Tech News Daily MORNING DISPATCH

2026-07-29

1. MCP 仕様 2026-07-28 が正式版に
2. OpenAI が書き起こしモデルを2種投入
3. llama.cpp 本体が MCP stdio に対応

8トピック

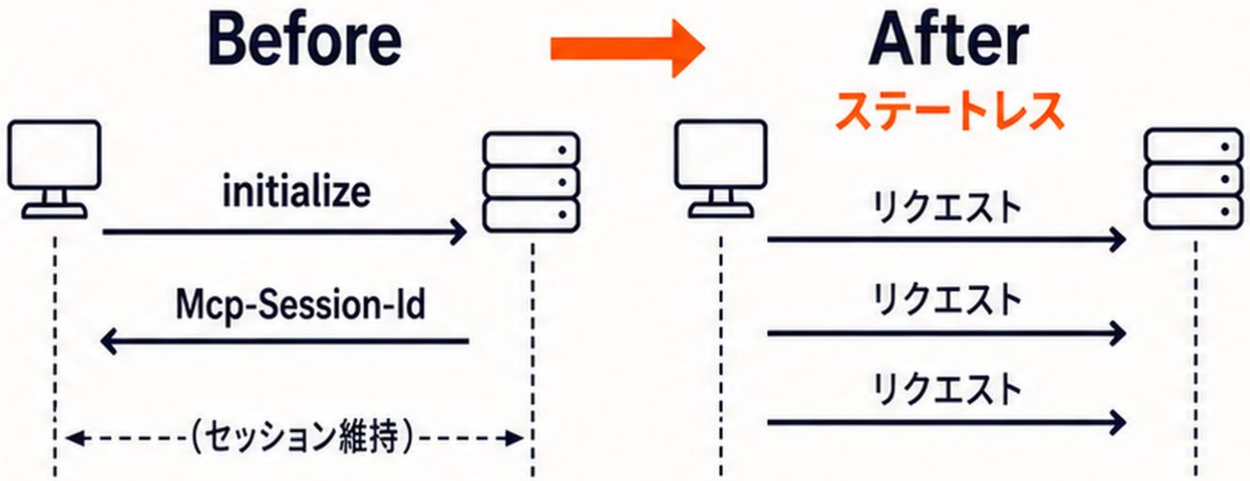
TEKION Group

MCP 仕様 2026-07-28 が正式版に

プロトコルがステートレスになり、initialize ハンドシェイクとセッションが消えた

Model Context Protocol の新仕様 2026-07-28 が正式版として公開された。

- 出典: Model Context Protocol 公式ブログ（2026-07-28）
- 分類: 📰 プロトコル / 開発者基盤
- `initialize`/`initialized` ハンドシェイクと `Mcp-Session-Id` ヘッダを撤廃。「どのリクエストも…



MCP サーバを自前でホストしている側には破壊的変更で、セッション初期化の除去・リクエストの自己完結化・DCR から CIMD への認可移行が実作業として降ってくる。

OpenAI が書き起こしモデルを2種投入

Whisper の誤り率を半分以下に、「文脈を渡せる」ASR へ

OpenAI が API に GPT-Live-Transcribe（低遅延のストリーミング書き起こし）と GPT-Transcribe（完了済み音声ファイル・バッチ向け）を追加した。

- 出典: OpenAI API Changelog / @OpenAIDevs（2026-07-28）
- 分類: 📰 モデル / 音声AI
- `gpt-live-transcribe`（マイク・通話など生の音声ストリーム向け、低遅延で transcript の差分を返す）と…

