Tech News Daily

MORNING DISPATCH

2026-09-02

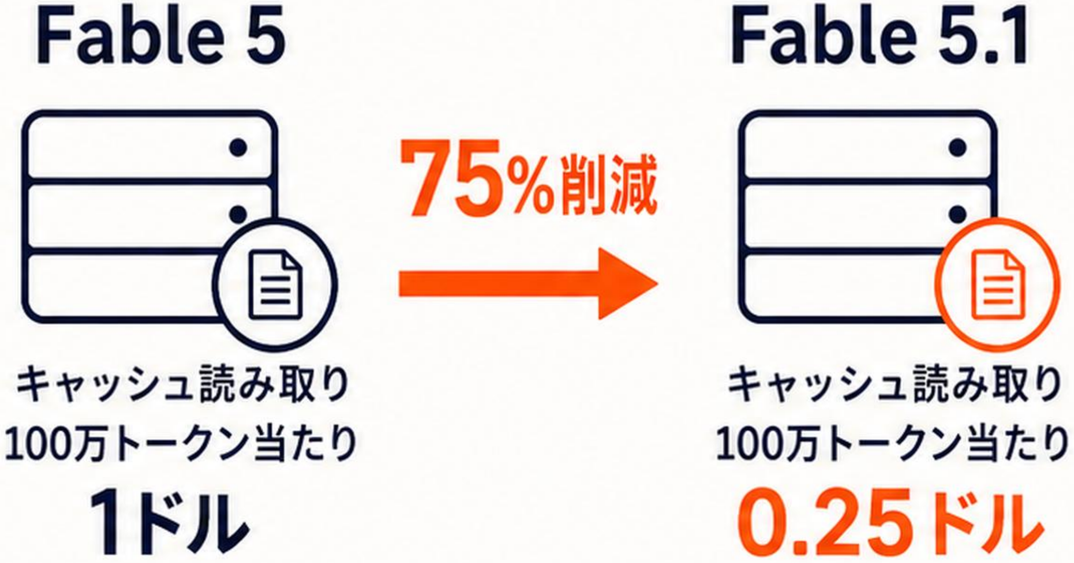
1. Claude Fable 5.1、長時間エージェントのキャッシュ料金を75%削減
2. Gemini、動画を全部読む方式から必要な区間を探す方式へ
3. Claude Code 2.1.257、auto modeの境界を具体化

8トピック

Claude Fable 5.1、長時間エージェントのキャッシュ料金を75%削減

AnthropicはClaude Fable 5.1とClaude Mythos 5.1を発表した。

- Fable 5.1は100万トークンのコンテキストと最大12万8,000トークンの出力に対応する
- 入力料金は100万トークン当たり10ドル、出力は50ドルでFable 5から据え置き
- キャッシュ読み取りは100万トークン当たり0.25ドルで、Fable 5より75%低い



