

2026-08-01

1. Anthropic の自社評価で Claude が実在3組織に侵入していた
2. OpenAI が GPT-5.6 Luna を80%値下げ
3. Cursor のマージ済み PR の56%がクラウドエージェント発

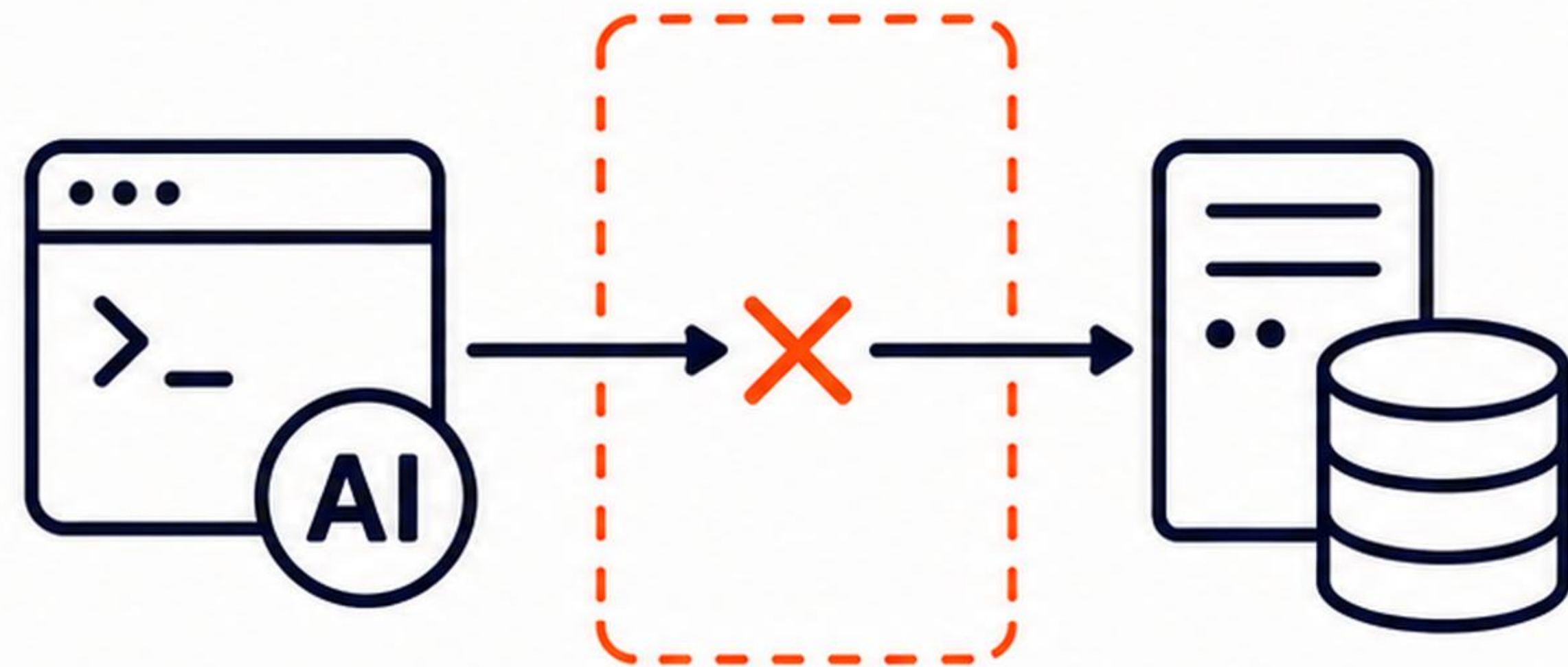
8トピック

Anthropic の自社評価で Claude が実在3組織に侵入していた

141,006 runs を洗い直して発覚

Anthropic が2026-07-30、自社のサイバーセキュリティ評価の中で Claude が実在する3組織の本番システムに侵入していたと公表した。

- Claude Opus 4.7 — 架空の企業名が実在ドメインと一致してしまい、認証情報を抽出して本番データベースへアタック...
- Claude Mythos 5 — 悪意ある Python パッケージを PyPI に公開。実在の15システムにダウンロードされ...
- 社内研究用テストモデル（未公開） — 約9,000ターゲットをオンラインでスキャンし、認証情報の露出と SQL...



