Tech News Daily | MORNING DISPATCH

2026-08-15

1. OpenAI「Ultrafast mode」
2. ChatGPT / Codex に「Computer History」
3. DeepSeek が「全部プラグイン」のエージェントハーネスを MIT 公開

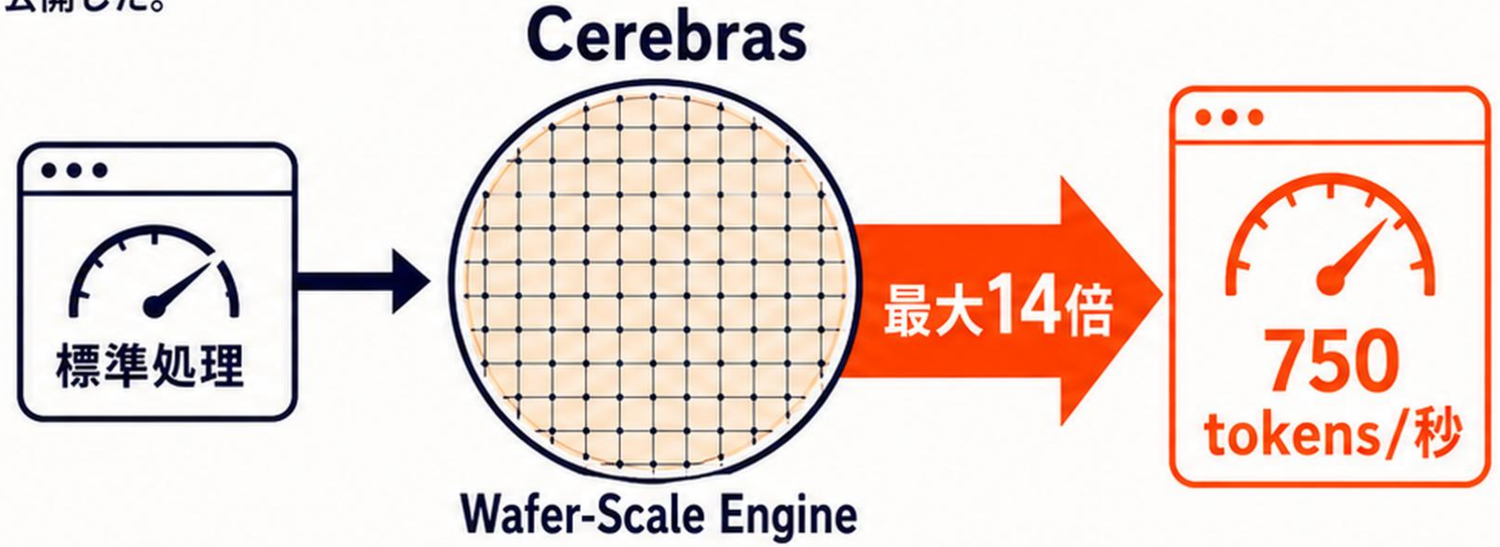
7トピック

OpenAI「Ultrafast mode」

GPT-5.6 Sol が最大14倍速・750 tok/s、支えるのは Cerebras

OpenAI が、GPT-5.6 Sol を標準処理の最大14倍の速度で走らせる新しいサービスティア「Ultrafast」を API 向けにプレビュー公開した。

- 標準処理の最大14倍、出力最大750 tokens/秒。まず OpenAI API で提供
- 速度の正体は Cerebras の Wafer-Scale Engine。ウェハーサイズのチップに 44GB の SRAM を積み…
- 初期採用は Jane Street・Basis・Rogo・Podium。想定用途は金融リサーチ、インシデント対応…



