Tech News Daily
MORNING DISPATCH

2026-10-30

Mistral、1兆パラメータの「Mistral Large 4」改良版を公開重みで公開

ChatGPT Pro と Business に毎月最大500ドルの API クレジット。Pro と Business…

OpenAI Codex 0.166.0、会話の記録をキーボードで選んで検索…

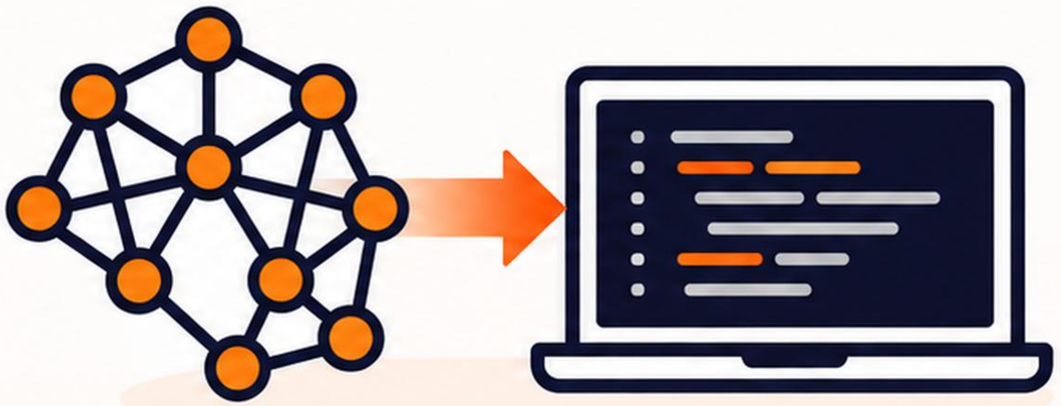
注目3件

TEKION Group

Mistral、1兆パラメータの「Mistral Large 4」改良版を公開重みで公開

Mistralが、10月上旬にパブリックプレビューで出したフラッグシップモデル「Mistral Large 4（愛称 Le Chonk）」を強化学習で3週間追加で鍛え、重みを公開しました。

- CognitionのFrontierCode 2.0で19.7%と、公開重みモデルで首位。Terminal-Bench…
- Artificial AnalysisのCoding Agent Indexは48%で、公開重みモデル首位のGLM-5.3と並ぶ…
- 推論の深さをLow・Medium・High（既定）の3段階から選べる

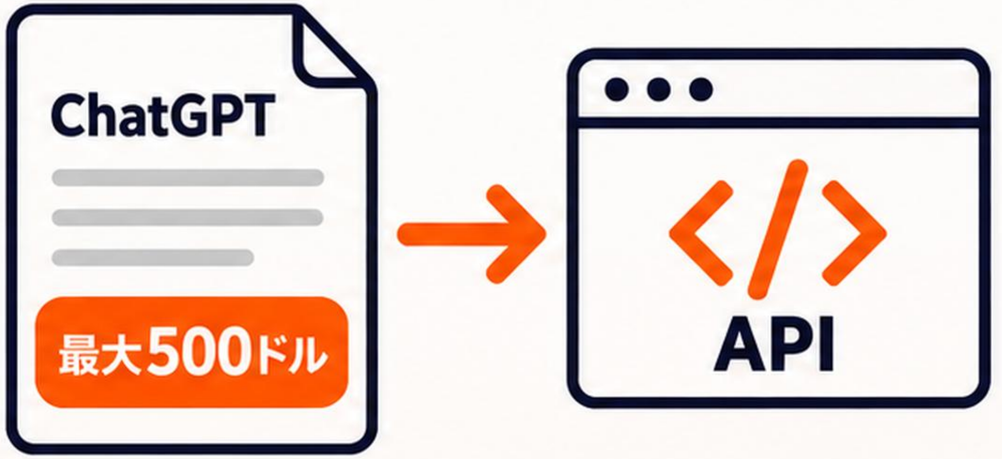


自社の環境で動かせる公開重みモデルが、コーディングエージェントの用途で上位に入ってきました。

ChatGPT ProとBusinessに毎月最大500ドルのAPIクレジット。ProとBusiness Premium…

OpenAIが、条件を満たすChatGPT ProとBusinessの契約者に、アプリやエージェントを作るためのAPIクレジットを毎月最大500ドル付与し始めました。

- Proのクレジットはプランに応じて毎月100・200・500ドル
- Businessは有料のStandard席1つにつき20ドル。Premium席1つにつき100ドル。ワークスペースごとに最大500ドルまで合算される
- 「条件を満たす」の細かい条件は告知に書かれていない。自分が対象かは、ChatGPTの請求設定の画面で確かめる



