

2026-08-09

1. Ollama Cloud の DeepSeek-V4-Flash が200 tok/s超へ、ZDRを維持
2. Gemini が Chrome の直近2リリースで1,000件超のセキュリティ修正を支援
3. Harrison Chase が整理した「Managed Agent」の3層構造

6トピック

Ollama Cloud の DeepSeek-V4-Flash が200 tok/s超へ、ZDRを維持

Ollama は DeepSeek-V4-Flash-0731 のクラウド出力速度が200 tok/sを超えたと公表した。

- 2026年8月8日時点の公表値は出力200 tok/s超
- DeepSeek-V4-Flash-0731 は Ollama Cloud の `deepseek-v4-flash` に全面展開済み
- 米国・欧州リージョンでゼロデータ保持 (ZDR) を掲げる

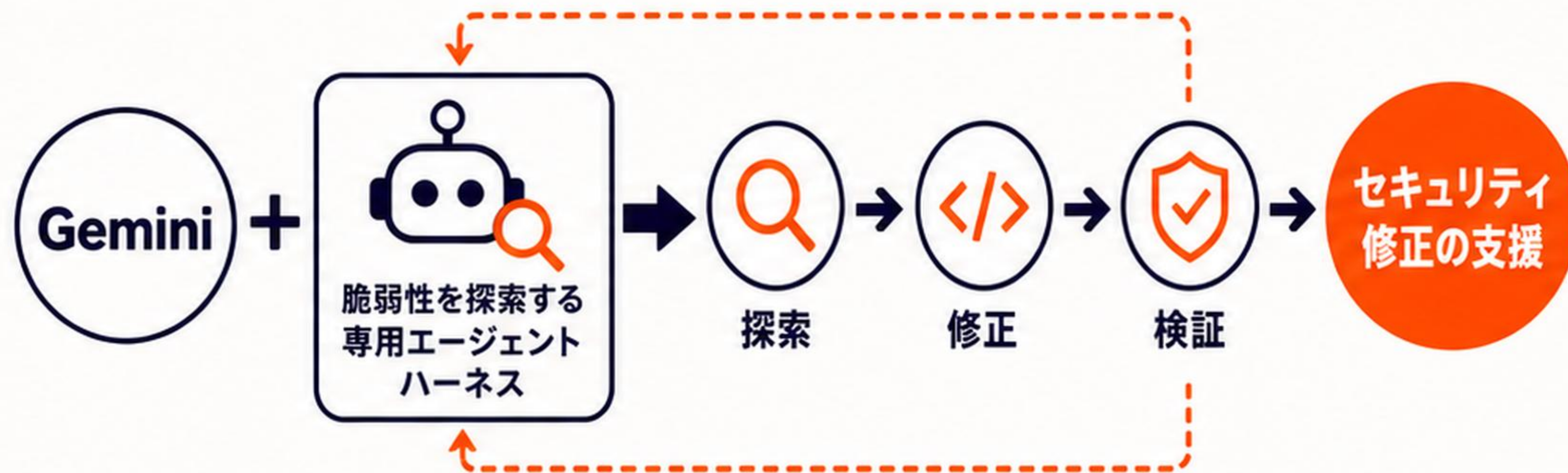


ローカルLLMの選択肢だけでなく、機密保持と速度を両立するマネージド推論経路が実用域に入った。

Gemini が Chrome の直近2リリースで1,000件超のセキュリティ修正を支援

Google は、直近2回の Chrome リリースでLLMが1,000件超のセキュリティバグ修正を支援し、それ以前23リリースの合計を上回ったと公表した。

- 直近2リリースで1,000件超のセキュリティバグ修正を支援
- その件数は、それ以前23リリースの合計を上回る
- Gemini単体ではなく、脆弱性を探索する専用エージェントハネスと組み合わせた