議事録・商談録音・セミナー書き起こしを回している運用で効くのは、モデルを差し替えることより事前に渡す文脈を用意すること。

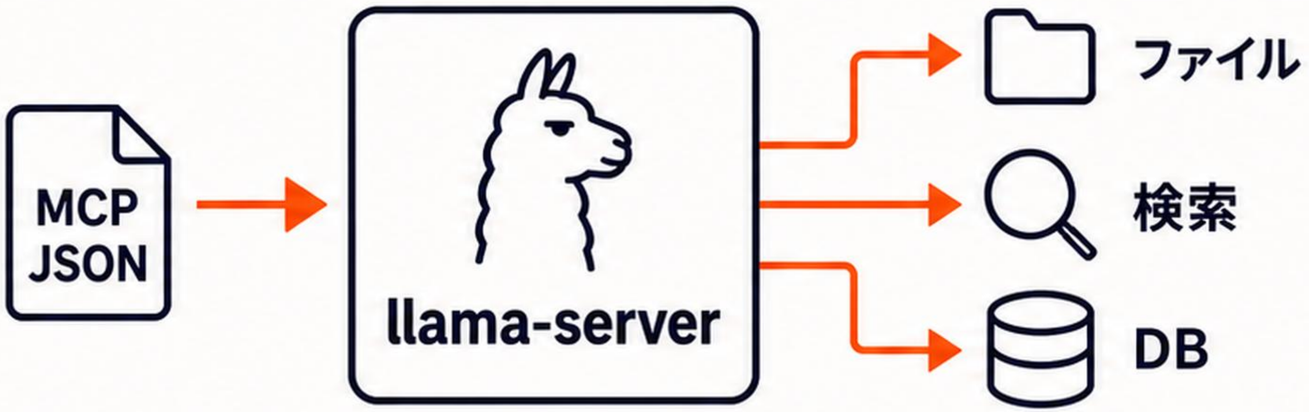
developers.openai.com

llama.cpp 本体が MCP stdio に対応

llama-server が単体でエージェントハーネスになった

ローカル LLM の実行エンジン llama.cpp に「server: support MCP stdio」（PR #26062）がマージされ、llama-server 自身が MCP の JSON 定義を受け取って MCP…

- 出典：ggml-org/llama.cpp PR #26062（2026-07-26）
- 分類：ローカルLLM / エージェント
- PR #26062「server: support MCP stdio」が 2026-07-25 23:08 UTC（JST 07-26 08:08）にマージ。先行 PR…



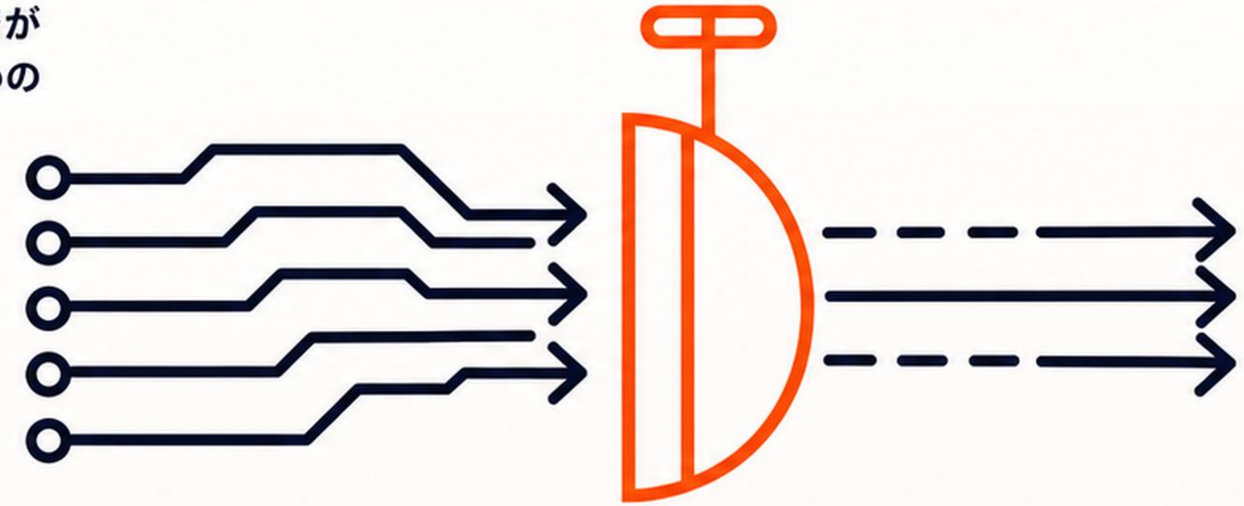
ローカル LLM を「チャットが動くだけ」から「ファイル・検索・DB を触るエージェント」に引き上げる最後のピースが、外部フレームワークではなくエンジン本体に入った。

Anthropic が「フロンティア開発を意図的に減速させる」請願を公式支持

CEO と複数の共同創業者が署名

OpenAI と Anthropic の従業員が回している請願——米政府が「自動化された AI 開発のフロンティアを意図的にペース調整するための技術的・ガバナンス的ツール…」

- 出典：@AnthropicAI / Bloomberg（2026-07-28）
- 分類：AI ガバナンス / 業界
- Anthropic 公式アカウントの表明：「We support this petition, signed by our CEO, several co-founders, and senior staff.」