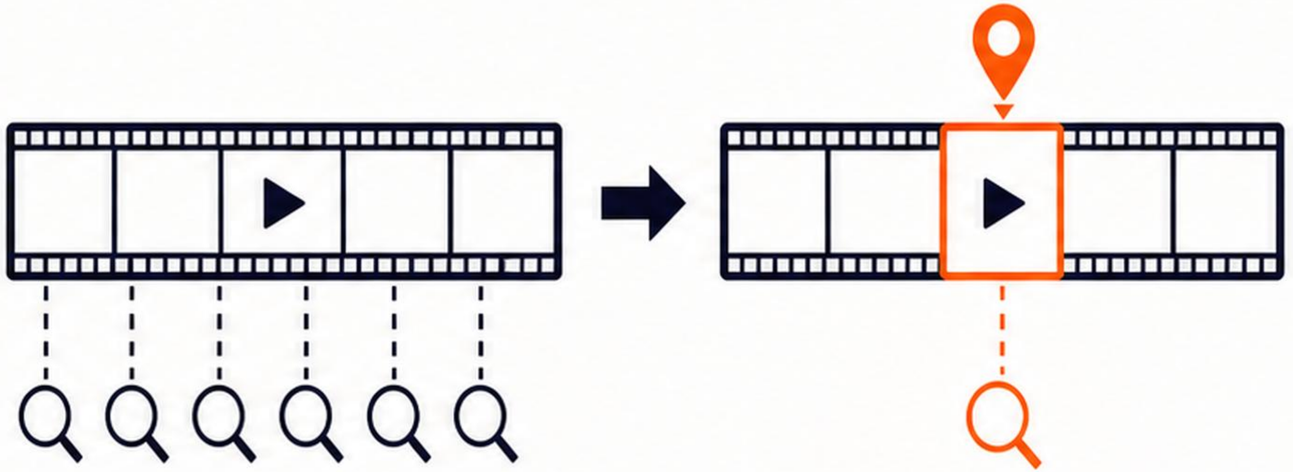
長時間ループでは、毎回読み直す仕様・履歴・ツール結果が請求の大半を占める。

anthropic.com

Gemini、動画を全部読む方式から必要な区間を探す方式へ

Google DeepMindはGemini 3.7 Flash、3.6 Flash、3.5 Flash-Liteにエージェント型動画理解を追加した。

- Googleの評価では、分析コストを最大66%、トークン消費を最大88%削減した
- 精度は最大7%向上。長尺動画ほど効率差が出やすい
- サブ秒の瞬間検索、異常検知、計数、長尺映像からの該当箇所探索を想定する

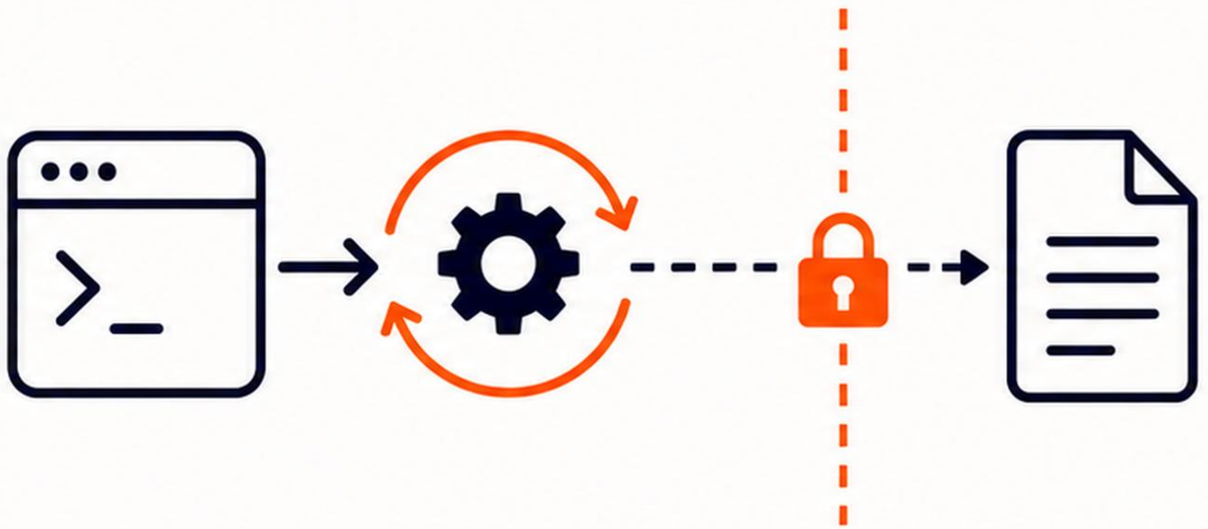


教材、会議録画、デモ動画を毎回すべて読ませる必要がなくなる。

Claude Code 2.1.257、auto modeの境界を具体化

Claude Code 2.1.257はFable 5.1をFable系の既定モデルにした。

- `claude-fable-5-1`をFable系の既定モデルにした
- `fable`と`best`のエイリアスは、未対応gatewayでは当面Fable 5のまま
- 作業ディレクトリ外のファイルを初めて読む前に確認し、設定で読み取り自体を拒否できる



