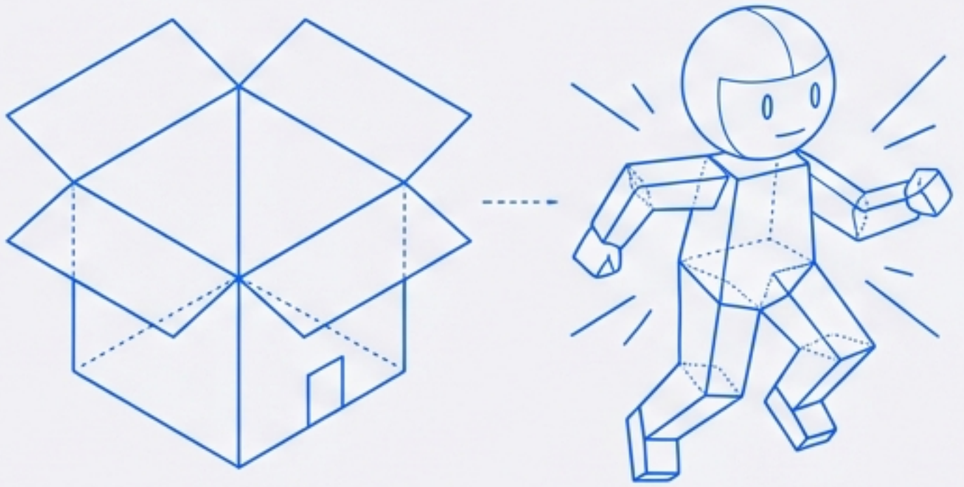
Gemini 3.1 Pro: 実用化された推論エンジン

Specs Grid

Context Window JetBrains Mono	Input 1M Tokens / Output 65k Tokens Noto Sans JP Regular
Input Media	Native PDF (100MB), Video, Audio, Code
Efficiency	出力トークン消費を10-15%削減（思考の洗練）

Feature: Vibe Coding

単なる構文ではなく、意図・雰囲気・文脈を汲み取るコーディング。



Example: テキストプロンプトから「変形する3D段ボール」や「アニメ風キャラ」のSVGアニメーションを直接生成。

JetBrains Mono, Noto Sans JP Regular

推論予算の制御：3段階の思考深度システム

高速・低コスト

- データ抽出、翻訳、単純なQA

標準品質（旧High相当）

- コードレビュー、アーキテクチャ設計

Deep Think Mini

- 複雑なデバッグ、数学的証明、自律エージェント計画

Strategic Value: 開発者は `thinking_level` パラメータで、タスクごとにコストと知能をプログラム可能。

競合分析：推論の「広さ」対「深さ」

Gemini 3.1 Pro Wins (Breadth & Orchestration)

GPQA Diamond (Science): 94.3%

LiveCodeBench: 2887 Elo

MCP Atlas: 69.2%

Geminiは「広範なシステム統合・構築」に優れる。

Claude Opus 4.6 Wins (Depth & Niche)

