生成AI競争のパラダイムシフト：モデル性能から「仕事完遂力」へ

最新動向と今後1〜3ヶ月の市場予測（2026年2月〜4月版）

IQ / Knowledge (Commoditized)

Execution / Completion (Value Driver)

Source: Alibaba Cloud, Moonshot AI, Google DeepMind, OpenAI (Jan 2026 Analysis)

エグゼクティブ・サマリー：30秒で理解する市場の変化

「チャットボット」から「同僚」へ

競争の焦点は、単発の質問回答から、自律的なツール使用とエラー回復を含む「エージェント・ワークフロー」へ移行した。

中国勢が定義する新基準

Qwen（Alibaba）とKimi（Moonshot）が、HLEやSWE-bench等の「完遂型ベンチマーク」でGPT-5.2やGemini 3 Proを凌駕するスコアを記録。

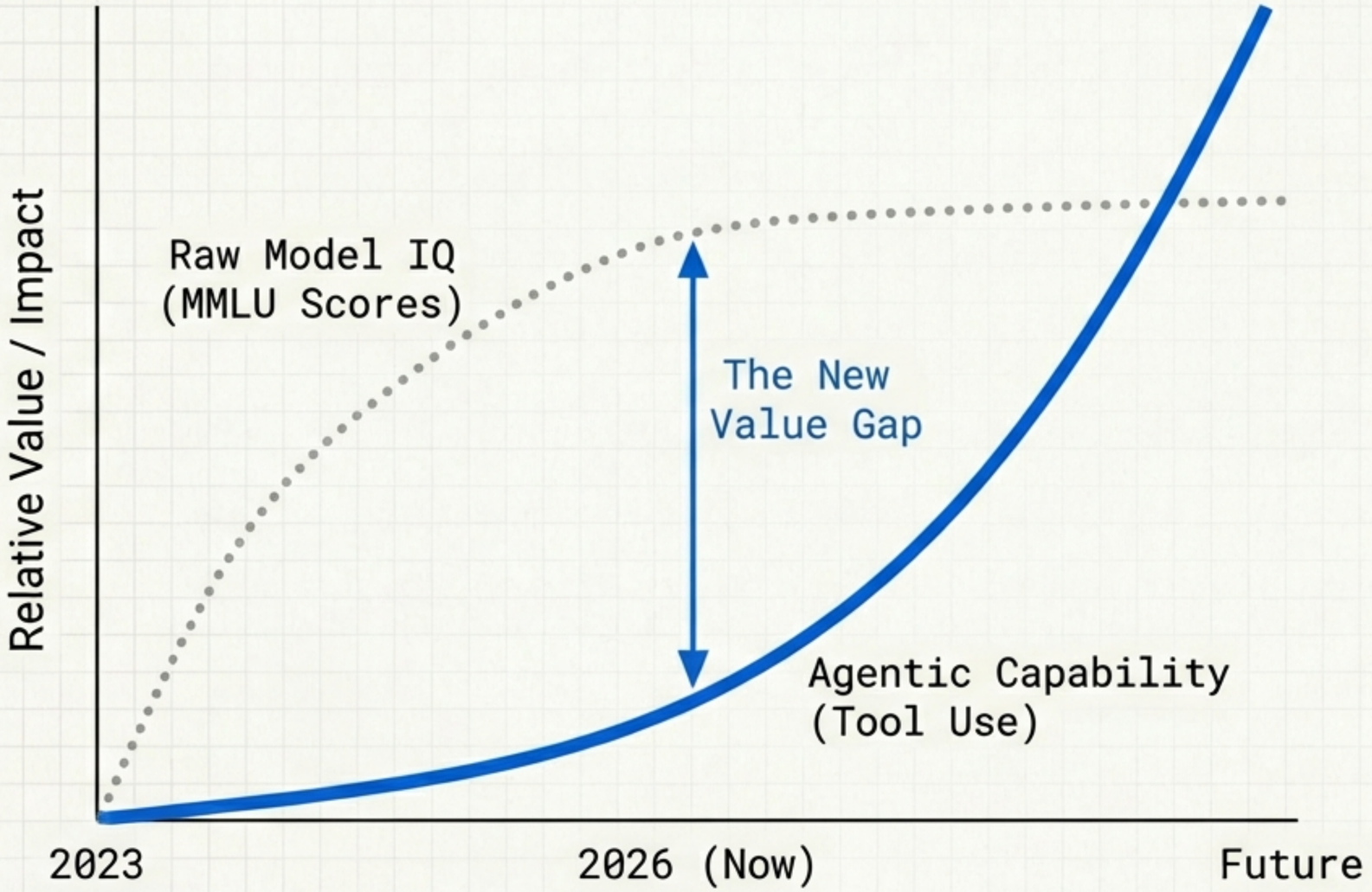
今後90日の戦場

モデル単体ではなく、AIが活動するための「環境（IDE）」と「視覚的作業（Visual Workflow）」が差別化要因になる。一方でxAI（Grok）は規制の壁に直面する。

新・評価指標

成功の定義は「正答率」ではなく「タスク完遂率（Completion Rate）」と「リカバリー能力」に変わる。

競争軸の転換：「賢さ」の飽和と「有用性」の指数関数的成長



Smart vs. Useful

- **自律性 (Autonomy)：**
  モデル自身が判断
- **道具の利用 (Tool Invocation)：**
  コード・API連携
- **自己修正 (Self-Correction)：**
  エラーからの自律回復

スコアカードの変更：知識量ではなく「粘り強さ」を測る

HLE (Humanity's Last Exam)