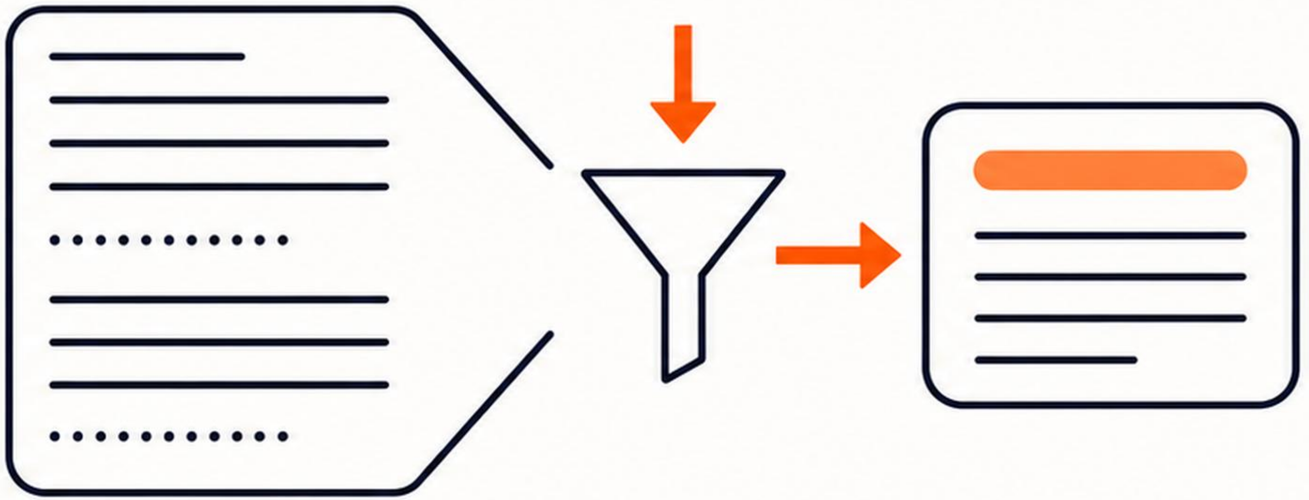
「自動で進める」と「境界を越えてよい」は別の設定だ。

github.com

Codex CLI 0.152.0、MCP出力をツール単位で制限

Codex CLI 0.152.0は、下書き内のVim検索や利用枠バナーからの操作を追加した。

- Vimモードで`/`と`?`による検索、`n`と`N`による候補移動に対応した
- レート制限バナーから利用状況、クレジット、上限リセット、プラン管理へ進める
- MCPサーバー名に`:@/.`を使える



巨大なMCP応答をそのまま会話へ入れると、必要な文脈が押し出される。

github.com

Vercel AI SDK、軽量コーディングエージェントfxを共通ハーネスへ

VercelはAI SDKのハーネス層へ、同社の軽量オープンソースエージェントfxを追加した。

- 公式アダプタは@ai-sdk/harness-fx
- @ai-sdk/harness-acpを使い、Agent Client Protocol経由でfxへ接続する
- Claude Code、Cline、Codex、Cursor、Deep Agents、Grok Build、OpenCode、Piと同じハーネス一覧に加わった