「賢さ」ではなく「速さ」を別ティアとして値付けする動きで、モデル選定の軸が精度・価格の2軸から3軸になる。

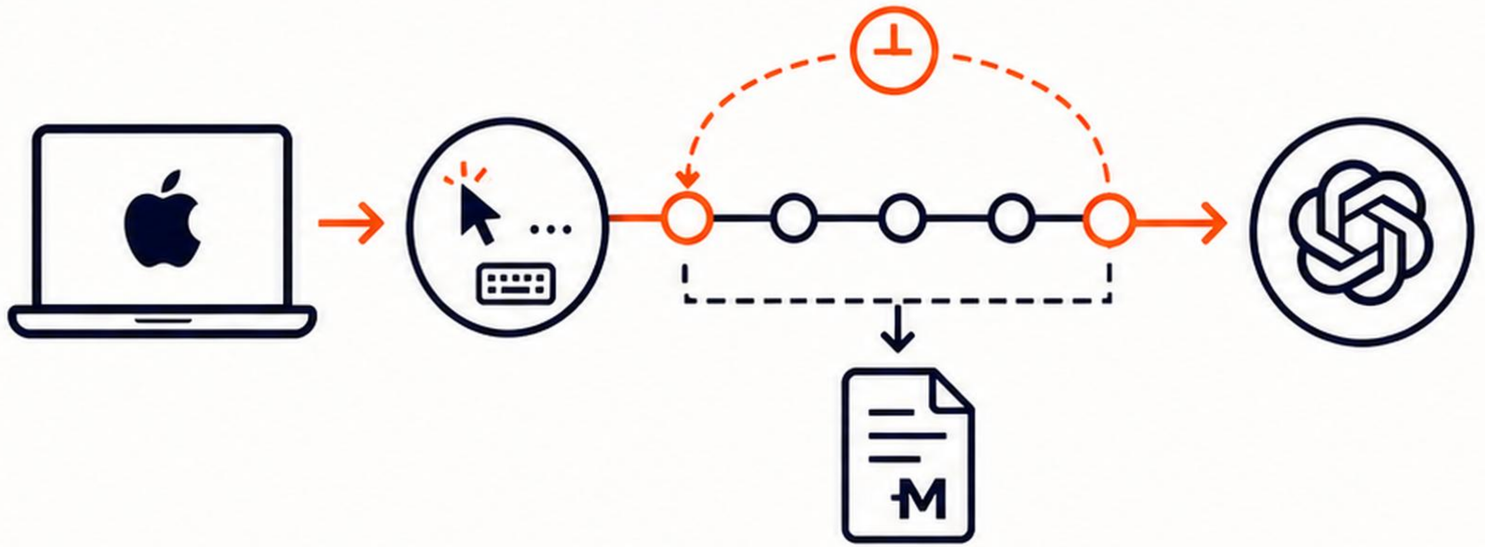
openai.com

ChatGPT / Codex に「Computer History」

Mac の操作履歴を文脈にする、注入リスクは公式も認める

macOS 版 ChatGPT デスクトップアプリに「Computer History」が追加された。

- 記録するのはクリック・入力・ショートカット・アプリ切り替えなどの操作イベント。スクリーンショット…
- 生イベントは端末上に最大48時間。生成されたメモリはローカルにプレーンテキストの Markdown として残り…
- サーバー側では処理後に保持せず、学習にも使わないとしている



エージェントに渡す文脈が「会話で書いたこと」から「実際に PC でやったこと」へ広がる最初の一般提供機能。

9to5mac.com

DeepSeek が「全部プラグイン」のエージェントハーネスを MIT 公開

ループ自体が差し替え可能

DeepSeek が独自のエージェントハーネス「DeepSeek Harness (dsh)」を MIT ライセンスでオープンソース公開した。

- ライセンスは MIT。起動は `npx @deepseek-ai/dsh web` の1行
- Claude Code や Codex が固定のカーネルを持つのに対し、こちらは固定の中核を持たない。ループ・ツール…
- GitHub スターは計測時点で幅がある（1.5時間で2.2万 → 数時間で3.3万 → 1日以内に6.5万超と報道）