SWE-Bench (Bug Fix): 80.8%

Gemini is 80.6%

GDPval-AA: 1606 Elo

Claudeは「局所的な深化・専門的デバッグ」に君臨する。

Verdict: 汎用性か、専門性か。

現状の課題とアーキテクチャ 上の留意点

Terminal Operations

Terminal-Bench 2.0: Gemini（68.5%）
vs GPT-5.3 Codex（**77.3%**）

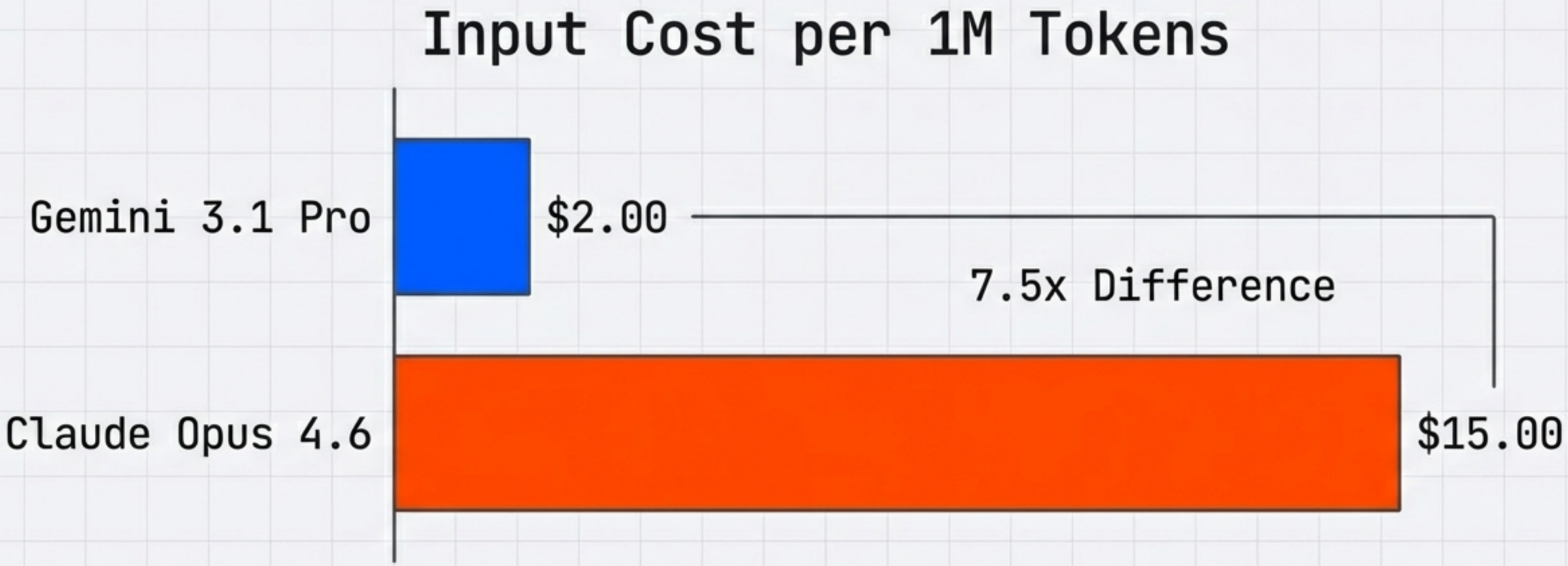
OSレベルの操作やシェルコマンドの実行では、OpenAIのCodexアーキテクチャが依然として堅牢。