ChatGPTの有料プランだけで、APIを使った試作まで手が届くようになりました。

OpenAI Codex 0.166.0、会話の記録をキーボードで選んで検索。安全機能が止めた理由も分かりやすく

OpenAIのコーディングエージェント「Codex」のCLIが0.166.0になりました。

- 会話の記録をキーボードで選択・検索できる。ショートカットは設定で変えられる
- 安全機能で止まったとき、TUIに理由の説明と、選べる安全な次の指示が出る
- `codex doctor --network-only` で、設定・接続・WebSocketだけを絞って診断できる



長い作業の記録を見返したり、つながらないときに原因を切り分けたりする手間が減ります。

Geminiアプリで、文書・PDF・スライド・表計算をチャットの中で作って、そのまま編集できるように

GoogleのGeminiアプリで、文書・PDF・プレゼン資料・表計算・コードなどのファイルをチャットの中で作り、プレビューして編集まで済ませられるようになりました。

- PDFのプレビューと編集はWebとモバイルの両方で使える
- Googleドキュメント、Googleスプレッドシート、スライド、Markdown、TXT、コードはWebのみ
- 全画面での直接編集、ハイライトした部分への質問、Geminiによる文章や図の手直しができる

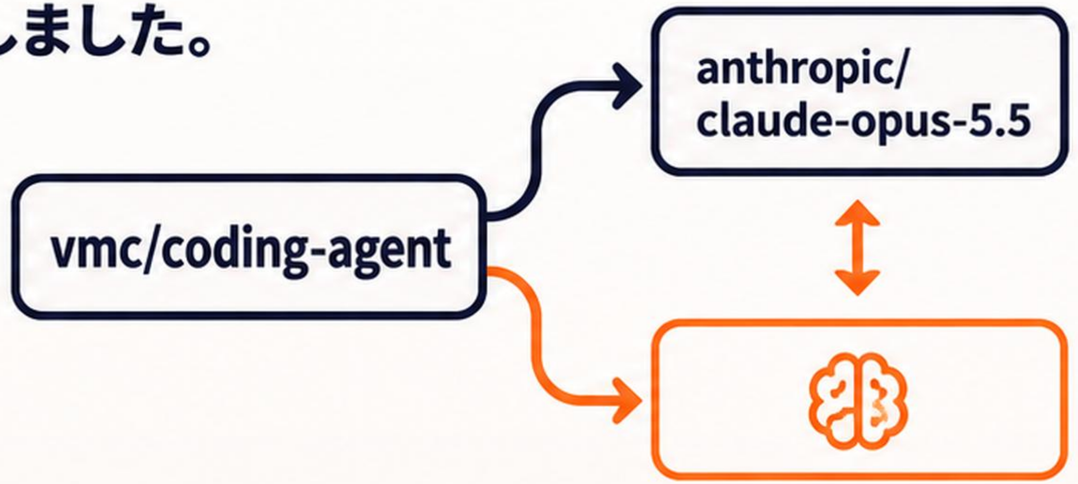


AIが作った資料を別のアプリに移さず、チャットの中で仕上げまでできるようになりました。

Vercel AI Gateway に「Virtual Models」。1つの ID の裏のモデルや設定を、コードを変えずに切り替え

VercelがAI Gatewayに「Virtual Models」を追加しました。

- 仮想モデルIDごとに、使うモデル、フォールバックと優先順位、Zero Data Retention（データを保存しない設定…
- 作業ごとにIDを作る使い方を想定。例は `vmc/coding-agent` を作って、モデルを `anthropic/claude-opus-5.5`…
- Vercel CLIの `vercel ai-gateway virtual-models create` と `edit` で作成・変更できる

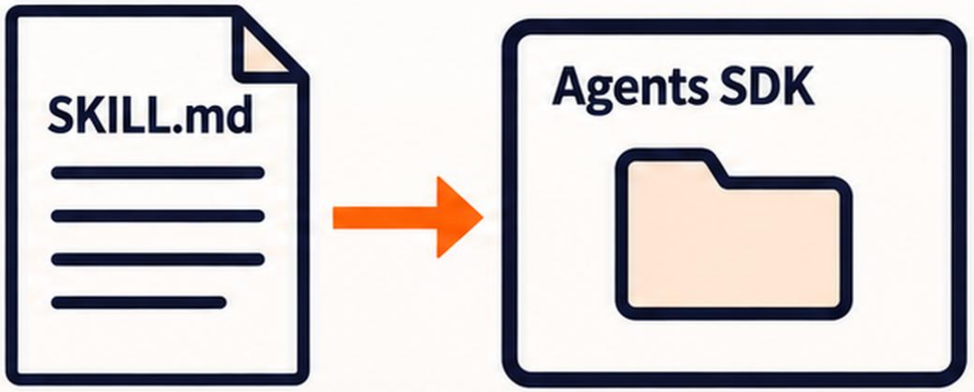


新しいモデルが週に何本も出るなかで、モデル名をコードに直接書いていると、試すたびに書き換えと再デプロイが要ります。

Cloudflare Agents SDK、エージェントの「スキル」をファイルとして置くだけで使えるように

CloudflareのAgents SDKで、Agent Skills（エージェントに手順や知識を渡す「スキル」）をファイルとして扱えるようになりました。

- スキルの読み込み元は、作業領域（Computer Workspace）、R2バケット、コードの3種類。毎回の手番の前に一覧を読み直す
- SKILL.mdはAgent Skillsの仕様に照らして確かめられ、ルールに反していても警告付きで読み込まれる…
- スキルに含まれるスクリプトは作業領域のファイルとして置き、モデルが作業領域の実行ツールで動かす

