

2026-08-16

1. Anthropic が2回目の Risk Report を公開
2. Claude Code 2.1.233
3. Qwen 3.8 27B が Ollama に

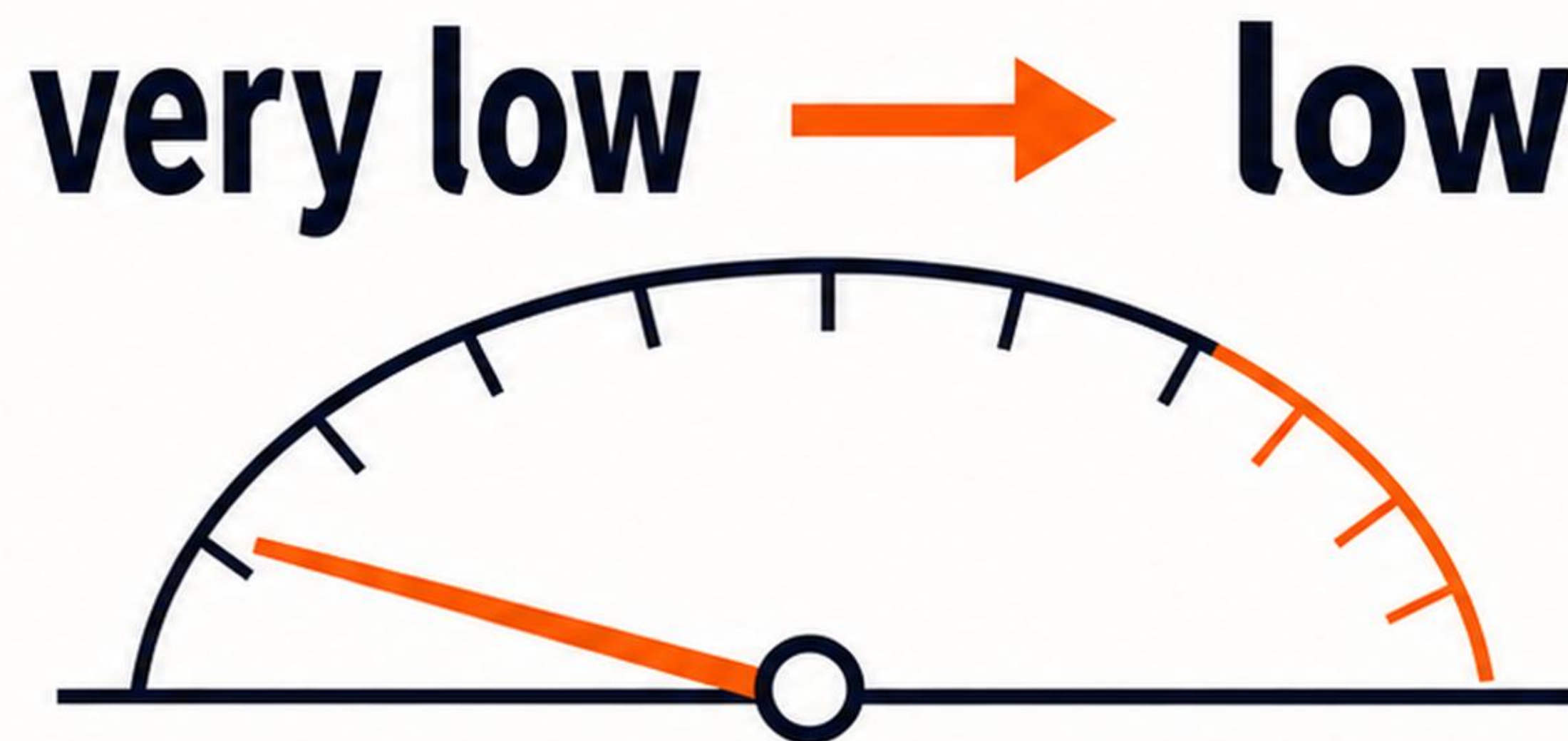


Anthropic が2回目の Risk Report を公開

自律性リスクを「very low」から「low」へ引き上げ

Anthropic が Responsible Scaling Policy に基づく Risk Report の2本目を公開した。

- 自律性の脅威モデル1（高リスク環境での misalignment）の総合評価を「very low」から「low」へ引き上げ…
- AI R&D 自動化の脅威モデルは「low」だが、最も具体的なタスクベース評価が『飽和…』
- カバー日時点で未公開の内部モデルが3つ — Claude Opus 5（当時内部展開・その後公開）、Model 1、Model…



