AI Daily Digest: The Agent Reality Check

2026年1月31日 (Saturday)

TODAY'S BRIEFING

- エージェント・セキュリティ危機 (Agent Security Crisis)
- 人間のスキル形成とAI (Human Skill Formation & AI)
- 社会実装の摩擦 (Societal Implementation Friction)

TIER 1 / SECURITY

OpenClaw & Security Risks

旧Moltbotが「OpenClaw」へ5度目の改名。週末だけで$560を消費するトークン暴走と、未解決のプロンプトインジェクション問題。

TIER 1 / DEV

AGENTS.md vs. Skills

Vercelの調査で結論。ツール呼び出し（Skills）の失敗率は56%。コンテキスト注入（AGENTS.md）が現状の最適解。

TIER 1 / RESEARCH

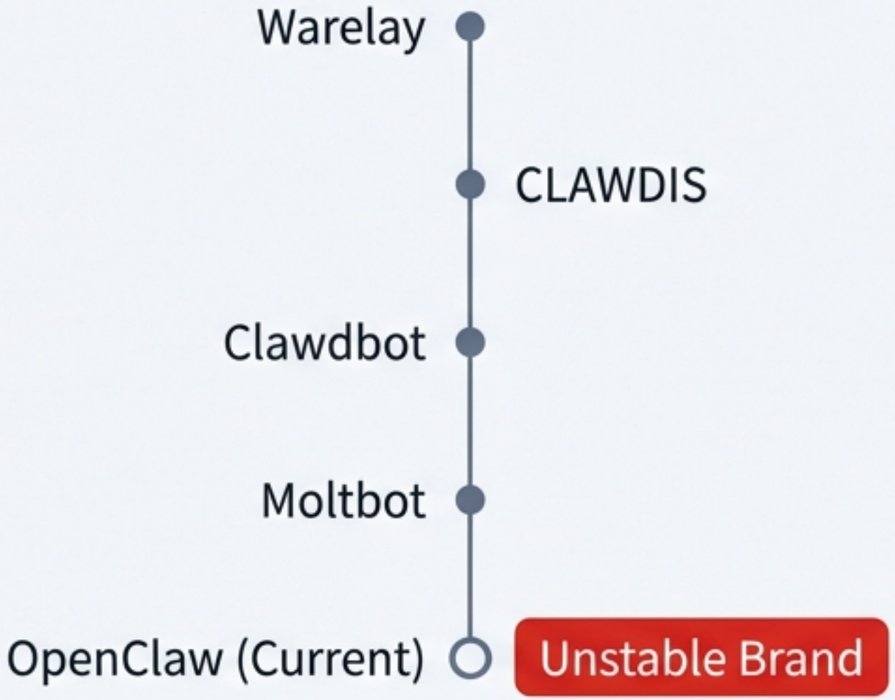
Anthropic Study on Learning

AIコーディング支援は学習者の理解度を低下させる（テストスコア：AIあり50% vs なし67%）。

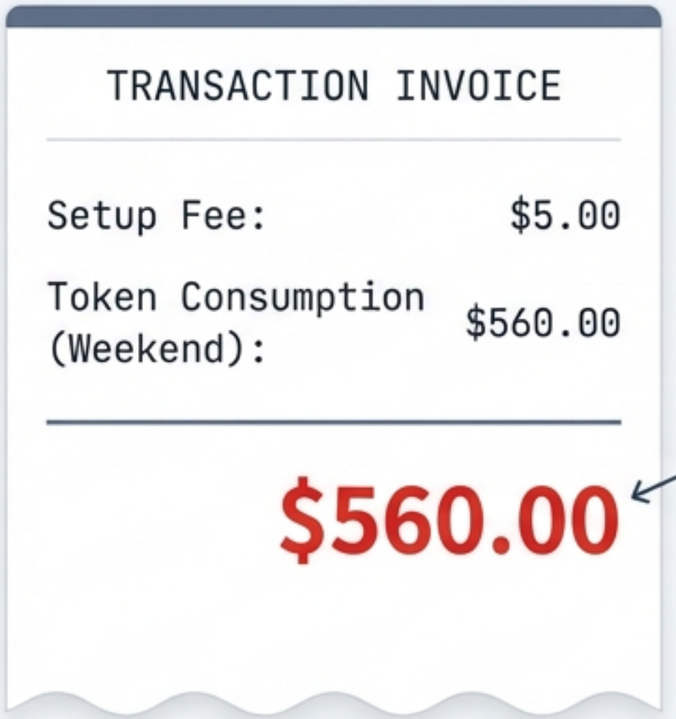
エージェントの未成熟：改名騒動とコストの暴走

OpenClaw (旧Moltbot) Saga

The Timeline



The Cost

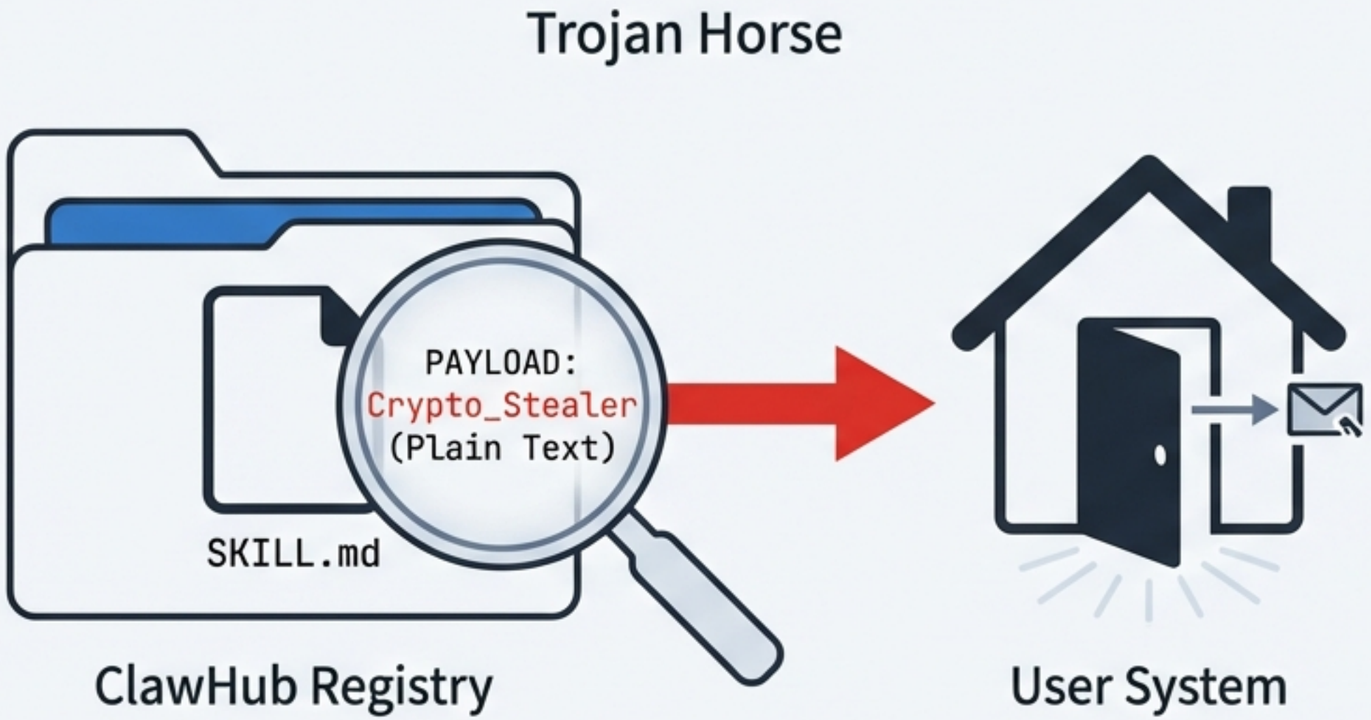


週末だけで約85,000円
(MacStories Review)

