Tech News Daily

MORNING DISPATCH

2026-08-25

1. NVIDIA、推論専用チップ「Groq 3 LPX」が量産開始
2. Anthropic、MCPコネクタの「会社まとめて認可」が正式提供に
3. OpenAI と AWS、Kiro 上の GPT-5.6 を最適化

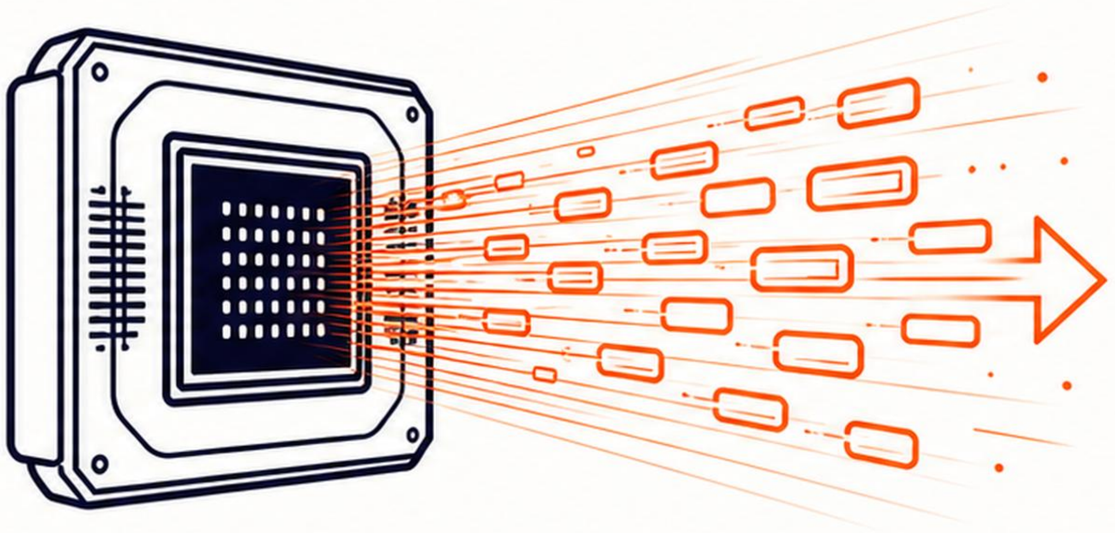
8トピック

NVIDIA、推論専用チップ「Groq 3 LPX」が量産開始

Gemma 4 31B で毎秒3,400トークン

NVIDIA が米国時間 8/24 の Hot Chips で、推論専用アクセラレータ「NVIDIA Groq 3 LPX」の量産開始を発表した。

- オープンモデル Gemma 4 31B・コンテキスト10万トークンで毎秒3,400出力トークン（計測は Artificial Analysis）
- NVIDIA の主張で応答性は競合プラットフォームの約4倍。エージェントのコーディングタスクが「数時間ではなく数…
- Vera Rubin NVL72 を推論ワークロード向けに拡張する位置づけ。7種のチップと5種のラックで構成（BlueField-4…