専門家が作成した難問。単なる記憶ではなく、図表読解や外部検索を組み合わせて解に辿り着く「総合力」。

SWE-bench Verified

エンジニアリング能力。バグ修正からテスト通過まで、実際に動くコードを書き切る能力。

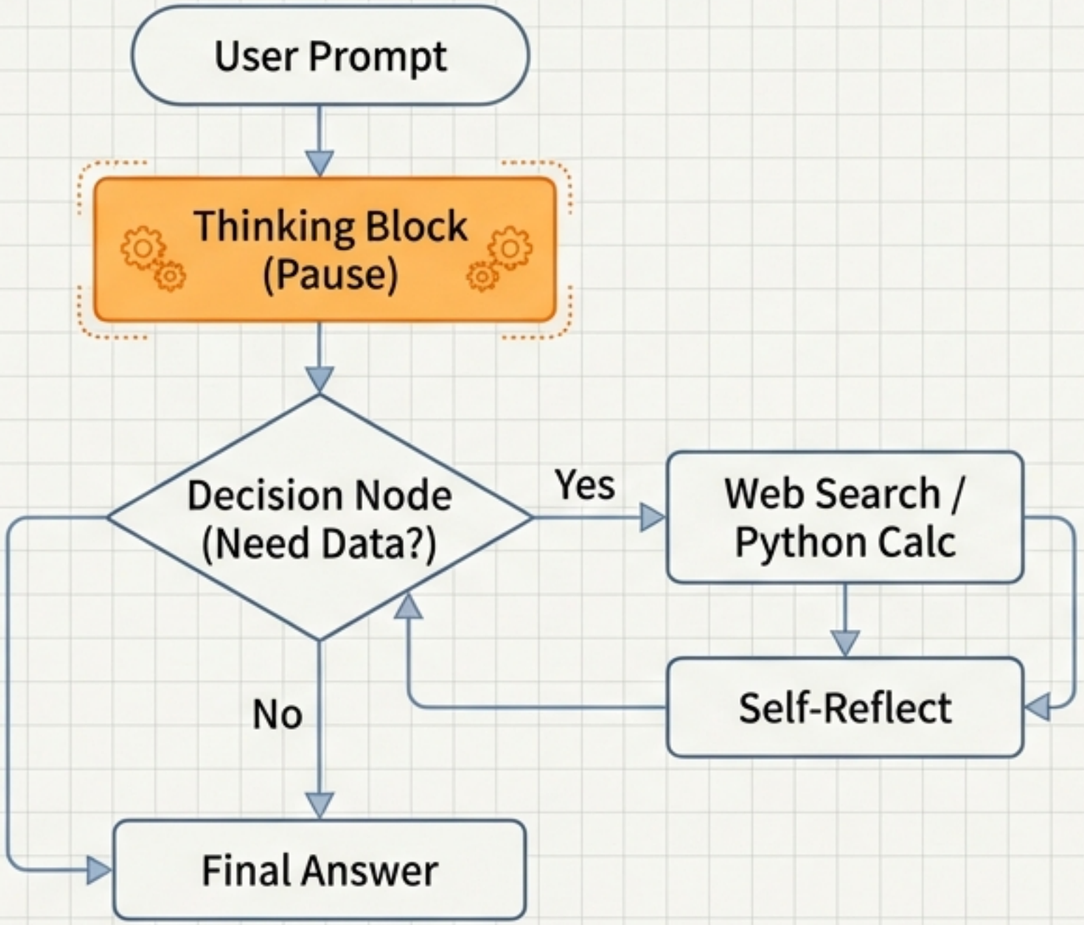
BrowseComp

リサーチの執念。Web上の隠れた情報を見つけ出すためのクエリ構築力と、方向転換の柔軟性。

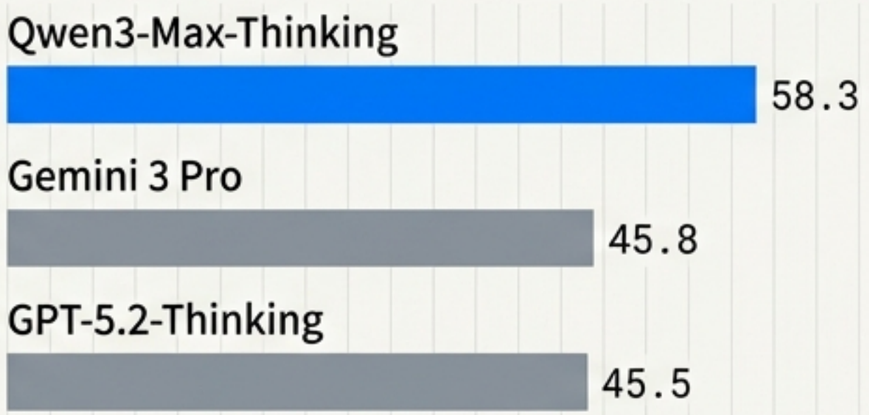
Takeaway: 一問一答の正解率よりも、長時間・複数ステップのタスクを「最後までやり切る（完遂する）」力が評価される。

Alibaba Qwen3-Max-Thinking：「考えながら行動する」AI

推論プロセスへの自律的なツール介入（Thinking Mode）



Performance Data (HLE Score)