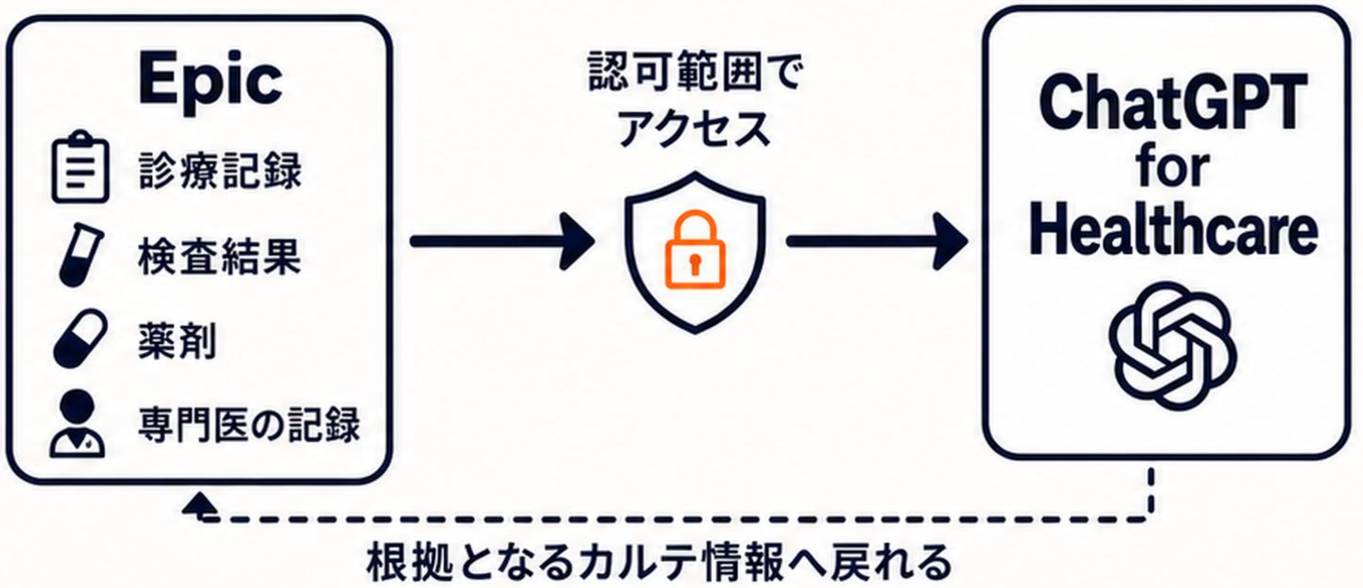
製品が特定のCLIへ直接依存すると、モデルや実行基盤を替えるたびに統合を書き直すことになる。

vercel.com

ChatGPT for Healthcare、Epicの患者情報を認可範囲で参照

OpenAIはChatGPT for HealthcareにEpic電子カルテ連携とHealthcare Public Dataプラグインを追加した。

- 診療記録、検査結果、薬剤、専門医の記録をまとめ、根拠となるカルテ情報へ戻れる
- 対応環境ではChatGPTを電子カルテ画面内へ組み込める
- 公的情報プラグインはClinicalTrials.gov、CMS Coverage、RxNorm、DailyMed、PubMedなど9ソースに接続する