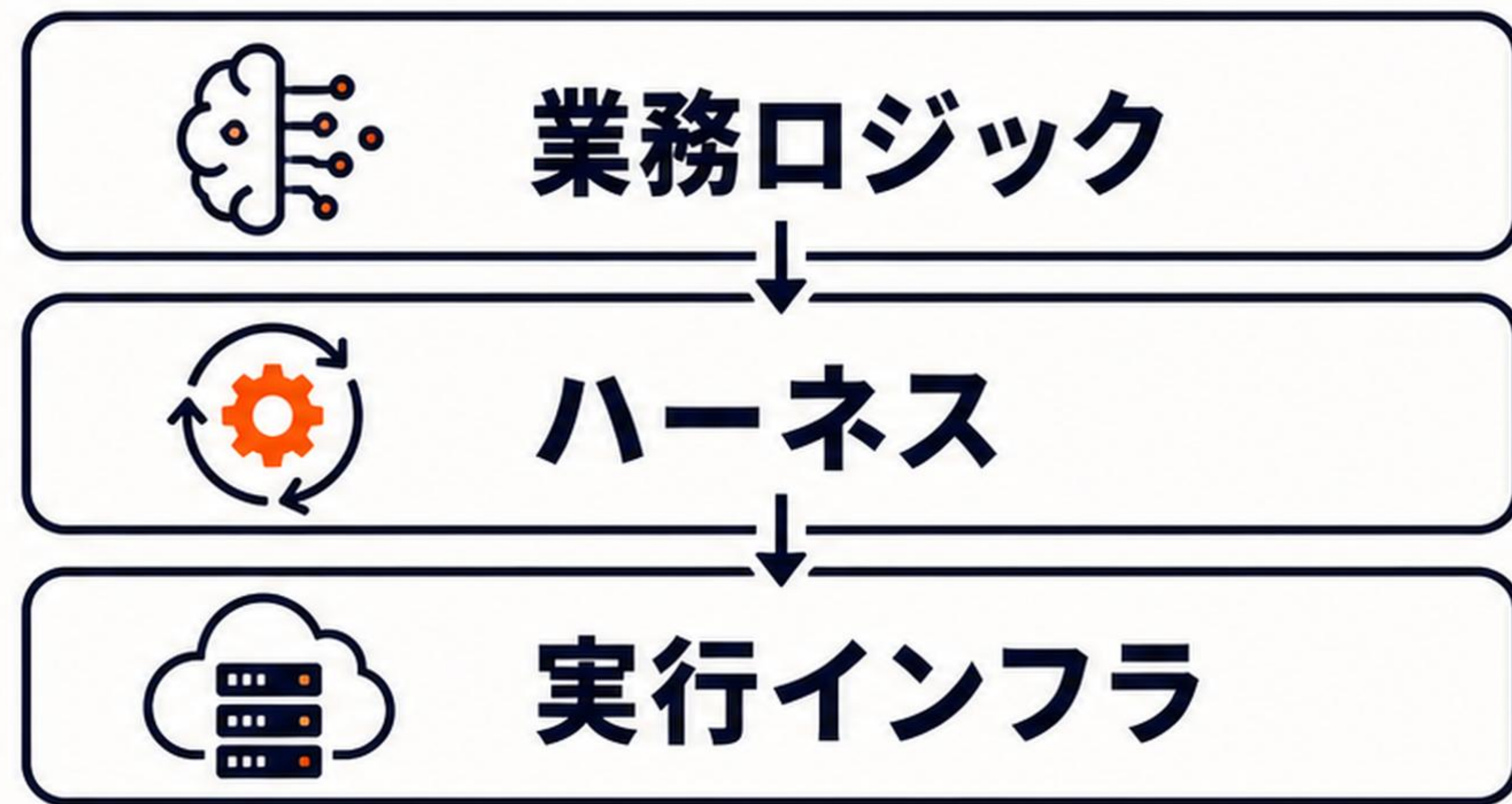


モデル単体ではなく、対象コードを探索・修正・検証する専用ハネスとの組み合わせが成果量を決める実例。

Harrison Chase が整理した「Managed Agent」の3層構造

LangChain共同創業者 Harrison Chase は、プロダクション用エージェントを「業務ロジック」「ハーネス」「実行インフラ」の3層に分解した。

- 業務ロジックは context・tools・instructions の組み合わせ
- ハーネスはLLMをループで動かし、ツール利用・計画・委任を担う
- 本番インフラには durable execution、checkpoint、streaming、sandbox が必要



エージェントをローカルで動かす段階から本番運用へ移すときに抜けやすい要件を、一枚のチェックリストとして使える。

公的ガイドをClaude Designのレビュー基準にするダッシュボード診断

デジタル庁の「ダッシュボードデザインの実践ガイドブック」を評価基準としてClaude Designに渡し、サンプル資料の診断シート作成からリデザインまで進める実践例が共有された。

- AIに「良くして」と頼まず、公的ガイドを評価ルーブリックとして渡す
- Claude Designに診断シートを作らせ、指摘からリデザインまで進める
- デジタル庁ガイドは要件整理・プロトタイプ・Do / Don'tを体系化



主観的なデザイン指示を、再現可能な検査項目へ変換する手法。

StreamCore

リアルタイム音声エージェントの媒体層をGoサーバー1つに集約

StreamCore は、WebRTC、割り込み、復旧、遅延制御といった音声エージェントの媒体層をまとめたApache 2.0のオープンソースGoサーバー。

- Go製のセルフホスト可能な音声エージェント基盤
- WebRTCとadaptive VADを内蔵
- ユーザーの割り込みを処理するclean barge-inと復旧を重視