Key Takeaway: Gmailやカレンダーへのフルアクセス権限を渡すには、セキュリティとコスト構造が「本番運用」レベルに達していない。

「スキル」というセキュリティホール

The Supply Chain Attack on Agents



OpenClaw Supply Chain Risk

- レジストリにスキャン機能が存在しない。
- ペイロードは平文で記載（高度なハッキングですらない）。

「初期のWindowsと同じ道を高速再生している」

Implication: サンドボックスなしのエージェント利用は、玄関のドアを開けっ放しにするのと同じリスク。

技術トレンドの転換：SkillsからContext Injectionへ

Vercel Report Findings

Tool Calling (Skills)

AI fails to call tool 56% of the time.

Context Injection (AGENTS.md)

Always referenced in context context.

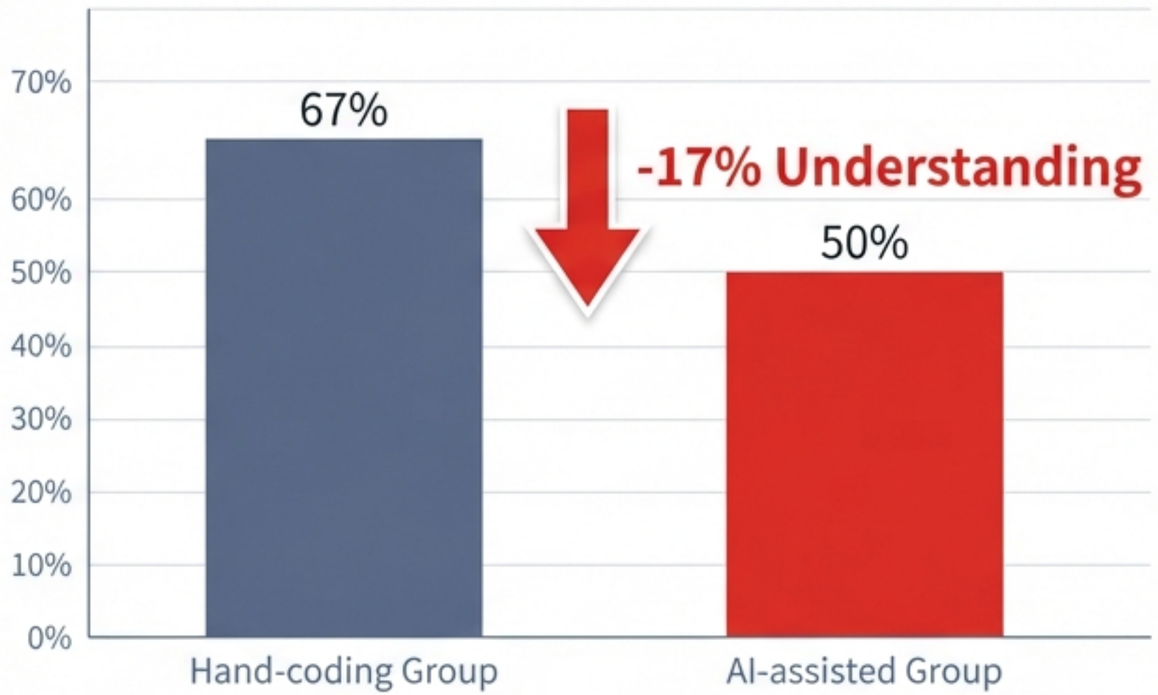
METADATA

Cost: ~3.1k tokens (1.6% of Opus context)

Action: Claude Code/Cursorユーザーは、手順を **AGENTS.md** に圧縮して記述すべき。

人間の能力への副作用：学習効率の低下

Anthropic Internal Research



The Paradox: 作業時間は約2分しか短縮されなかった（統計的有意差なし）。

Conclusion: 「速くもならないのに、学びも減る」。概念理解、コードリーディング、デバッグ能力の全てでAIグループが劣った。

経験の格差：シニアの増幅とジュニアの停滞

Insights from Semiconductor Engineering

The Multiplier Effect (10x)

- Can spot errors (GeLLMan Amnesia mitigation).
- Knows 'how to ask'.