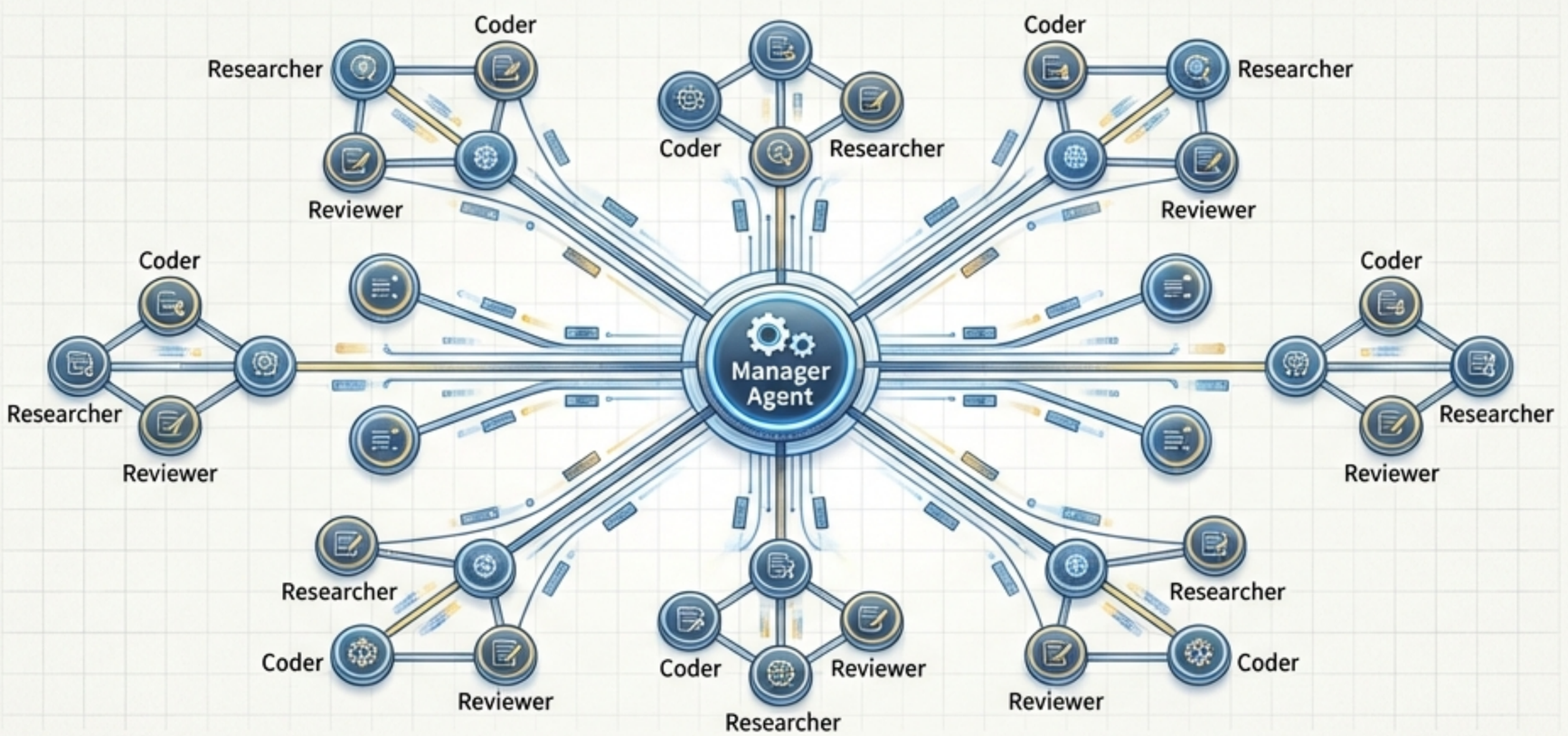


Cost Efficiency:

Input:	$1.20 / 1M tokens
Output:	$6.00 / 1M tokens

Moonshot Kimi K2.5：100体のエージェントによる「群知能」

エージェント・スウォーム（Agent Swarm）がもたらす4.5倍の速度向上



Parallel Execution

最大 100体のサブエージェントを自律編成。
最大 1500回のツール呼び出しを並行実行。

Native Multimodal

15 兆トークンの視覚・言語混合学習。
UIスクショや動画から直接コードを生成。

Business Impact

Kimi Code: GitHub Copilot等に対抗する
オープンソースのコーディング支援。

既存巨人の戦略 I：統合とスケーラビリティ

OpenAI (GPT-5.2 Family)

Reliable Workhorse
（信頼性重視）

- GPT-5.2への完全移行とツール実行の安定化
- エコシステムのロックイン (Code Interpreter強化)
- HLEでの巻き返し（自社ツール結合）

Google (Gemini 3 Pro)