エージェントの体感速度は賢さよりトークンの吐き出し速度で決まる場面が多い。

nvidianews.nvidia.com

本日のトピック一覧

01 Anthropic、MCPコネクタの「会社まとめて認可」が正式提供に

Claude Team / Enterpriseで、IT管理者によるMCPコネクタの一括認可が正式提供

Okta対応。Datadog・Notion・Slackなど対象拡大

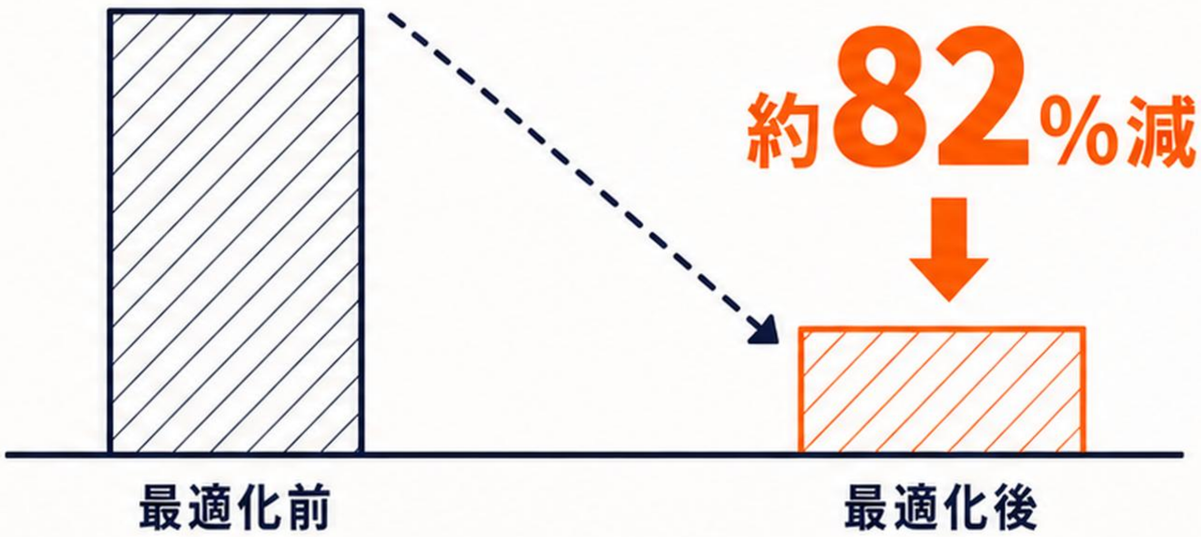
2026-08-25

OpenAI と AWS、Kiro 上の GPT-5.6 を最適化

タスク完了コストが約82%減

OpenAI が 8/24、AWS と共同で Kiro 上の GPT-5.6 を価格性能面で最適化したと発表した。

- Terminal-Bench 2.1 で、GPT-5.6 Terra がタスクを完了するまでのコストが約82%減（OpenAI 発表）
- Kiro のクレジット倍率は Sol 2.4x / Terra 1.2x / Luna 0.6x
- Coding Agent Index は Sol 80 / Terra 77.4 / Luna 74.6

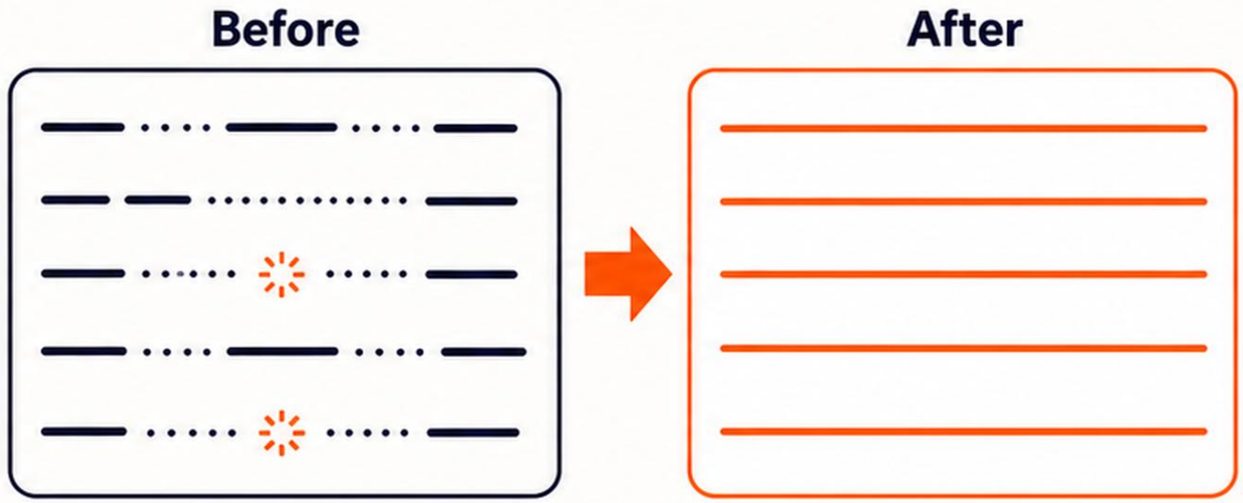


速いモデルや賢いモデルを選ぶより「文脈を先に固めてから渡す」ほうがコストに効く、という主張をモデル提供元がベンチ数値で出してきた。

Claudeの Web・デスクトップ、長文の描画を作り直して約4倍なめらかに

Anthropic が 8/24、Claude の Web 版とデスクトップ版の
ストリーミング描画を作り直したと告知した。

- 長い回答のストリーミング表示が
  約4倍なめらかに
- 非力なノートPCでは停止（stall）が
  9分の1、最悪のフリーズ時間は4.5分の1
- 120Hz の MacBook では最初から
  最後まで 120fps を維持