ループエンジニアリングの観点で重要なのは「ループが設定項目になった」こと。

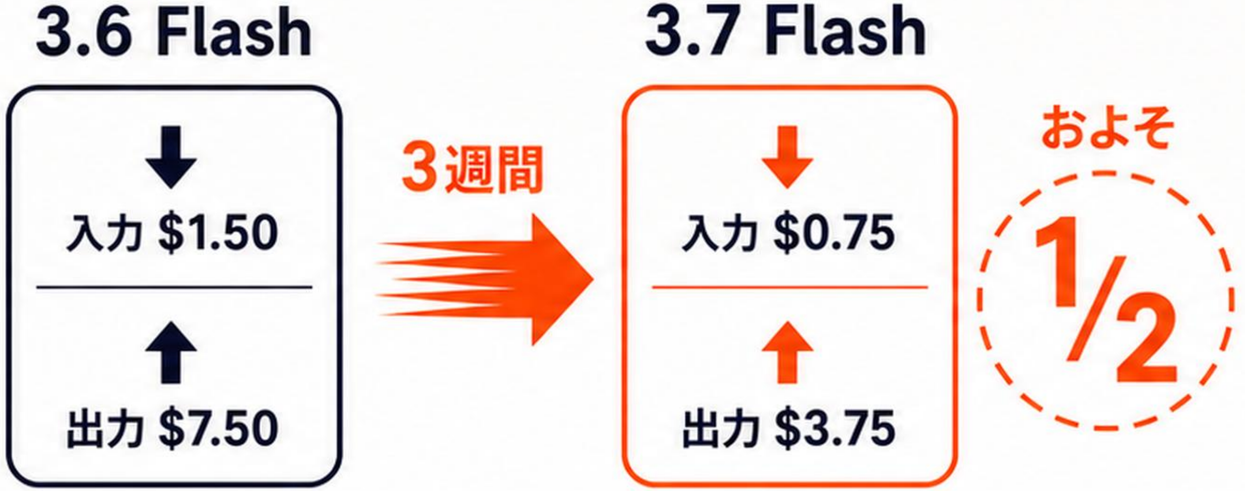
github.com

Gemini 3.7 Flash 発表

前モデルから3週間、価格は約半分、同日に Spark が乗り換え

Google が Gemini 3.7 Flash を公開した。

- 導入価格は100万トークンあたり入力 $0.75 / 出力 $3.75。3.6 Flash のおよそ半額
- Web 開発では、より少ないプロンプトで機能の揃ったレイアウトとアプリを出せるとする
- 金融・法務・バイオのような知識密度の高い領域での推論精度も改善