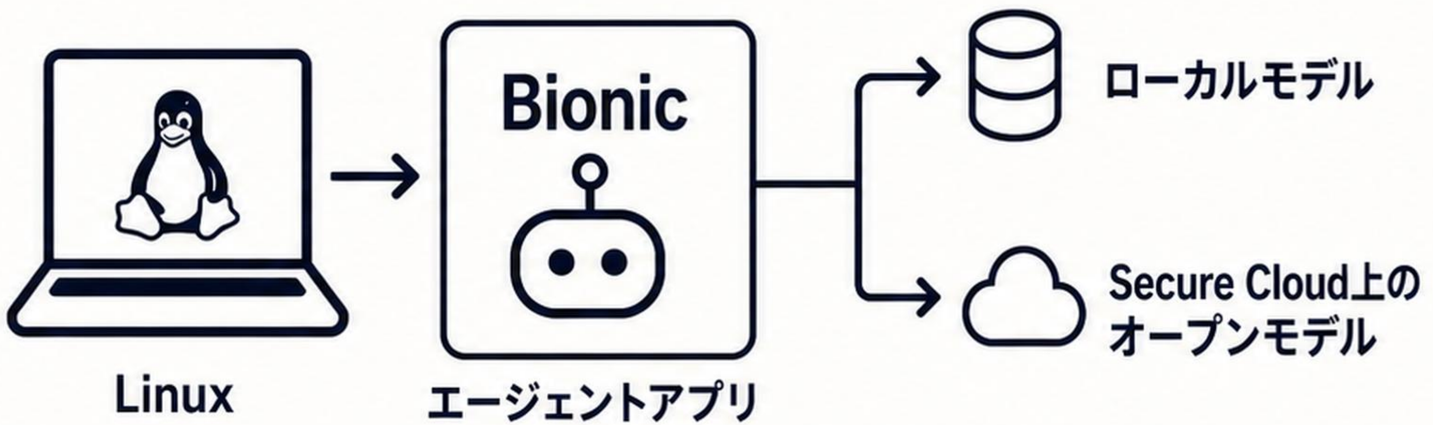


業務AIの差は、一般知識よりも、既存システムの権限を保ったまま固有データへ接続できるかに出る。

LM Studio Bionic、Linuxでローカルエージェントを実行

LM Studioはオープンモデル向けエージェントアプリBionicのLinux版を公開した。

- BionicはLM Studio本体とは別のエージェントアプリ
- Linux版はx64 AppImageとして配布される
- ローカルモデルとSecure Cloud上のオープンモデルを選べる

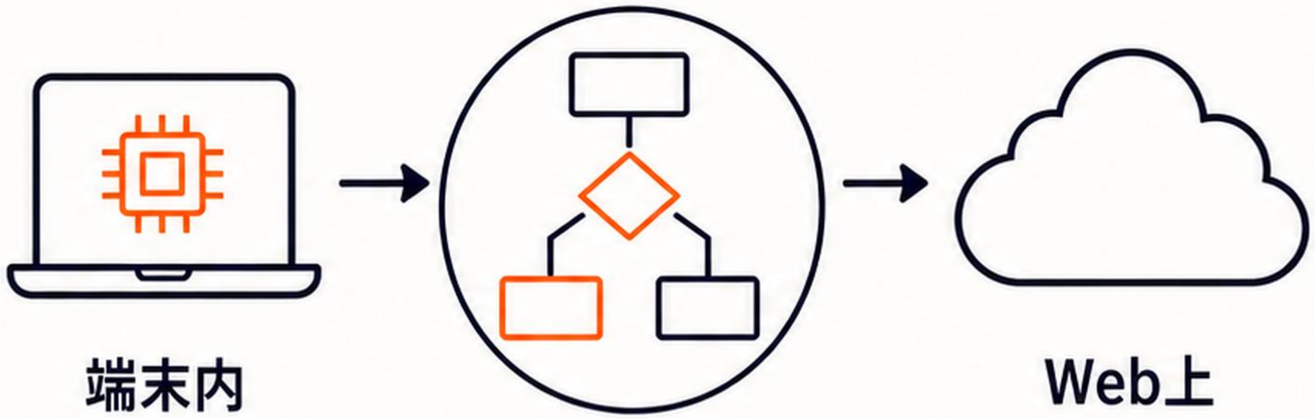


ローカルLLMの用途がチャットから、ファイルへ触るエージェントへ広がっている。

Perplexity Mac、機密工程をローカルモデルへ振り分け

Perplexity Macアプリは、クラウドとMac上のローカルモデルを組み合わせるハイブリッド計算に対応した。

- Perplexity Macアプリの全ユーザーへ提供する
- ローカル処理はApple Siliconを使い、その工程ではクラウド側のトークン料金が発生しない
- 端末内とWeb上の実行環境をタスクごとに使い分ける



ローカルかクラウドかをアプリ単位で決めるのではなく、工程単位で分ける設計だ。

本日のトピック一覧

1. Claude Fable 5.1、長時間エージェントのキャッシュ料金を75%削減
2. Gemini、動画を全部読む方式から必要な区間を探す方式へ
3. Claude Code 2.1.257、auto modeの境界を具体化
4. Codex CLI 0.152.0、MCP出力をツール単位で制限
5. Vercel AI SDK、軽量コーディングエージェントfxを共通ハーネスへ
6. ChatGPT for Healthcare、Epicの患者情報を認可範囲で参照
7. LM Studio Bionic、Linuxでローカルエージェントを実行
8. Perplexity Mac、機密工程をローカルモデルへ振り分け

2026-09-02