

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

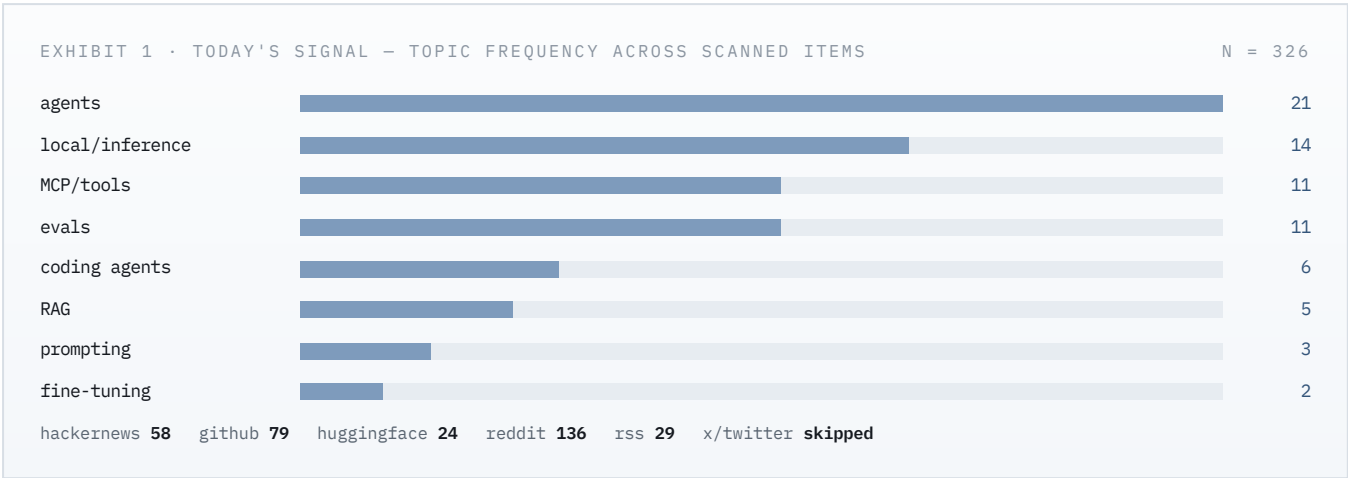
Today in AI engineering

2026-09-10

Daily AI-engineering signal (mechanical fallback — LLM error: 404 Client Error: Not Found for url: <https://api.groq.com/openai/v1/chat/complet>)

5 deep dives 3 builder stories 15 quick hits 326 items scanned mode: **mechanical fallback**

- agents
- local/inference
- mcp/tools
- evals
- coding agents
- rag
- prompting
- fine-tuning



DEEP DIVES

Show HN: Self-hosted company OS, Claude Code and Codex agents in departments

Show HN

Hi HN, I am Dimitris. This is a company OS that I built and use to run my business and anyone can install it and self host it for free. Think of it as Claude Code, Cowork and the cloud sessions in one self hosted application. It is a Multi-tenant application by design where many people can collabora

Why it matters. Flagged by the ranking heuristics as a real, artifact-backed build.

TRY THIS

Open the link and skim the repo / paper.

<https://github.com/OtoDock/oto-dock>

okf-memory/okf-agent-memory

[GitHub](#)

Git-native persistent memory for AI coding agents. Implements Google OKF v0.2 with sub-300µs in-memory BM25 search, embedded MCP server, and progressive disclosure. Slashes token bloat by 80% with zero external databases or dependencies. Built in pure Go.

Why it matters. Flagged by the ranking heuristics as a real, artifact-backed build.

TRY THIS

Open the link and skim the repo / paper.

<https://github.com/okf-memory/okf-agent-memory>

noskillish/bankmcp

[GitHub](#)

BankMCP™: your AI can now read your bank. Self-hosted, read-only MCP server for your own bank accounts via open banking (Enable Banking). Standard MCP; tested with Claude and Ollama.

Why it matters. Flagged by the ranking heuristics as a real, artifact-backed build.

TRY THIS

Open the link and skim the repo / paper.

<https://github.com/noskillish/bankmcp>

Qwen 3.8 follows GPT-5.5 Pro reasoning prefills

[HN](#)

(no summary)

Why it matters. Flagged by the ranking heuristics as a real, artifact-backed build.

TRY THIS

Open the link and skim the repo / paper.

<https://gist.github.com/wsxiaoy/e0286dc6bb624ff5fdf49e7f4c528ba3>

PillCrew/deedchain

[GitHub](#)

Benchmark of report fidelity for browser agents: does the agent's final report tell the truth about what it actually did?