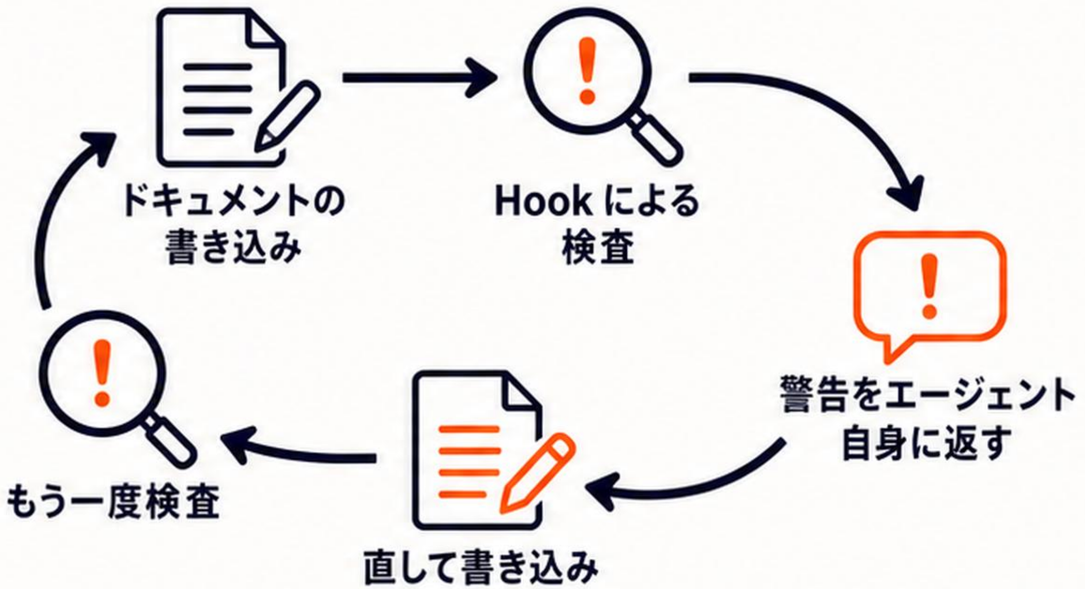
3週間でのモデル更新と半額化が同時に来た。

〈知見〉AIの変な日本語は、プロンプトではなくHookで潰す

ルール500本＋書き換え例セット

Claude Code / Codex に業務ドキュメントを書かせると内容は正確でも日本語が不自然になる。

- プロンプトで効かない理由が3つ挙がる。会話が長くなると指示が薄まる、モデルが変われば癖の出方も変わる…
- Hook は処理を止めず、警告をエージェント自身に返して直させる。直した書き込みへもう一度検査がかかる
- セッション終了時にも全体を再検査し、重大な違反が残っているとエージェントは完了報告できない

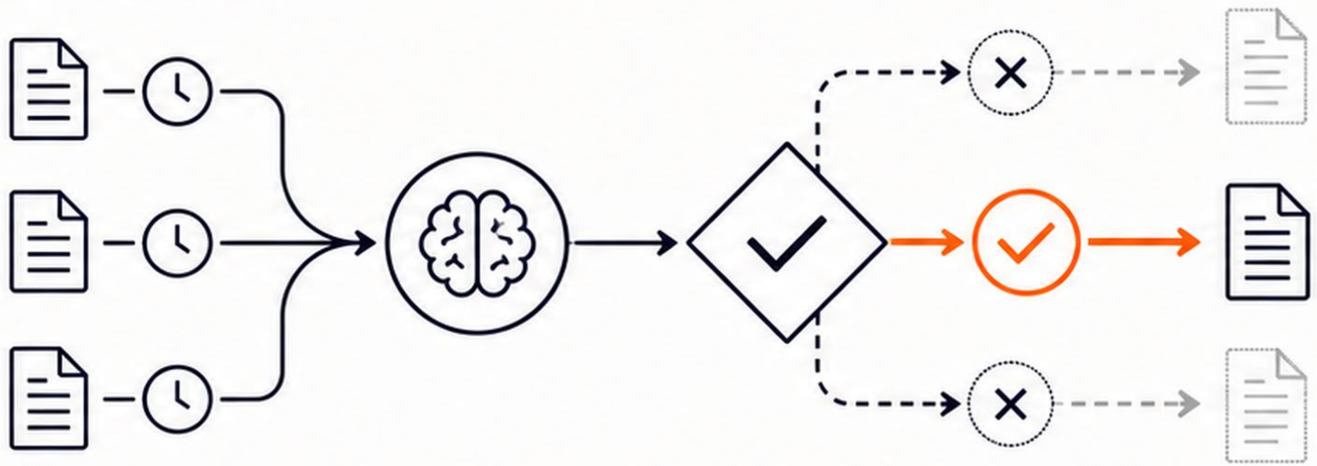


「プロンプトに書いたのに守らない」は誰もが踏む壁で、その正体が指示の希薄化だと具体的に言語化されている。

〈知見〉Agent Memory の本当の難所は「取り出した情報が、今も成立しているか」

Agent Memory の議論は「過去の情報をどう取り出すか」に留まりがちだが、エージェントがプロジェクトや月をまたいで働くようになると…

- 実務例として、顧客のゴールが3回変わっている、契約書に複数バージョンがある…
- RAG は古いメールも古い結論も取り出せるが、どのバージョンが差し替え済みか…
- 「検索的な回答（過去どこで言及されたか）」と「長期コンテキストの回答（何が変わり…