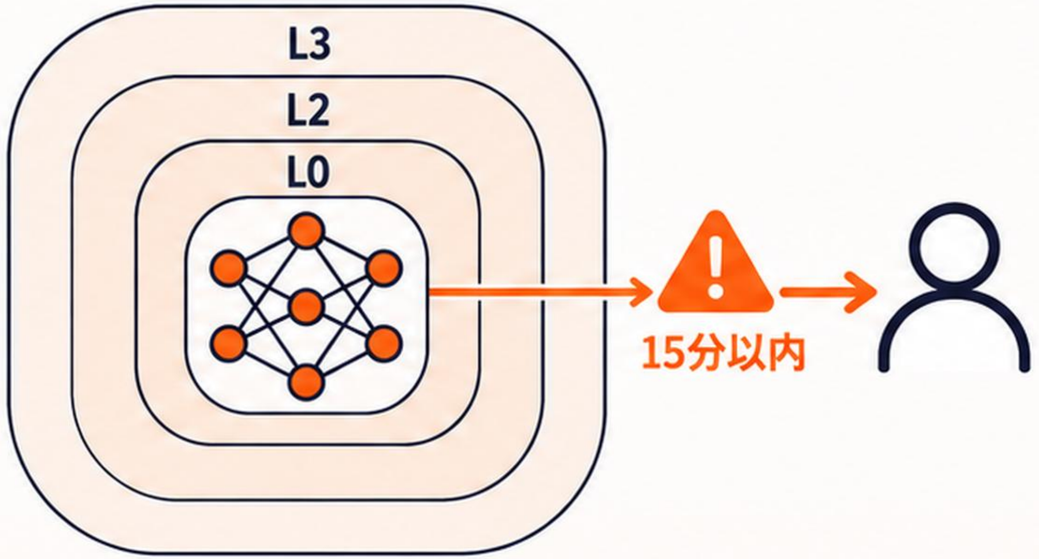


スキルの書式が、各社のエージェント基盤で「ファイルを置くだけ」の形で使われ始めています。

OpenAI、訓練中のモデルの「想定外の行動」を見張る仕組みを公開…

OpenAIが、最先端モデルの訓練・評価・社内利用の最中に、モデルが意図しない行動をとっていないかを見張る仕組みを公開しました。

- 監視は4層。L0はモデル内部の活動から怪しいパターンを拾う軽い分類器（activation probe）…
- L2は記録全体を読み返して経緯を組み立て直す調査用エージェント、L3はネットワークのログなど会話の外の証拠と…
- 呼び出しが有効な作業では、最優先の警告が通常15分以内にセキュリティと訓練の担当者に届く



エージェントに任せる範囲が広がるほど、越えてはいけない線を越えたときにすぐ気づける仕組みが要ります。

〈知見〉Anthropic、Claude 自身が会話を振り返って悪用の兆候を知らせる「self-assessment」

Anthropicが新しい「Safeguards」ブログで、Claudeが返答した直後に、その会話が安全かどうかを自分で簡単に評価する技術「self-assessment」を紹介しています。

- Claude Fable 5では、より強い分類器が「害の可能性あり」とした会話の78%、優先度の高いものの94%を…
- 評価には、害のある依頼かどうか、害の種類、人が確認する優先度、Claudeが断ったかどうか、判断の理由1文が含まれる
- 会話がすでにKVキャッシュに載っているので、短い指示と数トークンの出力だけで評価でき、計算が少なくて済む



AIの安全確認を、外付けの分類器だけでなく、AI自身の判断にも担わせる考え方の具体例です。

本日のトピック一覧

1. Mistral、1兆パラメータの「Mistral Large 4」改良版を公開重みで公開
2. ChatGPT Pro と Business に毎月最大500ドルの API クレジット。Pro と Business…
3. OpenAI Codex 0.166.0、会話の記録をキーボードで選んで検索…
4. Gemini アプリで、文書・PDF・スライド・表計算をチャットの中で作って…
5. Vercel AI Gateway に「Virtual Models」。1つの ID の裏のモデルや設定を…
6. Cloudflare Agents SDK、エージェントの「スキル」をファイルとして置くだけで使えるように
7. OpenAI、訓練中のモデルの「想定外の行動」を見張る仕組みを公開…
8. 〈知見〉Anthropic、Claude 自身が会話を振り返って悪用の兆候を知らせる「self-assessment」