Why it matters. Flagged by the ranking heuristics as a real, artifact-backed build.

TRY THIS

Open the link and skim the repo / paper.

<https://github.com/PillCrew/deedchain>

WHAT PEOPLE ACTUALLY BUILT

Show HN: AgentPulse – Claude Code and Codex status in tmux

[Show HN](#)

I use Claude Code and Codex every day, often across several tmux panes. I kept switching windows to check which agents were still working and which needed an answer. I built AgentPulse so I could see that at a glance. It is a TPM plugin that uses lifecycle hooks to show working,

STEAL THIS

Flagged by keyword heuristics as a first-person build write-up — read the post for the details.

<https://github.com/jerriclynsjohn/tmux-agent-pulse>

Show HN: Persistent Jupyter kernel execution and live output streaming in VSCode

[Show HN](#)

Hi HN, Every time I ran a training loop for a few days via Jupyter Notebook and ran into issues (network disconnected, my laptop suddenly died, etc.), I suffered from losing my ipykernel and cell outputs. I wanted to keep seeing my live outputs right inside the IDE without consta

STEAL THIS

Flagged by keyword heuristics as a first-person build write-up — read the post for the details.

<https://github.com/inozo/tithon>

Show HN: Compute polynomials twice as fast

[HN](#)

A few years ago my coauthor and I was wondering if we could reduce the number of multiplications used for hashing algorithms. We had a construction and a 100 page proof, but we were not 100% sure it was correct. Now we have a full Lean proof, so we decided to publish it. I made t

STEAL THIS

Flagged by keyword heuristics as a first-person build write-up — read the post for the details.

<https://thomasahle.com/fast-polynomials/>

QUICK HITS

hsj576/virtual-ai-infra-team [GITHUB](#)

An open-source virtual AI Infra team that discovers, benchmarks, and safely upgrades local LLM inference—with independent quality gates, a s
<https://github.com/hsj576/virtual-ai-infra-team>

Show HN: Geiger – See every AI agent on your machine and what it can touch [SHOW HN](#)

<https://github.com/Atomburstofficial/geiger>

Show HN: Hydra – Open-source agentic terminal with a PTY daemon [SHOW HN](#)

<https://github.com/hydraterm/hydra-local>

openbmb/MiniCPM5-2B [HF](#) [MODEL](#)

Trending model on Hugging Face— text-generation, transformers; 1013 likes, 42,289 downloads.
<https://huggingface.co/openbmb/MiniCPM5-2B>

XHToken/Spark-X2.5-4B [HF](#) [MODEL](#)

Trending model on Hugging Face— text-generation, transformers; 1029 likes, 15,930 downloads.
<https://huggingface.co/XHToken/Spark-X2.5-4B>

Claude, change the “Add to Cart” button to blue [HN](#)

<https://opusfived.dev/>

Show HN: Self-host open-source LLMs on AWS with scale-to-zero [SHOW HN](#)

<https://github.com/paguasmar/veloxml-deploy>

deepseek-ai/DeepSeek-V4.1-Flash [HF](#) [MODEL](#)

Trending model on Hugging Face— image-text-to-text, transformers; 667 likes, 6 downloads.
<https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash>

Show HN: AI-first open-source alternative to Microsoft and Google Office [SHOW HN](#)

InjOffice — AI-first open-source alternative to Microsoft Office and Google Workspace
<https://github.com/injectinglabs/injoffice>

Show HN: What if the speed of light was 5 km/h? SHOW HN

I've always wanted to make a visualization where the speed of light was scaled down to human speeds, so that we could intuit relativistic effects
<https://rivendell.dmitrybrant.com/relativity/>

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 14566 likes, 7,322,476 downloads.
<https://huggingface.co/Qwen/Qwen3.8-27B>

ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF HF MODEL

Trending model on Hugging Face — image-text-to-text, gguf; 754 likes, 614,850 downloads.
<https://huggingface.co/ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF>

chadhurley25075-png/pd-bridge GITHUB

Heterogeneous prefill/decode for DeepSeek-V4-Flash: CUDA prefill (DGX Spark, vLLM) -> Metal decode (Mac Studio, oMLX) over plain 10GbE
<https://github.com/chadhurley25075-png/pd-bridge>

nex-agi/Nex-N2.5-mini HF MODEL

Trending model on Hugging Face — text-generation, transformers; 619 likes, 2,444 downloads.
<https://huggingface.co/nex-agi/Nex-N2.5-mini>

aaravshah1311/Agent-2 GITHUB

Agent2 — An autonomous AI agent for coding, cybersecurity, terminal automation, and intelligent task execution.
<https://github.com/aaravshah1311/Agent-2>