Lost in the Middle

1Mトークン入力時の情報検索精度は **26.3%** まで急落（128kまでは84.9%）

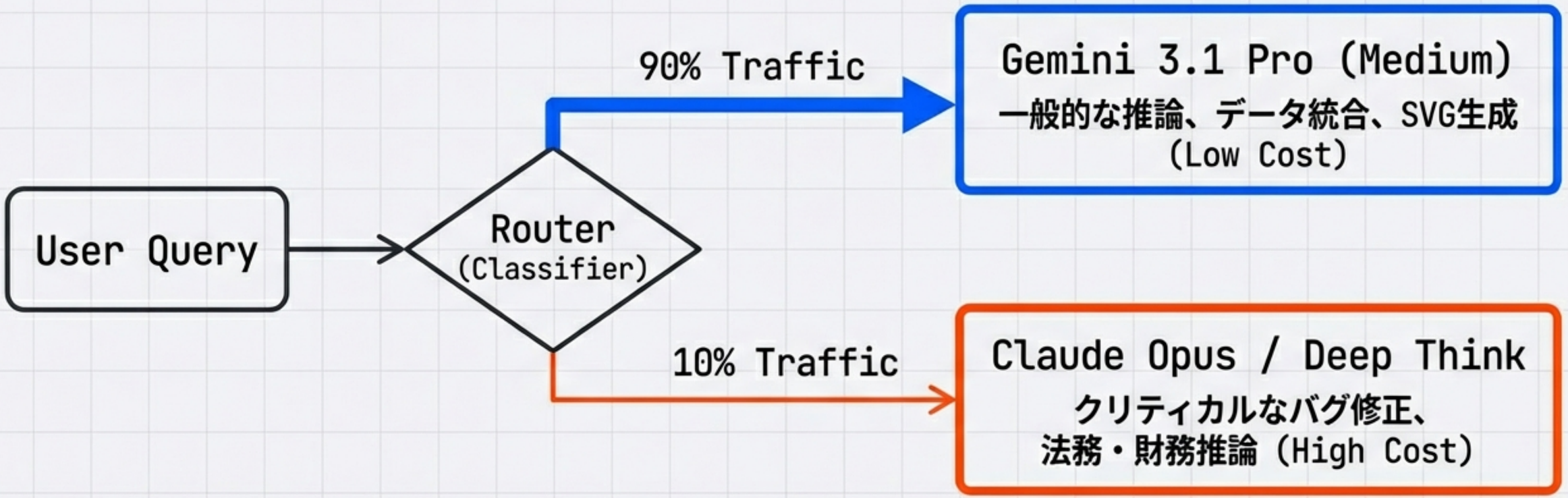
コンテキストウィンドウに依存せず、RAG（検索拡張生成）やチャンキング技術の併用が不可欠。