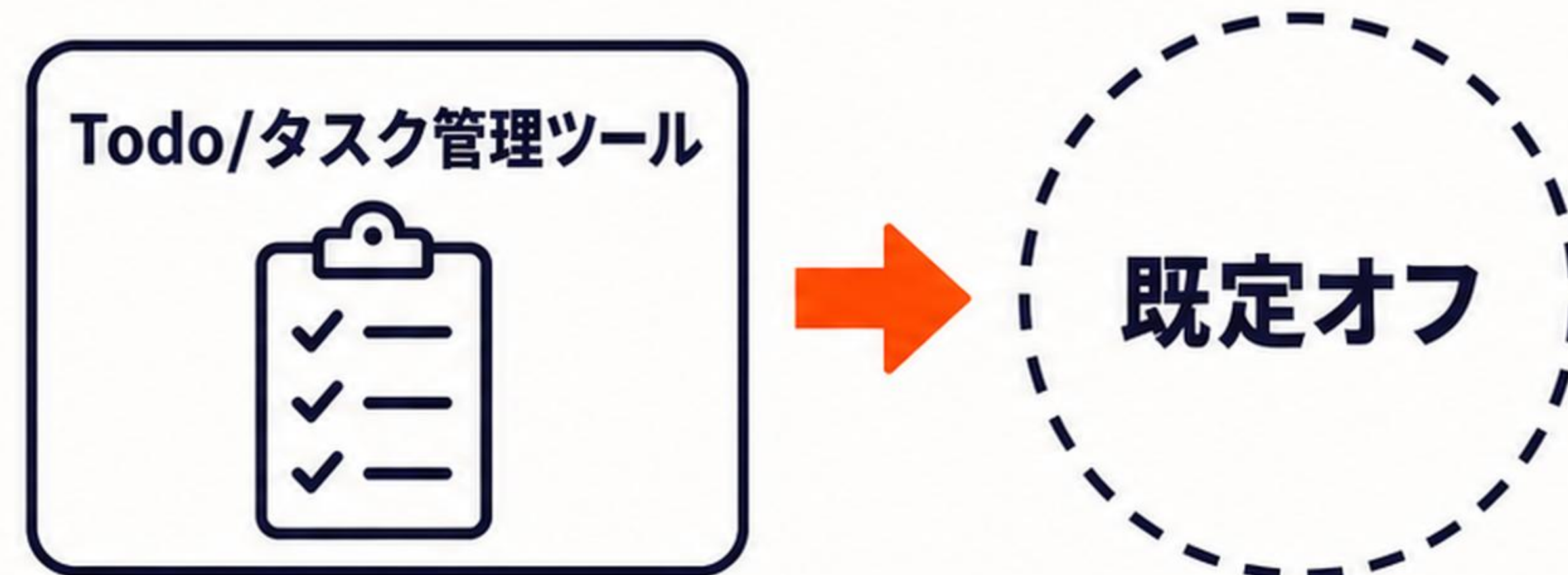
「安全である」という宣言ではなく、評価手法が能力に追いつかなくなった事実（評価の飽和）と…

Claude Code 2.1.233

Todo/タスク管理ツールが新しいモデルで既定オフに

Claude Code v2.1.233 で、TaskCreate / TaskGet / TaskUpdate / TaskList / TodoWrite
といったタスク管理ツール群が Opus 4.8 ・ Sonnet 5 ・ Fable 5 ・ Mythos 5...