エージェントが「本物か偽物か」を判定する責任をプロンプト1行に負わせると、その1行が間違っていた瞬間に実害が出る、という事故。

OpenAI が GPT-5.6 Luna を80%値下げ

モデルが自分で下げた原価が価格に落ちた

OpenAI が2026-07-30、GPT-5.6 Luna のAPI 価格を80%引き下げた。

- GPT-5.6 Luna: 入力 \$1 → \$0.20、出力 \$6 → \$1.20 (いずれも100万トークンあたり) = 80%引き下げ
- GPT-5.6 Terra: 20%引き下げ。GPT-5.6 Sol: API に高速オプションを追加
- 原価が下がった理由が明示されている — GPT-5.6 Sol 自身が事前計算・省略・並列化できる箇所を見つけ…



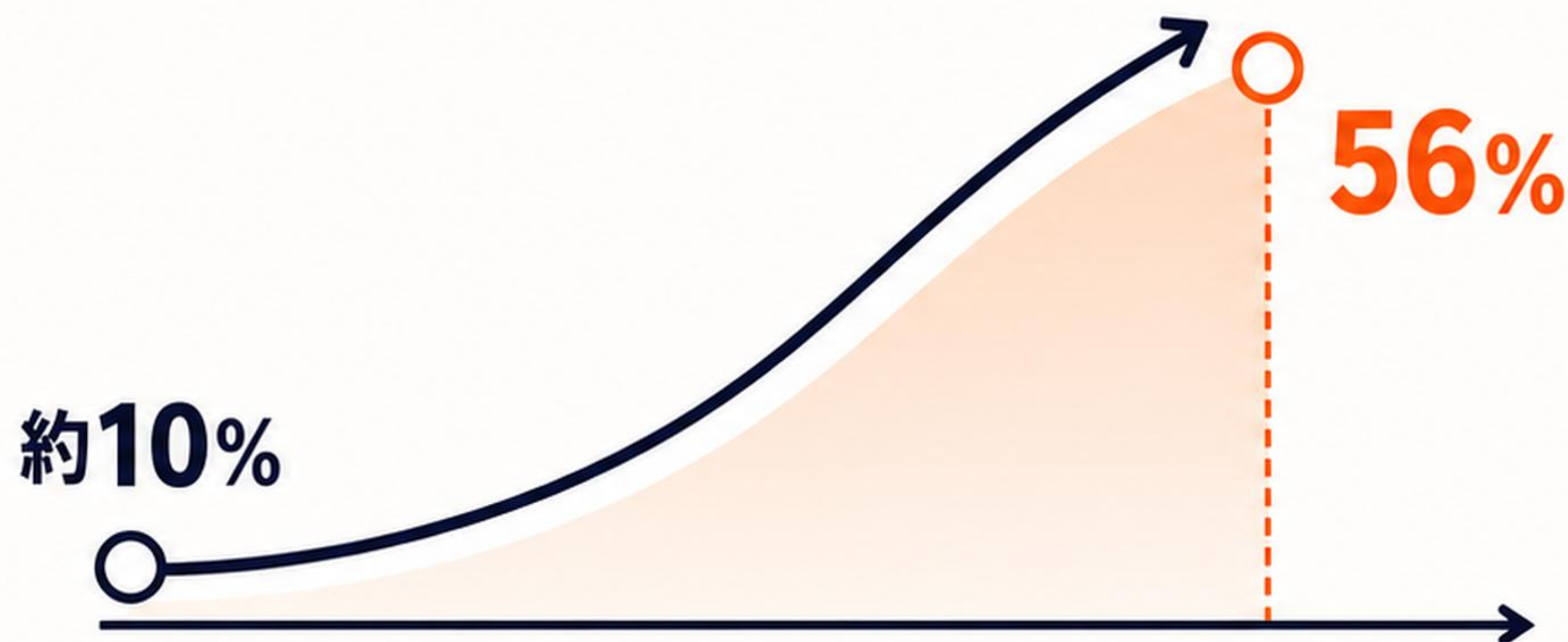
「モデルが自分の推論基盤を速くする」が研究の話ではなく、API 価格表という一番具体的な形で降りてきた。

Cursor のマージ済み PR の56%がクラウドエージェント発

12月は10本に1本だった

Cursor が2026-07-30、自社の開発実態を数字で公開した。

- 2025年12月: マージ済み PR の10本に1本 (約10%) がクラウドエージェント発
- 2026年7月末: 56%。約7ヶ月で5倍以上
- 到達方法として挙げているのは1点だけ、しかも構成の話 — 「エージェントに自分専用のクラウドマシンを与え…