従業員の署名運動を CEO と共同創業者が追認するのは、社内の温度と経営方針が同じ方向を向いていることの公開表明にあたる。

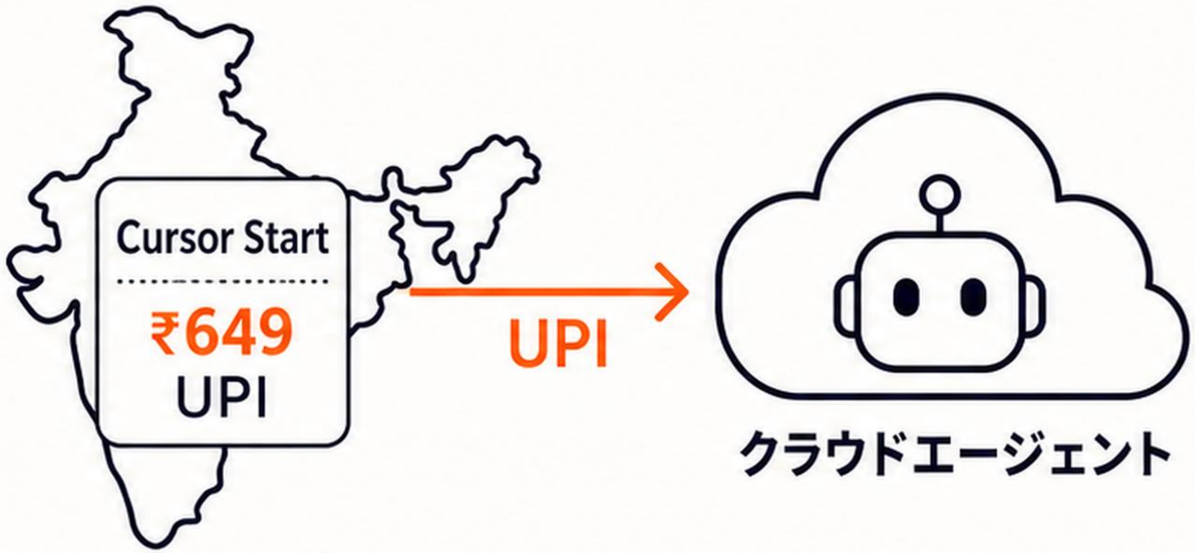
x.com

Cursor が地域別プラン「Cursor Start」をインドで開始

月₹649・UPI 決済、常時稼働クラウドエージェント込み

Cursor がインド向けの地域別プラン「Cursor Start」を月額 ₹649（税込）で開始した。

- 出典: Cursor 公式プログ / @cursor_ai（2026-07-28）
- 分類: 💼 AI×ビジネス / 価格戦略
- 月額 ₹649（税込）、INR 建て・月次自動更新、UPI またはカードで支払い。2026-07-28 提供開始



注目すべきは価格そのものより、常時稼働のクラウドエージェントが最安の有料段に降りてきたこと。

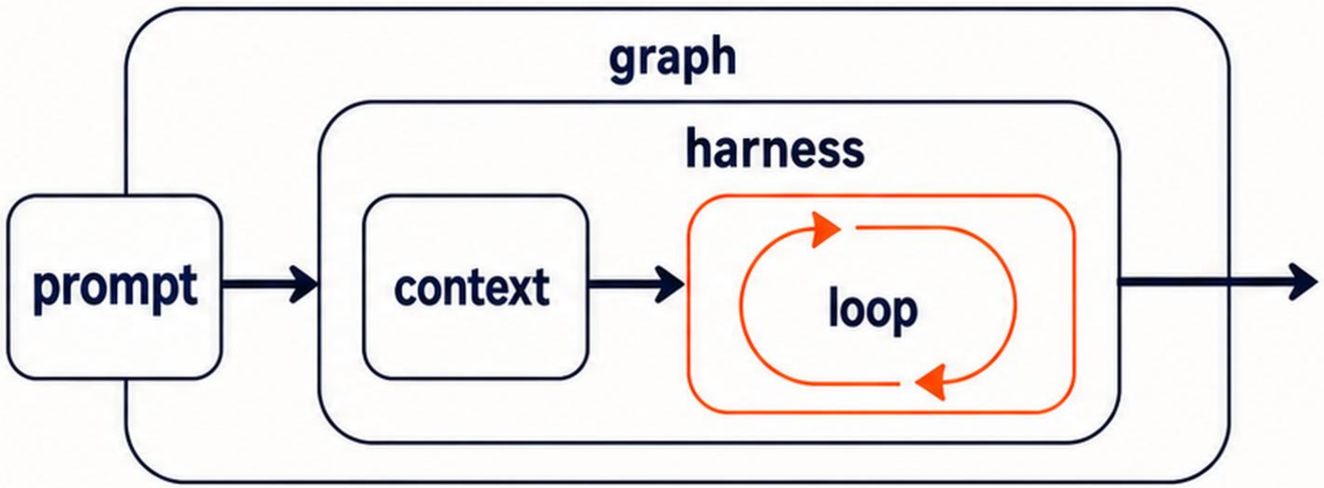
cursor.com

〈知見〉prompt → context → harness → loop → graph

5層を「作業の単位」で切り分ける

増え続ける「○○エンジニアリング」という用語は互いを置き換える競合概念ではなく入れ子構造であり、
見分ける唯一クリーンな基準は「1単位の作業（unit of work）が何か」だと整理した投稿。