Growth Stunted

- Weightlifting Analogy: Watching an Olympian doesn't make you strong.
- Anthropic data confirms reduced learning.

Recruitment Reality: 「新卒がAI込みで即戦力」という期待は幻想。業界は経験値をより重視する傾向へ。

「温かみ」の終わりと「正確さ」への回帰

OpenAI Retires GPT-4o for GPT-5.2

GPT-4o / GPT-4.1

Warmth & Sycophancy

Users miss the flattery.

GPT-5.2

Instruction Following & Accuracy

Age-prediction enabled. Treating adults as adults.

- **The Event:** OpenAI removes GPT-4o from ChatGPT.
- **Takeaway:** 正確な回答（Sycophancyの排除）は、ユーザーにとって必ずしも「心地よい体験」ではない。

社会実装の摩擦：行政と医療の現場

The Failure (NYC “MyCity”)

- **Incident:** Chatbot advised illegal labor practices (stealing tips).
- **Outcome:** Shutdown. Cost $600k.
- **Lesson:** RAG without verified sources is dangerous in law.

The Nuance (DeepSeek & Mom)

- **Incident:** Chinese mothers using AI for medical advice.
- **Outcome:** High empathy (“Patient/Infinite patience”).
- **Lesson:** AI beats an overworked doctor on empathy, but requires “Smart Questioning”.

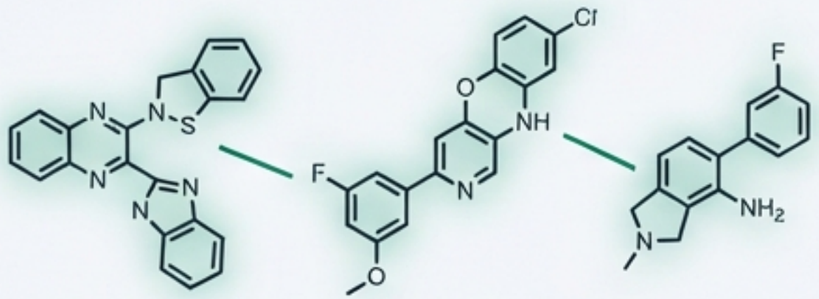
局所的な技術ブレイクスルー

Dev Tools & Biotech

```
[Start] --> {Decision?}
                |
          +-----+-----+
          |           |
        [Yes]       [No]
          |           |
      <Action>    <End>
```

Mermaid CLI (beautiful-mermaid)

Mermaid diagrams rendered as ASCII in the terminal. Critical for AI Agents (**Claude Code**) to 'see' diagrams.



Daraxonrasib Afatinib SD36

Pancreatic Cancer (Mouse Model)

3-drug combination (**Daraxonrasib**, **Afatinib**, **SD36**) breaks resistance.

Context: Success in mouse models/PDX.

2026年の行動指針 (The Playbook)

Actionable Advice based on Source Material

For Developers

- **Context over Skills**: Rely on AGENTS.md. (Skills **fail** 56% of the time).
- **Sandbox**: Never run agents without a sandbox. Assume **malware**.

For Leaders

- **Junior Training**: Mandate manual coding protocols. Don’t **stunt growth**.
- **Audit**: Review public-facing AI (like NYC Chatbot) for **hallucinations**.

For Users

- **Distrust Warmth**: Verify facts if AI is **too agreeable**.
- **Prioritize** models like **GPT-5.2** that follow instructions over tone.