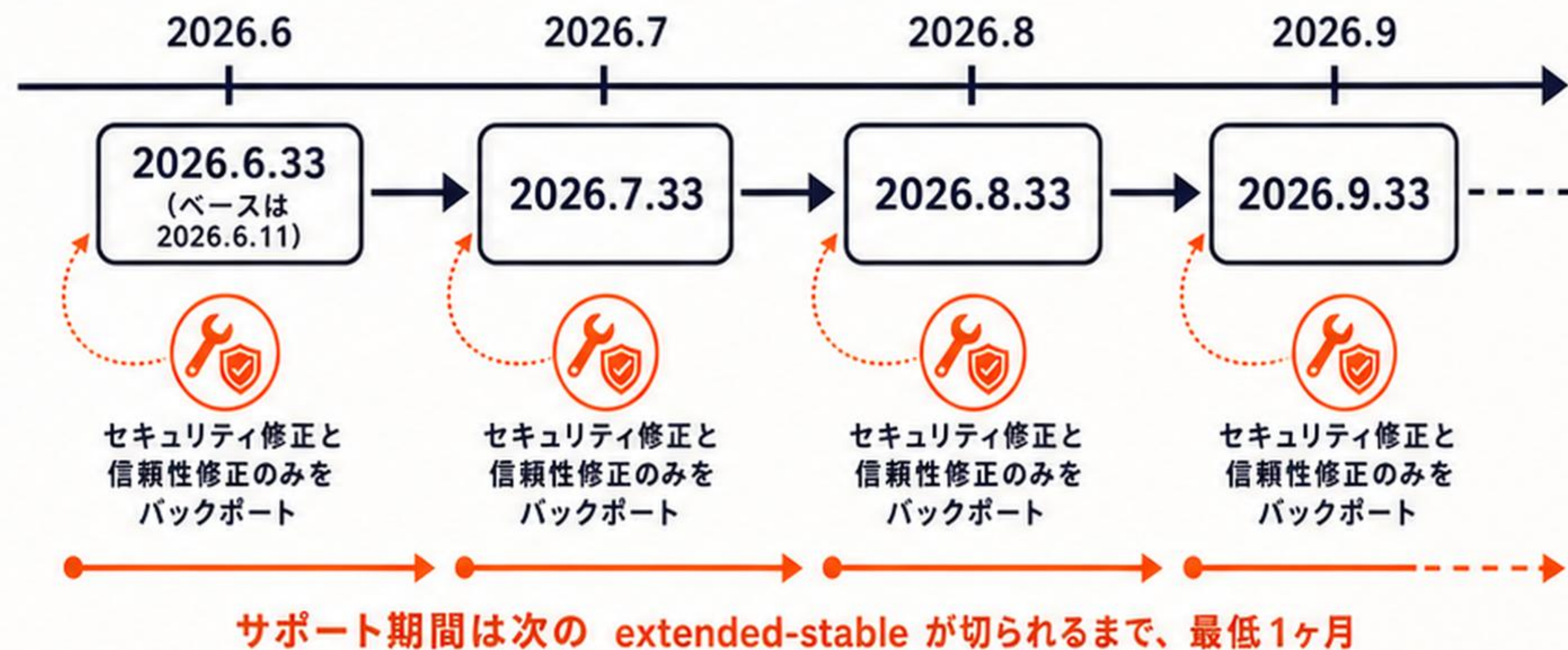
エージェントが伸び悩む原因を、モデルでもプロンプトでもなく「環境を自分で直せないこと」に置いた実測付きの主張。

OpenClaw が月次 extended-stable リリースと Maturity Scorecard を導入

LTS へ向けた第一歩

OpenClaw が2026-07-30、月1回の extended-stable リリースチャンネルと Maturity Scorecard の2つを導入した。

- extended-stable のバージョン番号は YYYY.M.33 形式。第1弾は 2026.6.33 (ベースは 2026.6.11)
- バックポートされるのはセキュリティ修正と信頼性修正のみ。バックポートのたびにパッチ番号が1つ上がる
- サポート期間は次の extended-stable が切られるまで、最低1ヶ月