- 廃止対象は TaskCreate / TaskGet / TaskUpdate / TaskList / TodoWrite。対象モデルは Opus 4.8 ・ Sonnet...
- 直前の 2.1.232 でサブエージェントの fork が既定オンに。
subagent_type: "fork"...
- 同じ 2.1.232 で、プロンプトに @ を打って別の Claude セッションを名前呼び、SendMessage...



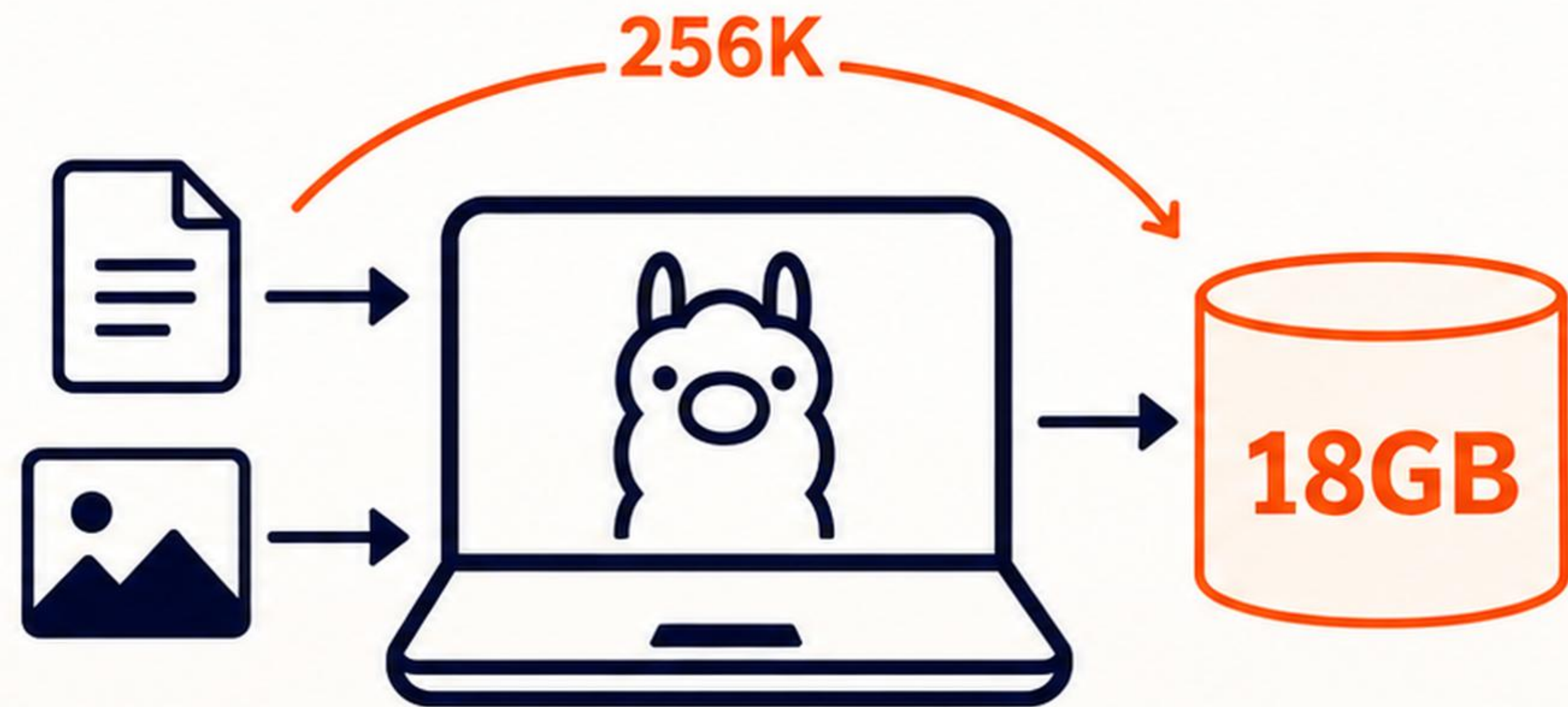
足場（scaffold）を外す方向の変更。

Qwen 3.8 27B が Ollama に

256K コンテキスト・画像も読む・18GB でローカル完結

Alibaba の Qwen 3.8 27B が Ollama で配布開始した。

- 270億パラメータ・配布サイズ 18GB。バリエーションは qwen3.8:27b と Apple Silicon 向け qwen3.8:27b-mlx
- コンテキスト長は 256K トークン
- 入力はテキストと画像の両方。thinking（推論）は既定オンで、リクエスト単位に切れる