知能の経済性：7.5倍のコスト・アドバンテージ



高度な推論ワークフローをスケールさせるための唯一の経済的解。

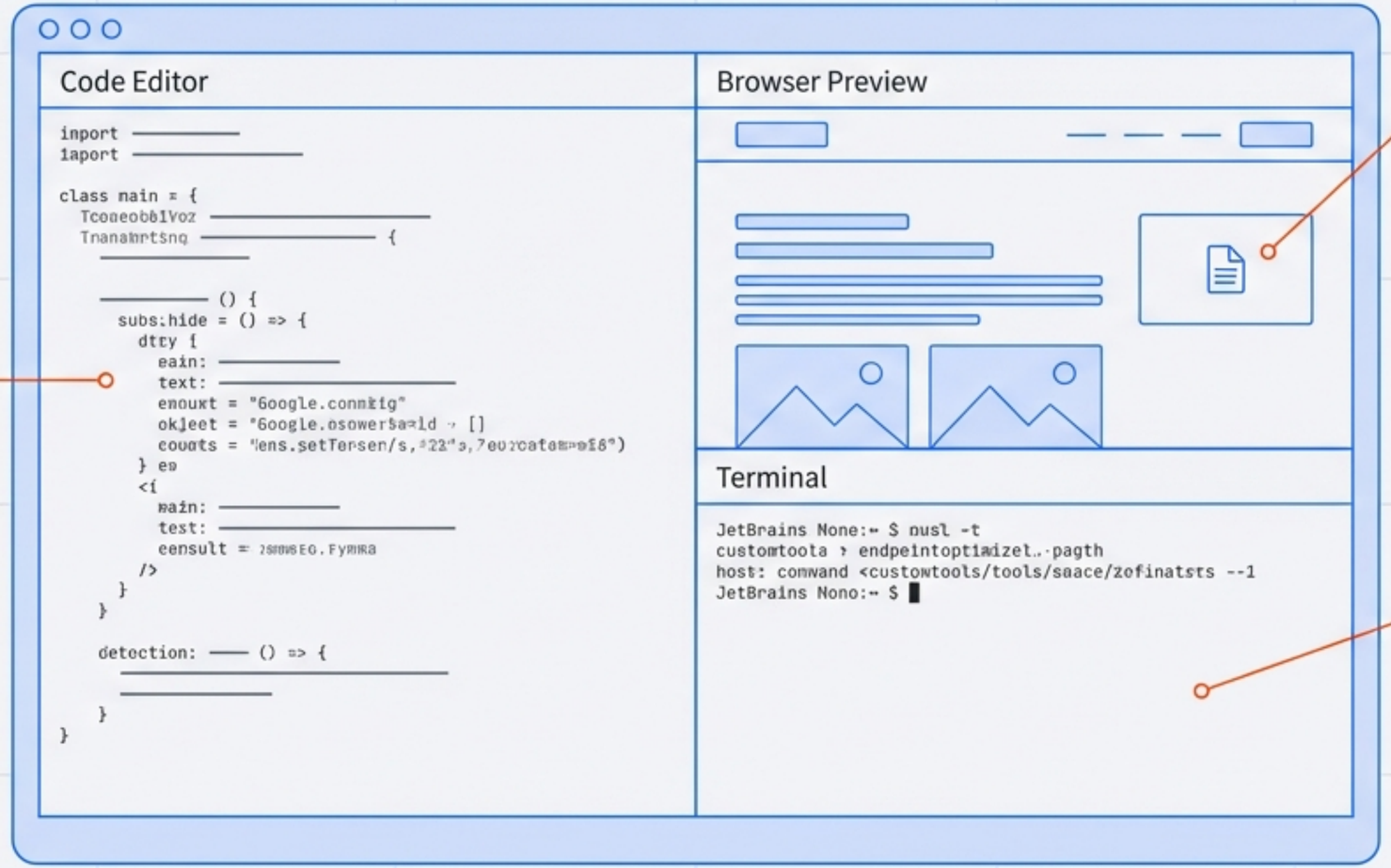
戦略的モデル・ルーティング



適材適所のハイブリッド構成が、2026年のプロダクション環境におけるベストプラクティス。

エージェント基盤：Google Antigravity

「チャット」から「自律的な仕事」へ。

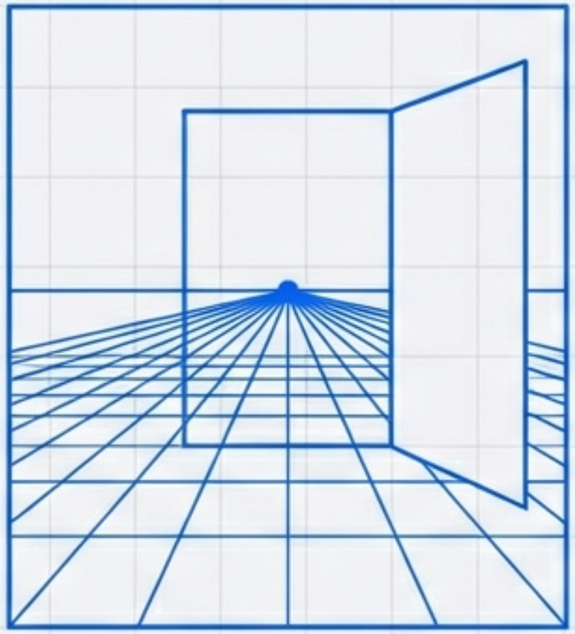


Multi-Agent Orchestration

Artifacts
検証可能な成果物を生成

Custom Tools
`customtools`
endpoint optimized

結論：思考するAIと共に歩む



1.

Paradigm Shift:
スケール競争から「推論時間計算量」の時代へ。

2.

The Engine:
Gemini 3.1 Proは、コスト・速度・性能のバランスにおいて現時点の最適解。

3.

Action Plan:
モデル・ルーティング、思考深度制御、Antigravityの導入。

AIはもはや「検索する道具」ではない。
「共に考え、問題を解決するパートナー」である。

Appendix / Sources

1. **Google DeepMind Technical Reports (Feb 2026)**
2. **ARC-AGI-2 Benchmark Results**
3. **Digital Applied: Gemini 3 Deep Think Analysis** (JetBrains Mono)
4. **Terminal-Bench 2.0 & SWE-Bench Verified Datasets** (Noto Sans JP Regular)
5. **Google Cloud Antigravity Documentation** (JetBrains Mono Regular)