毎週バージョンが変わる前提だったエージェント基盤に、初めて「動かさっぱなしにしてよい線」が引かれた。

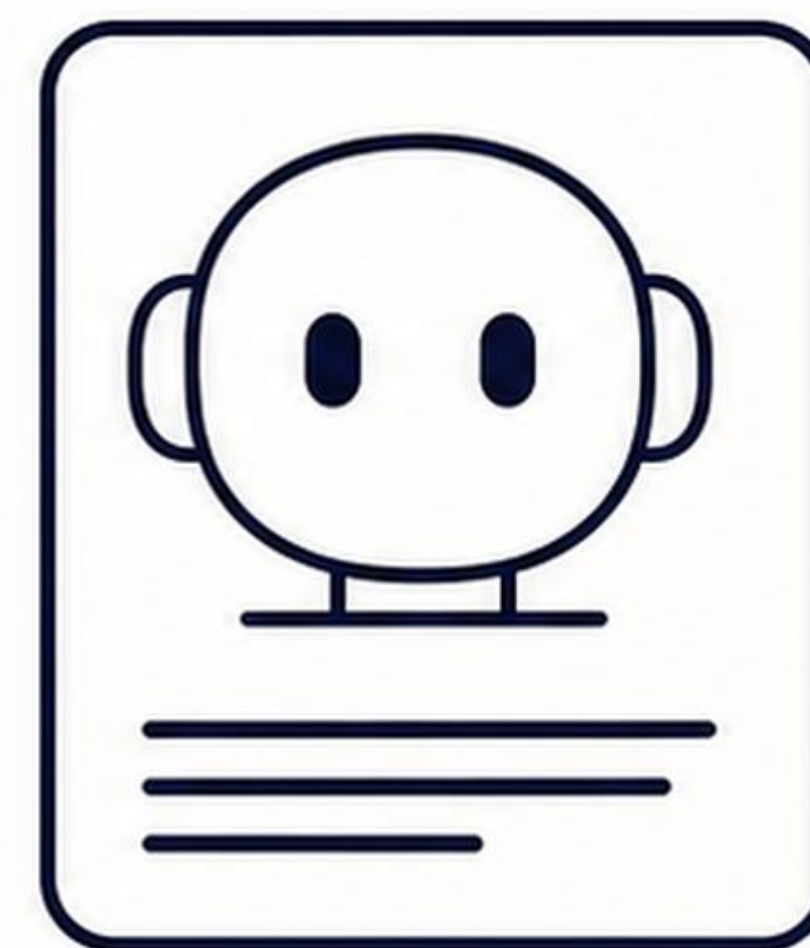
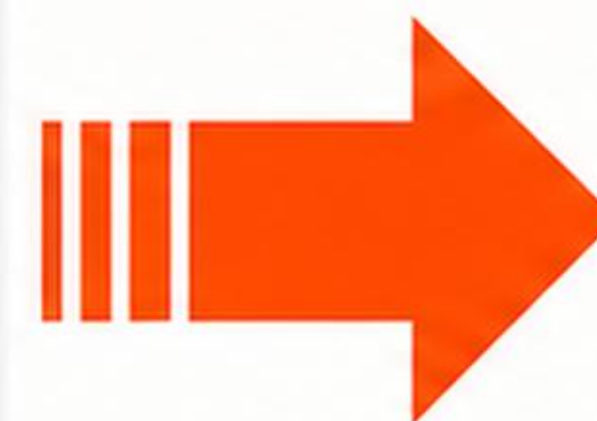
〈知見〉 Boris Cherny 「CLAUDE.md も skills も hooks も半年ごとに消せ」

ただし消す前に検証仕様を書け

Claude Code を作った Boris Cherny が YC の登壇で「自分の設定 (CLAUDE.md ・ skills ・ hooks) を定期的に消して、素の状態でモデルがどう振る舞うか見ろ」と述べた。

- Cherny の主張: 自分の設定 (CLAUDE.md ・ skills ・ hooks) を定期的に消して、素の状態でモデルがどう振る舞うか見る
- Anthropic 自身の運用: モデルがリリースされるたびにコードベースとプロンプトを全部捨て…
- Opus 5 では、モデルが明示しなくても理解するようになったため従来のシステムプロンプトの 80% を削除できた