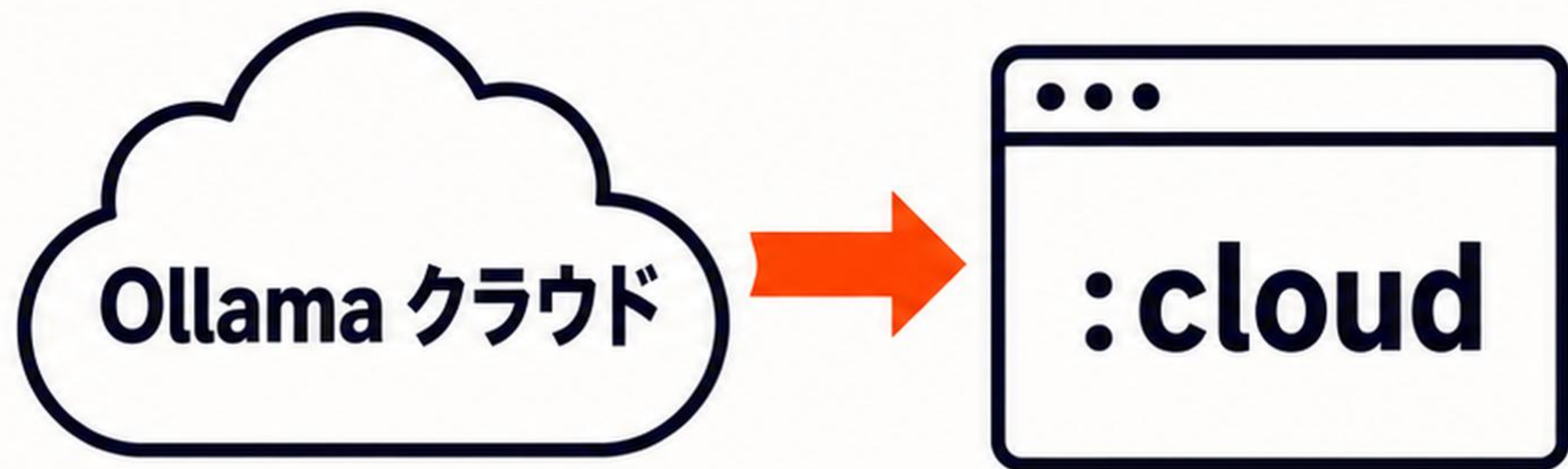
18GB は 32GB メモリのマシンで現実的に回るサイズで、256K コンテキストと画像入力がローカルに降りてきた意味は大きい。

DeepSeek-V4-Pro (0813) が Ollama クラウドで正式提供

ハーネスも1日で主要ランタイムに載った

DeepSeek-V4-Pro の 0813 版が Ollama のクラウドで全面展開され、Pro / Max プランに含まれるようになった。

- モデルは **総パラメータ 1.6T・アクティブ 49B** の Mixture-of-Experts、コンテキスト長 **100万トークン**
- 推論の強度を3段階で切り替えられる — no thinking (速い直感的回答) / thinking (丁寧な論理分析) / max...
- 配布されるのは **:cloud 系バリエーション**のみ。ローカル実行ではなく Ollama のクラウド経由である点に注意

