

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

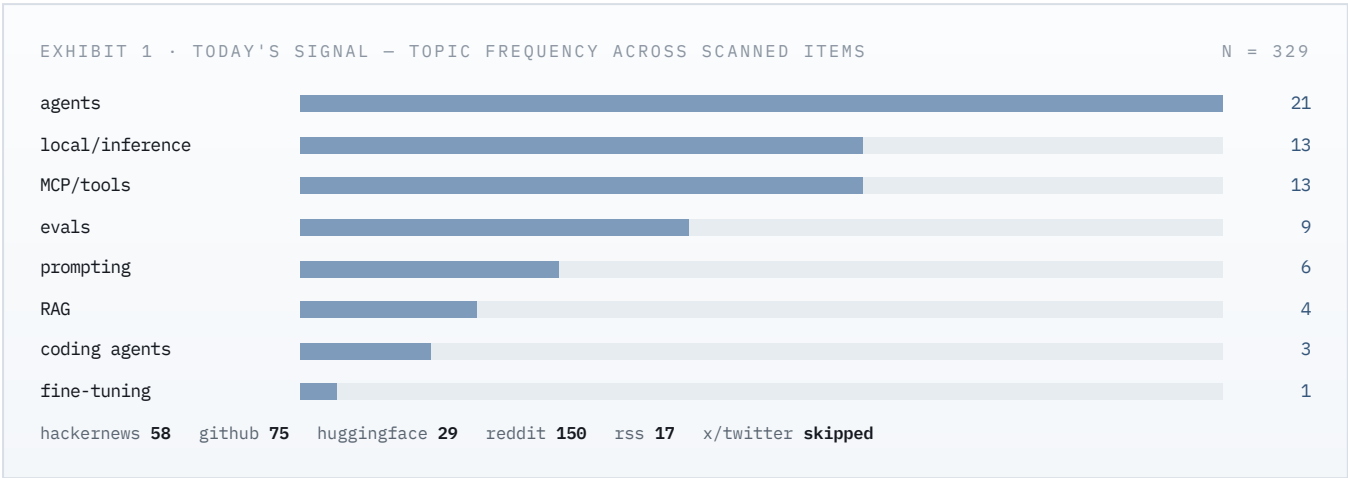
Today in AI engineering

2026-09-02

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 329 items scanned mode: **mechanical fallback**

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



DEEP DIVES

Show HN: Running 104GB Qwen3.8-Flash-Next on 48GB Mac with at ~12 tok/s

Show HN

I built slotstream, a way to run Qwen3.8-Flash-Next 4-bit on a low-memory mac starting from 16GB, a 125B parameter model that would need 100GB+ memory/RAM, thanks to expert-offloading/ssd-streaming. Easy to install/update, and mac-native using MLX and Swift. It ships with auto-mode, which makes a 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/carloslifu/slotstream>

Ryze-AI-Adgent/open-seo-mcp-skills

[GitHub](#)

Open-source SEO + GEO skills for Claude — keyword research, rank tracking, site audits, backlinks, competitor gaps, AI visibility. Runs on your real Search Console / GA4 / ads data via MCP, with DataForSEO built in. Free, MIT.

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/Ryze-AI-Adgent/open-seo-mcp-skills>

Qwen/Qwen3.8-Flash-Next

[HF model](#)

Trending model on Hugging Face — image-text-to-text, transformers; 4687 likes, 207,941 downloads.

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://huggingface.co/Qwen/Qwen3.8-Flash-Next>

zai-org/GLM-5.3-Flash

[HF model](#)

Trending model on Hugging Face — image-text-to-text, transformers; 1912 likes, 441,348 downloads.

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://huggingface.co/zai-org/GLM-5.3-Flash>

zai-org/GLM-5.3

[HF model](#)

Trending model on Hugging Face — text-generation, transformers; 1481 likes, 94,403 downloads.