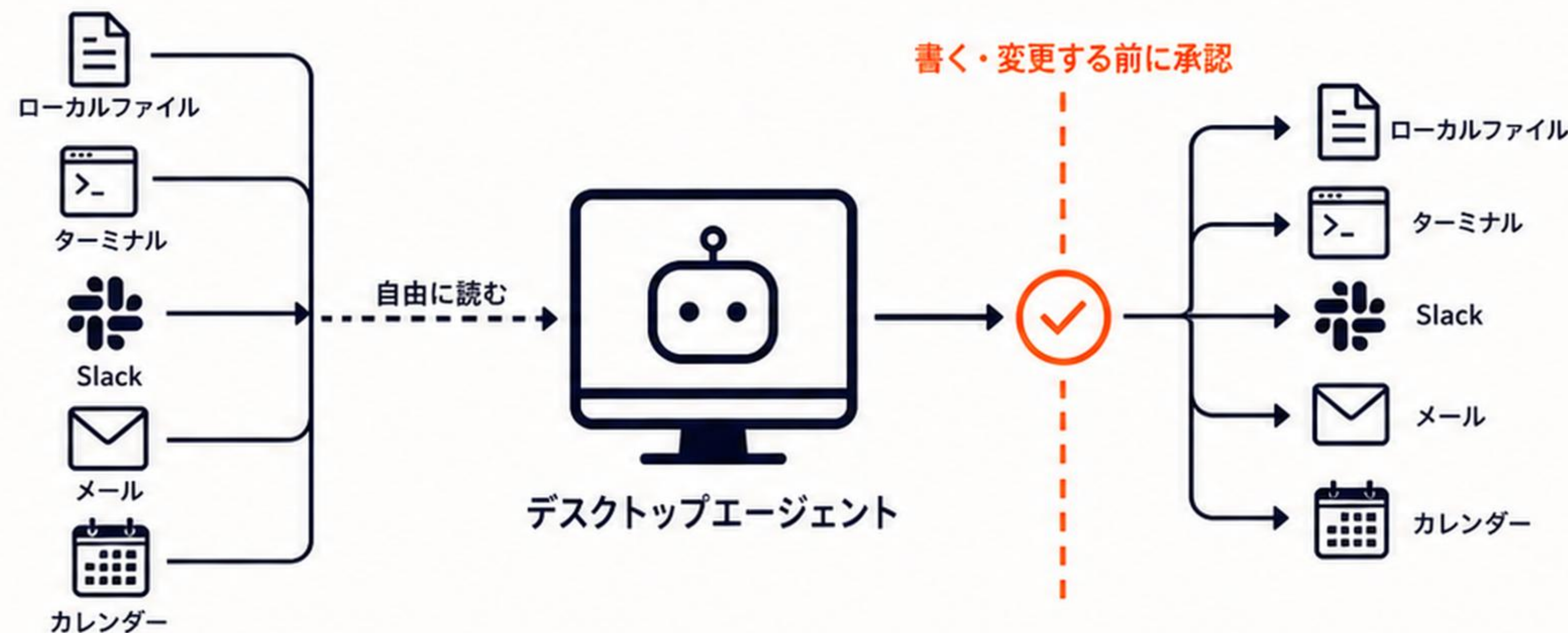


音声エージェントで難しいのはLLMの返答より、ユーザーが途中で話し始めたときに止めて再開する媒体制御。

〈知見〉 OpenWorker の「自由に読む・書く前に承認」境界

OpenWorker は、ローカルファイル、ターミナル、Slack、メール、カレンダー等を横断し、チャット回答ではなく文書や更新済み予定などの完成物を返すデスクトップエージェント。

- オープンソースでローカル優先のデスクトップエージェント
- OpenAI・Anthropic・Google等のAPI、またはOllamaを選択できる
- 会話ではなく文書・返信案・更新済み予定などの完成物を返す



デスクトップエージェントの有用性と安全性を両立するには、ツールごとの細かな許可より「読む」と「外部状態を変える」を分離するほうが説明しやすい。

本日のトピックス一覧

1. Ollama Cloud の DeepSeek-V4-Flash が200 tok/s超へ、ZDRを維持
2. Gemini が Chrome の直近2リリースで1,000件超のセキュリティ修正を支援
3. Harrison Chase が整理した「Managed Agent」の3層構造
4. 公的ガイドを Claude Designのレビュー基準にするダッシュボード診断
5. StreamCore
6. 〈知見〉 OpenWorker の「自由に読む・書く前に承認」境界