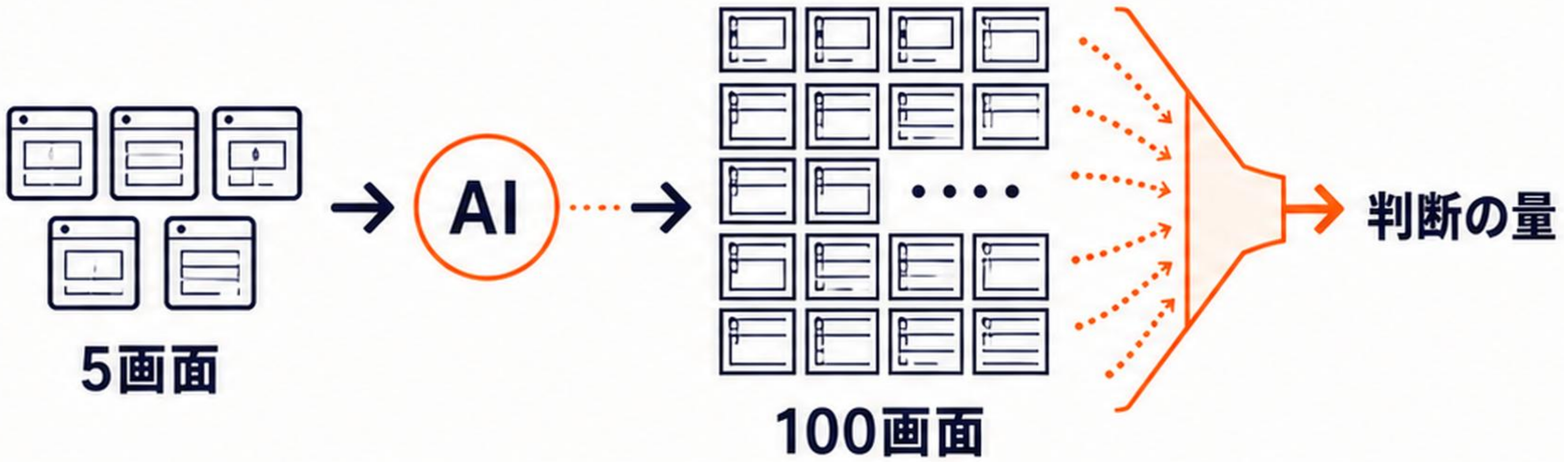


記憶を「書き込みと呼び出し」の2操作で設計している限り、履歴が長くなるほど古い結論が現行のものとして混ざる。

〈知見〉デザイナーの手が速くなると、次に足りなくなるのは「判断の量」

Figma と AI をつないで5画面から100画面へ展開したとき、
作業は減ったのに「これでいいか」を決める回数が激増した、という現場の観察。

- 5画面 → 100画面の展開で、作業量は減り、判断の回数が激増した
- 先に効いた対策は、AI が読める形に Figma を整えておくこと
- レイヤー名が乱れたファイルを渡すと、AI は乱れたまま量産する



「AI で速くなる」がそのまま生産性にならない構造を一言で説明している。

本日のトピック一覧

1. OpenAI「Ultrafast mode」
2. ChatGPT / Codex に「Computer History」
3. DeepSeek が「全部プラグイン」のエージェントハーネスを MIT 公開
4. Gemini 3.7 Flash 発表
5. 〈知見〉AI の変な日本語は、プロンプトではなく Hook で潰す
6. 〈知見〉Agent Memory の本当の難所は「取り出した情報が、今も成立しているか」
7. 〈知見〉デザイナーの手が速くなると、次に足りなくなるのは「判断の量」

2026-08-15