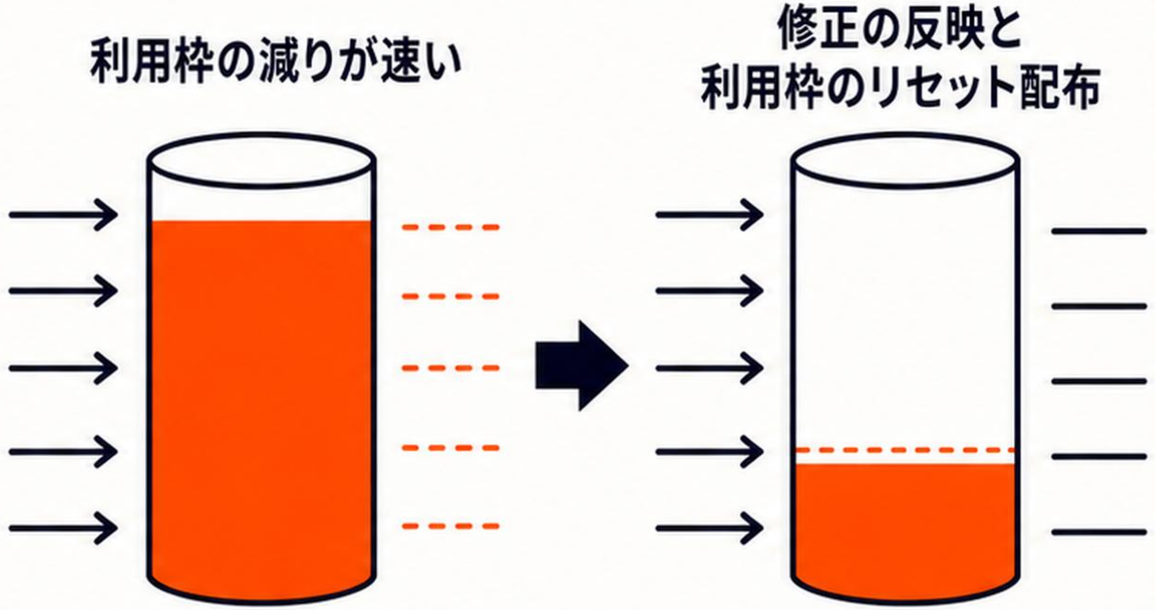
出力が速くなったのではなく「読める形で出てくる」ようになった改善。

x.com

Codex の利用枠が急に減る問題、原因3つを特定して修正

「Codex の利用枠の減りが速い」という報告が続いていた件について、OpenAI の Thibault Sottiaux が 8/23 に原因3点を公表し、8/24 に修正の反映と利用枠のリセット配布を告知した。

- 原因① 長いセッションで画像を扱い、compaction（圧縮）を複数回はさんだときの非効率
- 原因② Computer History 機能の p95 以上の使用量が高かった
- 原因③ 会話タイトルを自動生成する機能が想定より枠を消費していた

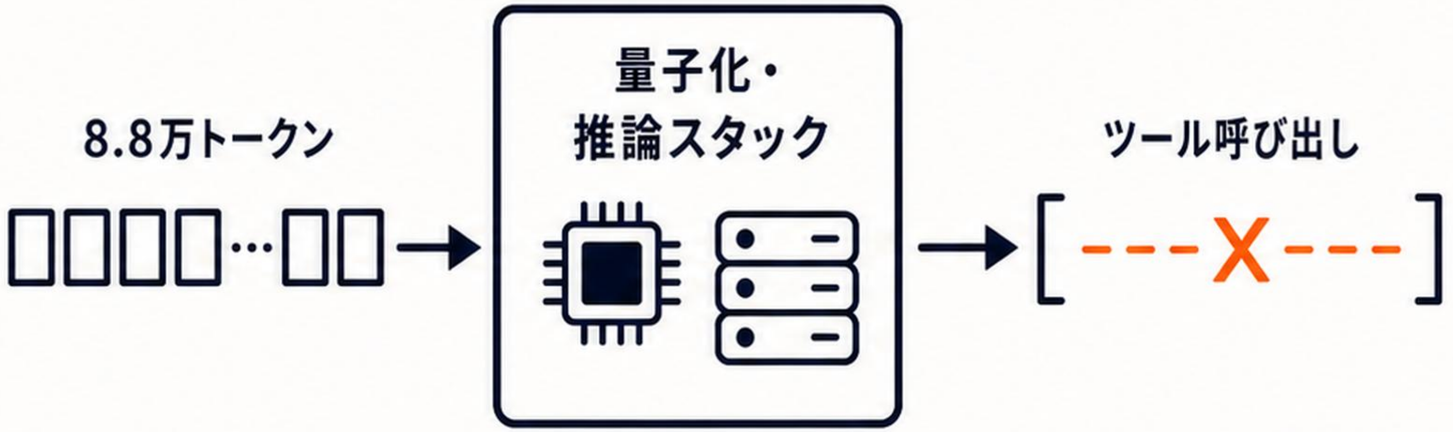


上限に当たるとまず自分の使い方を疑うが、今回は実装側の消費が原因だった。

〈知見〉ローカルLLMが「バカに感じる」のはモデルのせいではない

同じモデル・同じ重みでも、量子化の形式と推論スタックの設定で出力が別物になるという検証記事。

- NVFP4（混合）が最悪。8.8万トークン地点で約50%のトークンが BF16 と食い違い、ツール呼び出しが完了しなかった
- AWQ W4A16 も約40%のトークンが食い違い、ツール呼び出しを正しく閉じられなかった
- 公式 FP8 での 8.8万トークン地点の不一致は約15〜20%