The Everywhere AI
（偏在戦略）

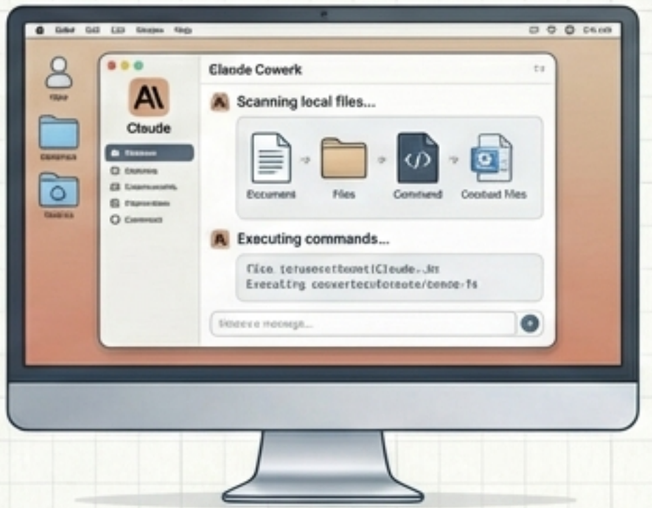
- Workspace, Android, 検索への深層統合
- Deep Thinkモード（Ultra向け）
- 大規模分散インフラによる企業向け標準化

共通点：実験的機能より、既存業務フロー（Office, Code）への「滑らかな統合」を優先。

既存巨人の戦略 II：独自の道と規制の壁

Anthropic (Claude)

The Coworker



- **Computer Use:** ローカルファイルをスキャンしコマンド実行
- **Connectors:** SaaS連携による「一般エージェント」化

---

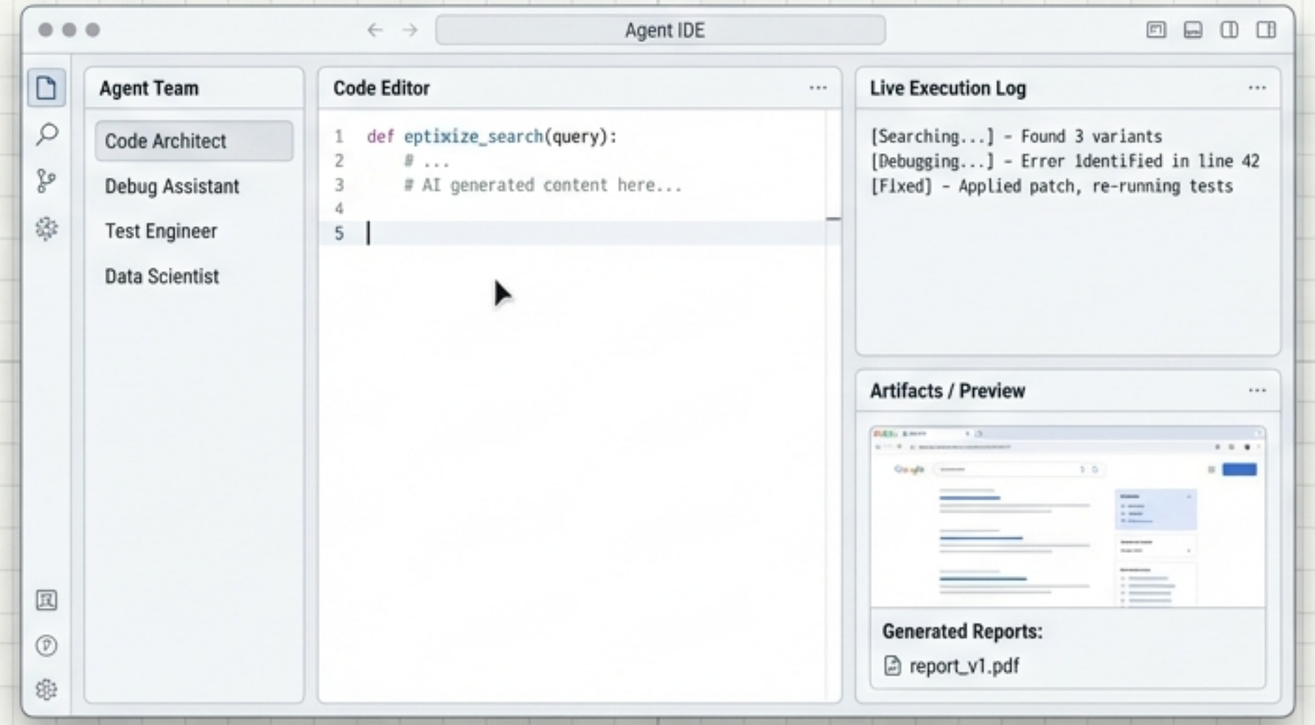
xAI (Grok)