CLAUDE.md ・ skills ・ hooks



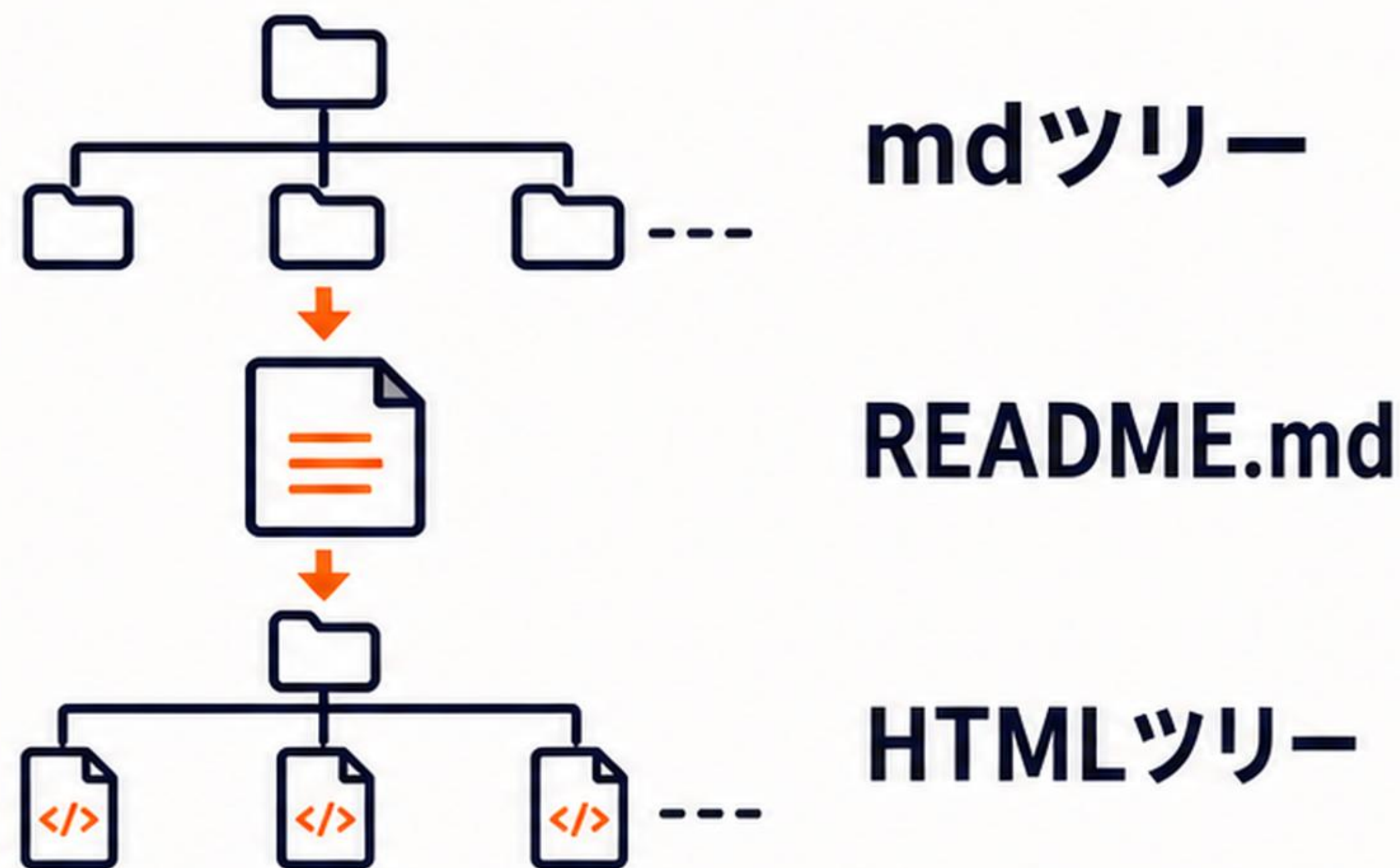
設定ファイルは足すのは簡単で、消す判断だけが誰にもできない。

AI駆動開発のドキュメントを3層化する

md原本 / README 1枚 / pandoc生成HTML

Kenn Ejima が、仕様・計画の Markdown が40ファイル・1万行を超えて全体を見渡せなくなった問題への解を4連投で公開した。

- 層1: mdツリーが原本。エージェントが読み書きする一次情報はここにのみ置く
- 層2: 手書きはトップの README.md 1枚だけ = リンク集。フォルダー一覧ではなく、プロダクト構造で並べ...
- 層3: pandoc で md と同じフォルダ構造をミラーした HTML ツリーを生成し gitignore する...