ローカルLLM は動いた時点で使えると判断しがちだが、短いやり取りでは差が出ず、長い文脈のツール実行で初めて壊れる。

forum.level1techs.com

〈知見〉2〜3時間で作ったスキルが社内利用1位に

効いたのは4,400行の社内知識

タイミーの AI/データ部門が公開した、Claude Code スキル「bq-analysis」が社内で最も使われるまでの実データ付きレポート。

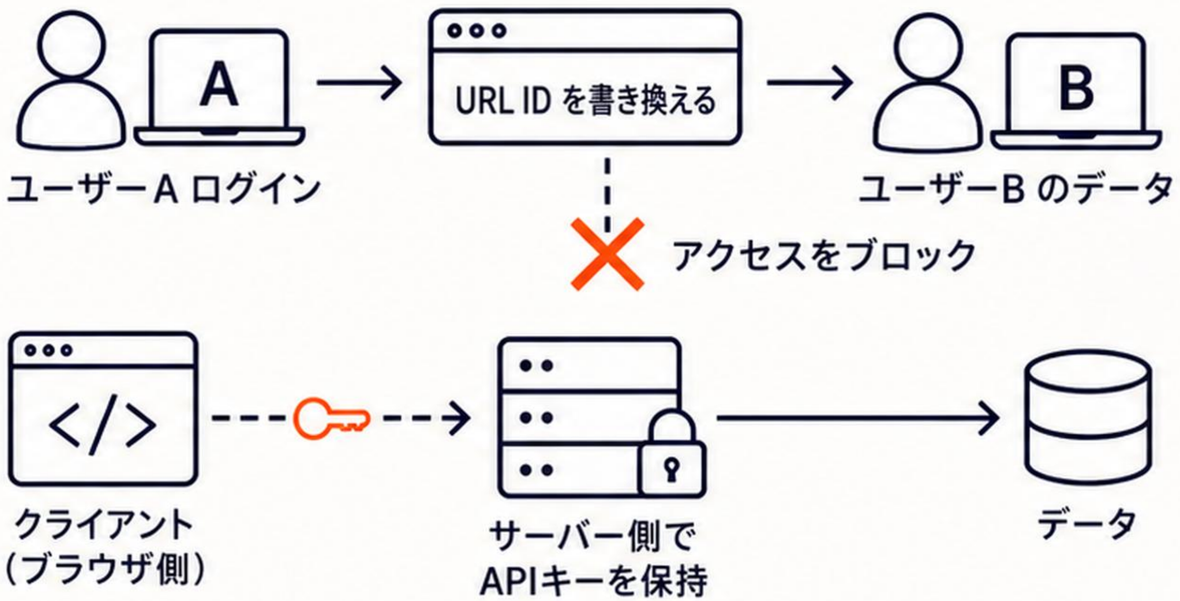
- 初回コミットは 2/9、3ファイル・582行。実働は2〜3時間
- 6月までに 582行 → 1,823行（3.1倍）。最終構成は約4,400行の社内参照資料
- 内訳は用語辞書20% / 指標定義10% / 計測仕様60%超。コードではなく社内の前提が本体

スキルを作る価値は手順を書くことではなく、社内にしかない前提を言語化することにある、という実例。

Claude Code で作ったツールを公開する前に確認すべき8項目

Claude Code で作ったツールを一般公開する前に潰すべき穴を8つに整理した X Article。

- 最頻の事故は認可の抜け。ログインは通るが URL の IDを書き換えると他人のデータが見える…
- 検証は人力で足りる。テストアカウントを2つ作り、A でログインしたまま B のデータ URL を開いて中身が見えたらアウト
- APIキーはブラウザ側に置かない。クライアントコードに置くと開発者ツールで誰でも読める…



バイブコーディングで作ったものを外に出す段階でそのまま使えるチェックリスト。

本日のトピック一覧

01 NVIDIA、推論専用チップ「Groq 3 LPX」が量産開始

02 Anthropic、MCPコネクタの「会社まとめて認可」が正式提供に

03 OpenAI と AWS、Kiro 上の GPT-5.6 を最適化

04 Claude の Web・デスクトップ、長文の描画を作り直して約4倍なめらかに

05 Codex の利用枠が急に減る問題、原因3つを特定して修正

06 〈知見〉ローカルLLMが「バカに感じる」のはモデルのせいではない

07 〈知見〉2〜3時間で作ったスキルが社内利用1位に

08 Claude Code で作ったツールを公開する前に確認すべき8項目

2026-08-25