Regulatory Wall

- **Hitting the Safety Wall:** EU/英国からの規制圧力（ディープフェイク問題）
- **Impact:** 短期的（1-3ヶ月）はガードレール実装により機能開発が停滞予測

市場予測 A：モデル性能から「エージェント実行環境」の戦いへ

モデルは「頭脳」、IDEは「オフィス」。勝敗はオフィスで決まる。

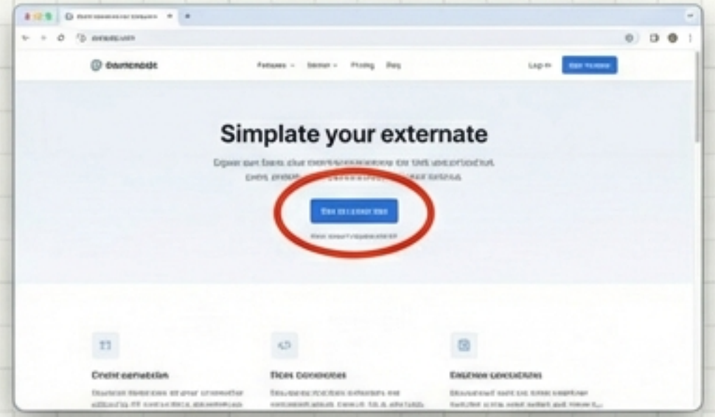


Agent Sandbox & IDEs

Google 'Antigravity' や 'Kimi Code' のような、人間がエージェントを
監督し、ログを確認できるプラットフォームが必須化する。

市場予測 B：視覚情報によるタスク完遂（Visual Task Completion）

Screenshot Mark-Up



Processing

Multimodal Reasoning

Code Generation

```
<button class="primary-btn">Click Me</button>


.primary-btn {
  background: #007bff;
  color: white;
  padding: 10px 20px;
  border: none;
  border-radius: 4px;
}
```