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://huggingface.co/zai-org/GLM-5.3>

WHAT PEOPLE ACTUALLY BUILT

Show HN: Weedout – Safari extension that hides YouTube AI-labeled videos

[Show HN](#)

I'm the developer. I built this because AI-generated conspiracy videos were getting out of control in my own YouTube feed. Weedout is a \$1.99 Safari extension for macOS that removes videos YouTube labels “Made with AI” from feed, search, related videos, playlists and Shorts. It u

STEAL THIS

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

<https://masteranza.github.io/weedout/>

Show HN: Shed – Git Repo Management for Terminal Agents

[Show HN](#)

Hi HN, Shed is a git repo management system for terminal agents. I built it because I wanted a complete git repo and workspace management system that was agent-first by design and portable across agent harnesses. Tools like ghq, worktrunk, and vanilla git worktrees fill some of t

STEAL THIS

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

<https://github.com/AndrewHannigan/shed>

Show HN: Kaden Terminal – Bloat / AI Free Terminal for Linux

[Show HN](#)

Kaden is a Terminal Emulator and fork of warp with clipboard support and configurable keybindings, I made Kaden because every other Linux terminal lacks proper text selection. It's impossible to select user input with a single hotkey on Linux, requiring mouse use and awkward mult

STEAL THIS

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

<https://github.com/glitcher255/Kaden-Terminal>

QUICK HITS

Show HN: OSS, K8s-native AI platform for distributed multi-model inference [SHOW HN](#)

Hi everyone, I'm one of the co-founders of axem. We recently open sourced Shaide, a project we've been working on to make running multiple L

<https://github.com/axem-solutions/shaide>

JordyZomer/lemmalog [GITHUB](#)

A Datalog engine for LLM agent memory: stratified rules, provenance-tracked facts, incremental derivation, and an MCP server that lets your

<https://github.com/JordyZomer/lemmalog>

Claude Fable 5.1 and Claude Mythos 5.1 [HN](#)

What's new in Claude Fable 5.1 — <https://platform.claude.com/docs/en/models/fable-5-1/whats-n...> System Card: <https://www-cdn.anthropic.com/>

<https://www.anthropic.com/claude-fable-and-mythos-5-1>

Vuk97/forward-implementation-first [GITHUB](#)

Stop your coding agent from stalling real work on self-invented bookkeeping - receipts, hashes, locks, certification rituals. Ship first, th

<https://github.com/Vuk97/forward-implementation-first>

unsloth/Qwen3.8-Flash-Next-GGUF [HF](#) [MODEL](#)

Trending model on Hugging Face — image-text-to-text; 697 likes, 431,339 downloads.

<https://huggingface.co/unsloth/Qwen3.8-Flash-Next-GGUF>

Qwen/Qwen3.8-27B [HF](#) [MODEL](#)

Trending model on Hugging Face — image-text-to-text, transformers; 13622 likes, 4,960,483 downloads.

<https://huggingface.co/Qwen/Qwen3.8-27B>

rokbenko/quackd [GITHUB](#)

 Give your Microduck a brain. Tell a small robot with two legs what you want in plain language. An LLM (Claude, OpenAI, Gemini, Grok, or a

<https://github.com/rokbenko/quackd>

deepseek-ai/DeepSeek-V4-Flash-Vision-Exp [HF](#) [MODEL](#)

Trending model on Hugging Face — image-text-to-text, transformers; 474 likes, 17,893 downloads.

<https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-Vision-Exp>

Show HN: Eatheria – Self-hosted AppSec platform with AI false-positive filtering SHOW HN

<https://github.com/MartinCrespoC/Security-Team---Workspace->

sunyuzheng/lizheng-open-context GITHUB

Source-grounded public context for 立正: Superlinear posts, selected comments, YouTube, Public Axioms V1, and 《真本事》.

<https://github.com/sunyuzheng/lizheng-open-context>

Show HN: Newton's Orchard – Browser-based space/gravity playground SHOW HN

Hi HN! My son spent a lot of time in 5th grade playing with the 2-dimensional PhET gravity/space sims[1]. His STEM class uses it, and he was

<https://newtonsorcharard.app>

I trained a small transformer in 1.5hrs and it beats many LLMs HN

<https://mvakde.github.io/blog/44-on-arc-1/>

tencent/Hy4-preview HF MODEL

Trending model on Hugging Face — text-generation, transformers; 391 likes, 3,516 downloads.

<https://huggingface.co/tencent/Hy4-preview>

baidu-baige/LoongSage GITHUB

A production-grade, high-efficiency Agentic RL framework for frontier LLMs with validated recipes.

<https://github.com/baidu-baige/LoongSage>

ashutoshsinghpr7/wikiskill GITHUB

WikiSkill (arXiv:2608.27454) for Hermes Agent — self-evolving agent skills via a persistent knowledge wiki. Faithful Algorithm 1 implementat

<https://github.com/ashutoshsinghpr7/wikiskill>