- 出典: @akshay_pachaar（2026-07-26）
- 分類: 🧠 コンテキスト / ループエンジニアリング
- prompt はメッセージ。モデルはこの呼び出し以前を何も覚えていないので…



自分のループ運用の障害切り分けチェックリストとしてそのまま使える。

〈知見〉ローカル LLM の「遅い」を3区間に分解する

Ollama の API が返す3つの duration を残す

ローカル LLM が「遅い」と感じるとき、①モデルをストレージから読み込む時間 ② プロンプトを読み取る時間 ③ 回答を生成する時間、の3区間に分けると原因を追いやすくなる、という計測の作法。

- 出典: @murasametech（2026-07-27）
- 分類: ローカルLLM / 計測
- 3区間に分ける: ① モデルのストレージからの読み込み ② プロンプトの読み取り（prefill） ③ 回答の生成（decode）

「遅いから量子化を強くする」「遅いからモデルを小さくする」という反射的な対処は、ボトルネックがモデルのロード時間だった場合ほとんど効かない。

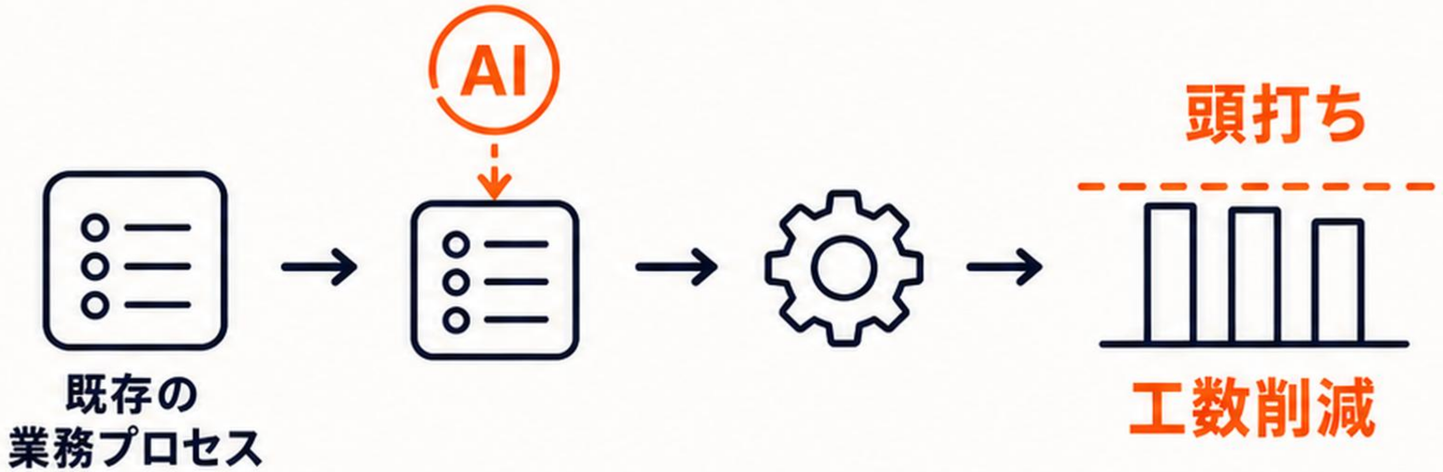
x.com

〈知見〉AIを「既存業務に足す」のと「仕事の設計をやり直す」のは別物

前者は工数削減で頭打ちになる

AIを既存の業務プロセスに追加するやり方は工数削減で終わる。

- 出典：@dg_kikushu（2026-07-28）
- 分類：AI×ビジネス活用
- 既存業務にAIを追加 → 得られるのは工数削減。上限は「今やっている仕事の時間」で頭打ちになる



社内で「AIを導入したのに効果が薄い」となる典型が前者で止まっているケース。

本日のトピック一覧

1. MCP 仕様 2026-07-28 が正式版に
2. OpenAI が書き起こしモデルを2種投入
3. llama.cpp 本体が MCP stdio に対応
4. Anthropic が「フロンティア開発を意図的に減速させる」請願を公式支持
5. Cursor が地域別プラン「Cursor Start」をインドで開始
6. 〈知見〉prompt → context → harness → loop → graph
7. 〈知見〉ローカル LLM の「遅い」を3区間に分解する
8. 〈知見〉AI を「既存業務に足す」のと「仕事の設計をやり直す」のは別物

2026-07-29

TEKION Group