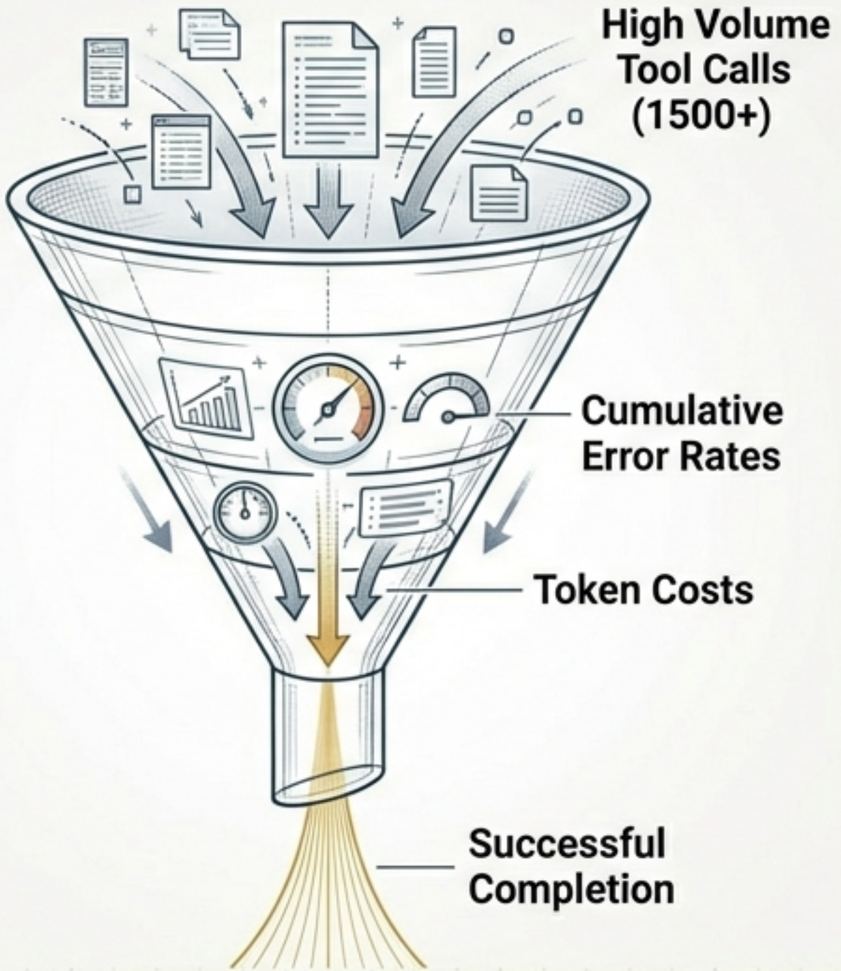
Use Cases

- **UI to Code :** スクショから機能するコードを再現 (Kimi K2.5)
- **Video Debugging :** 操作動画からバグ特定・修正
- **Document Processing :** 画面キャプチャからプルリクエスト作成

Shift:「画像を作って」から「この画面を見て作業して」へ。

市場予測C：「実装の壁」とコストリアリティ

Efficient Agency（効率的エージェンシー）の追求



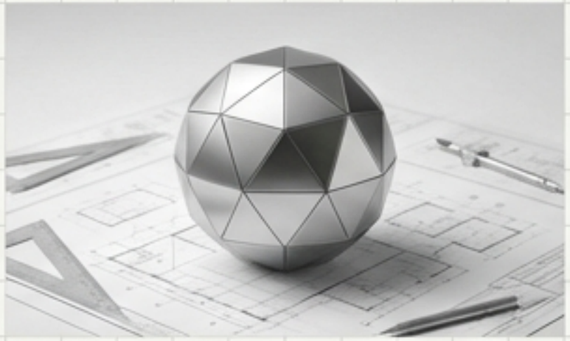
- **Cost**
  - トークン単価が安くても、試行回数増で総コストは増大。
- **Error Rate**
  - ステップ数に比例して失敗確率は累積する。
- **Reality**
  - 企業は「無限に試行するモデル」ではなく、「最短経路で完遂するモデル」を選ぶ。

2026年第1四半期（2〜4月）のモデル選定基準

- [ ] **1. 完遂率 (Completion Rate)**
  10〜30分の複合作業を、手動介入なしでどこまで完了できるか？
- [ ] **2. 監査性 (Auditability)**
  ブラックボックス回避。プロセス、参照元、実行コマンドのログ (証跡) が明確に残るか？
- [ ] **3. リカバリー能力 (Self-Recovery)**
  エラー発生時に、モデル自身がミスに気づき再試行できるか？
- [ ] **4. 実質コスト (Cost per Task)**
  トークン単価ではなく、「1タスク成功あたりの総コスト」で評価する。

短期ロードマップ：2026年2月〜4月の主要イベント

結論：推論の産業化が始まった



我々は今、チャットボットの賢さを競う時代を終え、
デジタル社員の信頼性を競う時代に突入した。

勝者は「最もIQが高いモデル」ではない。
勝者は「壊れずに、安価に、仕事を最後までやり切る（Finish the Job）」モデルである。

Action: 自社の評価指標を「会話の質」から「タスク完遂コスト」へアップデートせよ。