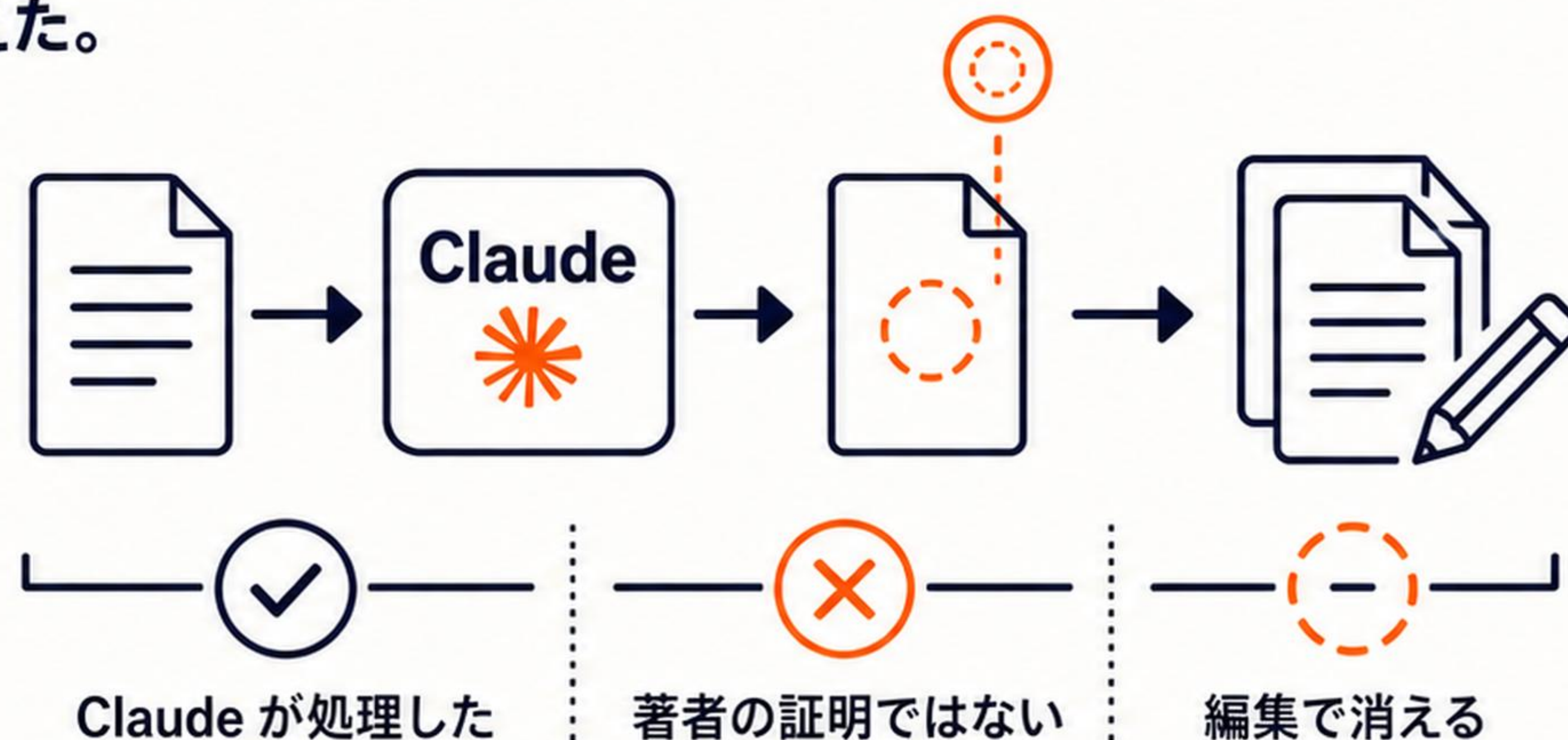


先週 MIT 公開された DeepSeek Harness が、公開から1日で主要ランタイムに載った。

Claude の電子透かし、FAQ で分かったのは「検出器はまだ無い」「著者の証明にはならない」

8月上旬に始まった Claude 生成テキストへの不可視の電子透かしについて、Anthropic が FAQ を公開して運用上の疑問に答えた。

- 検出ツールはまだ公開されていない。「ユーザーや第三者が透かしとプロヴェナンス情報を検出できるようにす…
- 検出された印が示すのは「Claude が処理した」ことであって「Claude が著者である」ことではない — 校正…
- 透かしはコピー&ペーストに追従し「一部の編集には耐える」が、大幅な編集・言い換え・翻訳…