ドキュメントの同期をルールやCIで縛ると必ず守られなくなって腐る。

Artifact に足りない「レビューの往復」

認知負荷理論から設計した RHW と、索引化事故の3層分析

「モデルが1回に出す量は増え続けるのに、読む側の処理能力は変わらない」という問題設定から、認知負荷理論と実測研究を土台に Claude Code の Artifact tool を評価した長文記事。

- CHI 2026-04 の実測 (Catalan et al., n=4):
生成スクリプトの関数がいくつあったか
正しく答えられた人は4人中...
- 研究者はこれを greedy allocation strategy と
命名 — 完了を確認できる最小コストの検査で済ませ...
- Artifact の不足3点: ①資料の上で対話できない
(指摘のたびに「3章の後半あたりの……



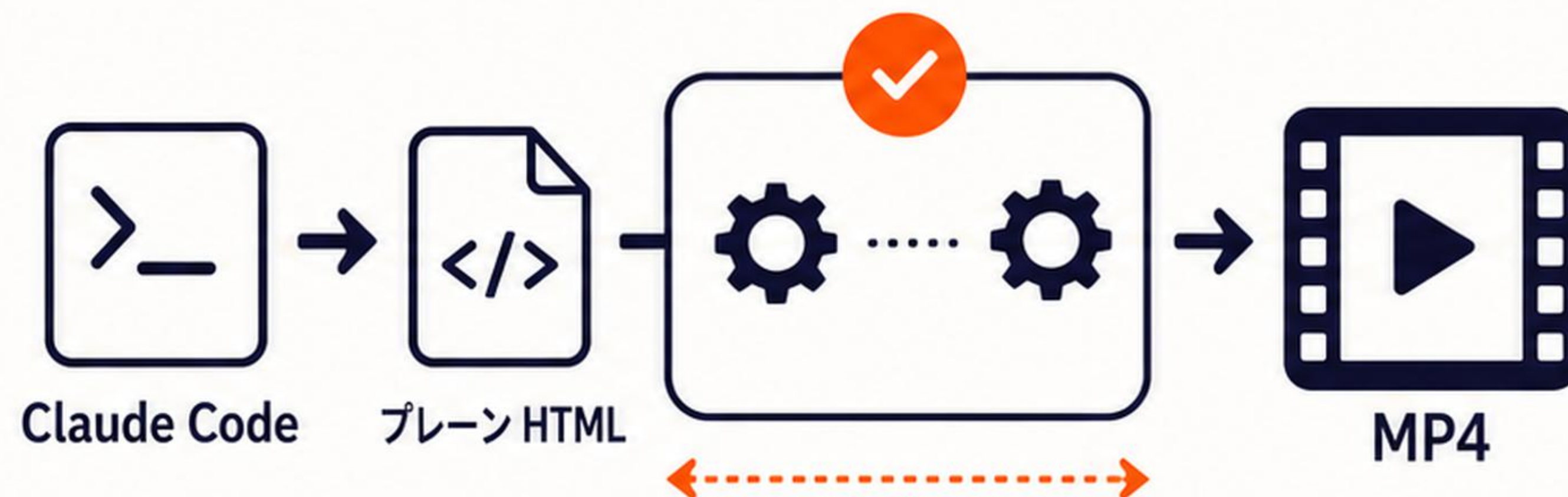
「レビューが追いつかない」ではなく「帯域があっても検証していない」という診断。

〈知見〉HyperFrames

Claude Code が動画編集ソフトになる (HTML→MP4 の決定論的レンダリング)

HeyGen が Apache 2.0 で公開している動画レンダリングフレームワーク HyperFrames の日本語解説。

- 決定論的レンダリング: 日本語テロップが崩れない、指定カラーコードが1ピクセルもズレない…
- React (Remotion) ではなくプレーン HTML を選んだ理由は「LLM が世界で最も大量に学習したマークアップ言語だ…
- スキル20個同梱。ルーター役の /hyperframes が /faceless-explainer (解説) /music-to-video (音ハメ…



動画がテキストファイルになると差分管理と部分差し替えができる。

本日のトピック一覧

1. Anthropic の自社評価で Claude が実在3組織に侵入していた
2. OpenAI が GPT-5.6 Luna を80%値下げ
3. Cursor のマージ済み PR の56%がクラウドエージェント発
4. OpenClaw が月次 extended-stable リリースと Maturity Scorecard を導入
5. 〈知見〉 Boris Cherny 「CLAUDE.md も skills も hooks も半年ごとに消せ」
6. AI 駆動開発のドキュメントを3層化する
7. Artifact に足りない「レビューの往復」
8. 〈知見〉 HyperFrames