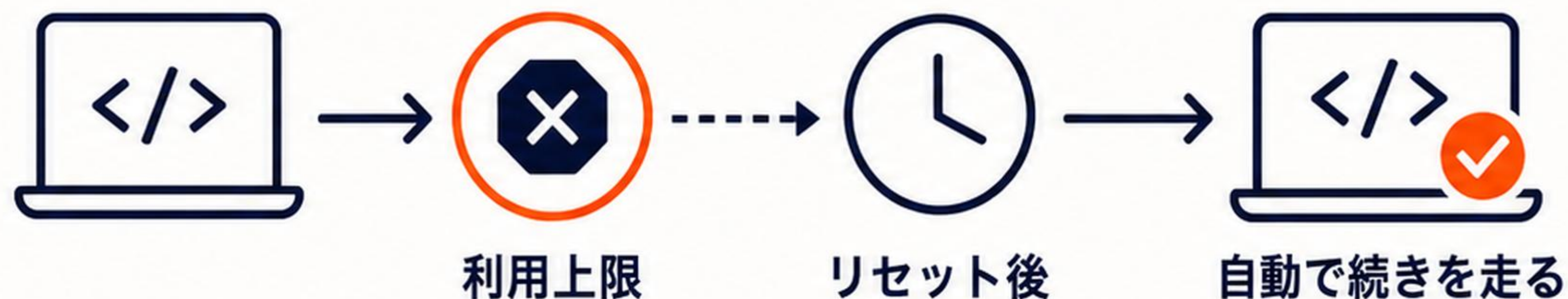
「AI が書いたかどうかを機械で判定できる」という期待に対し、公式が自分で限界を4つ（検出器未提供・処理の証明にすぎない・編集で消える・短文は不可）明示した。

Claude Code デスクトップ版に「自動再開」

利用上限は停止ではなく遅延になる

Claude Code のデスクトップアプリに Auto-continue のチェックボックスが追加された。

- 対象は **デスクトップアプリ**。CLI の CHANGELOG (v2.1.230~233) には該当項目が無いので...
- オンにすると、利用上限で停止したセッションがリセット後に自動で続きを走る
- 公式アカウントの投稿としては直近で最大級の反応 (Likes 15,000超)



地味だが、長時間走らせるループの設計に効く。

〈知見〉Seedance 2.5 は「描写」ではなく「素材の役割分担」で指揮する

Seedance 2.5 が1回の生成で画像30枚・動画10本・音声10本を参照として受け取れるようになった（ByteDance 公式、2026-07-31発表）。

- 残すものと変えるものを分けて書く — 残す（演技・タイミング・視線・立ち位置・カメラの動き…
- 固定度を3段階に分ける — 固定（人物・商品の形・動きの元素材＝作り替えさせない）／誘導（夜…
- PRIORITY 節で競合時の優先順位を先に決める — 全部を同時に守れない場面は必ず来るので、何を先に守るかを番号付きで書く



参照が1枚から30枚に増えると、うまい言い回しを探すプロンプト職人芸そのものが合理性を失う。

本日のトピック一覧

- 1 Anthropic が2回目の Risk Report を公開
- 2 Claude Code 2.1.233
- 3 Qwen 3.8 27B が Ollama に
- 4 DeepSeek-V4-Pro (0813) が Ollama クラウドで正式提供
- 5 Claude の電子透かし、FAQ で分かったのは「検出器はまだ無い」「著者の証明にはならない」
- 6 Claude Code デスクトップ版に「自動再開」
- 7 〈知見〉 Seedance 2.5 は「描写」ではなく「素材の役割分担」で指揮する

2026-08-16