

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

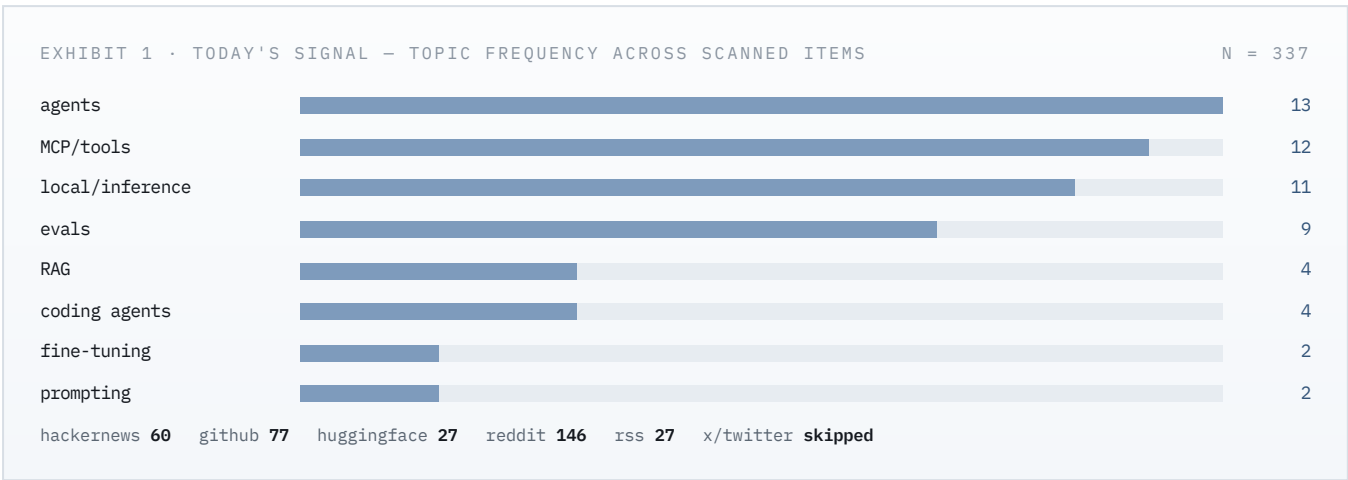
Today in AI engineering

2026-09-03

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 337 items scanned mode: **mechanical fallback**

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



DEEP DIVES

Show HN: Aura – a Rust agent that investigates and fixes production incidents

Show HN

We run a SaaS that handles petabytes of data. Our SRE team experimented with using claude, openclaw, langchain, etc. within our incident response workflows. We struggled with overflowing context, lethal trifecta vectors, hallucinations, and burned a lot of frontier tokens mostly on easy work. Approv

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/mezmo/aura>

Show HN: ZSvirt – A lightweight, scalable open source virtualization platform

[Show 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://github.com/ZSvirt/zsvirt>

Show HN: I built a version of Omarchy that runs on Apple Silicon

[Show HN](#)

I wanted to try out Omarchy before doing a full install, but I only have an M series MacBook so I would have to find a laptop and there's no official M support yet. So I decided to create something that allows people to test and run Omarchy on their devices so they can understand how it feels. So I

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/themartiano/try-omarchy>

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>

carloslfu/slotstream

[GitHub](#)

Run Qwen3.8-Flash-Next (125B MoE, 104 GB at 4-bit) on Macs with a fraction of that RAM by streaming experts from SSD. MLX + Swift, Ollama-compatible API.

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/carloslfu/slotstream>

WHAT PEOPLE ACTUALLY BUILT

Show HN: Review Markdown files like a document

[Show HN](#)

Markdown has become the preferred format for documentation and authoring skills across many teams. Yet, our tools to review markdown files haven't really kept up. Reviewing changes to a markdown file still requires you switch back and forth between the preview and source mode. Th

STEAL THIS

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

<https://github.com/shudv/easy-markdown-review-ado>

I built an open-source blind benchmark for AI agent harnesses - how would you make the comparison fair?

[r/LLMDevs](#)

[https://preview.redd.it/5cnaHgri3nh1.png?](https://preview.redd.it/5cnaHgri3nh1.png?width=1282&format=png&auto=webp&s=1a4af80d66754b8410f0af1337bf356d4e657e96)

[width=1282&format=png&auto=webp&s=1a4af80d66754b8410f0af1337bf356d4e657e96](https://preview.redd.it/5cnaHgri3nh1.png?width=1282&format=png&auto=webp&s=1a4af80d66754b8410f0af1337bf356d4e657e96) I've been trying to compare agent harnesses like Claude Code, Codex, Hermes, OpenClaw and others, and kept running into the same problem: most comparisons change t

STEAL THIS

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

https://www.reddit.com/r/LLMDevs/comments/1w56c7e/i_built_an_opensource_blind_benchmark_for_ai/

Model injection: making the model the payload for attacks

[r/LLMDevs](#)

This is my second post here! Cool if you check it out! Short version: i built a POC where a fine-tuned open model behaves as it should until it sees a specific trigger, then drops its guardrails and instructions. the keyword version is a toy, but the real point is that the trigge

STEAL THIS

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

https://www.reddit.com/r/LLMDevs/comments/1w5jaxb/model_injection_making_the_model_the_payload_for/

QUICK HITS

zai-org/GLM-5.3 HF MODEL

Trending model on Hugging Face — text-generation, transformers; 1551 likes, 151,021 downloads.

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

WebLLM: high-performance in-browser LLM inference engine HN

<https://github.com/mlc-ai/web-llm>

Show HN: Open source ML programming language playground SHOW HN

After joining a ML Study group and trying Google Colab I decided to create my own array programming language to help me learn and visualize

<https://sw-ml-study.github.io/sw-mlpl/>

Qwen/Qwen3.8-Flash-Next HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 4756 likes, 263,287 downloads.

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

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>

zai-org/GLM-5.3-Flash HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 1991 likes, 517,902 downloads.
<https://huggingface.co/zai-org/GLM-5.3-Flash>

Gemini 3.8 Flash and 3.8 Flash Cyber HN

<https://deepmind.google/models/model-cards/gemini-3-8-flash/>
<https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/>

rokbenko/quackd GITHUB

 Give your small robot a brain. Tell a Microduck, a Reachy Mini, a LeRobot arm or any ROS base what you want in plain language. An LLM (CI
<https://github.com/rokbenko/quackd>

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 13742 likes, 5,254,882 downloads.
<https://huggingface.co/Qwen/Qwen3.8-27B>

Show HN: Codeknow – Architecture health scores for any codebase, no LLM needed SHOW HN

<https://github.com/asalsali/codeknow>

deepseek-ai/DeepSeek-V4-Flash-Vision-Exp HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 517 likes, 54,571 downloads.
<https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-Vision-Exp>

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>

Show HN: Open-source transaction enrichment dataset SHOW HN

Many FinTech products display a nice merchant icon and human-readable name for transactions, instead of the plain text SQ *BRICKFIELDS CHIPP
<https://github.com/steveharrison/openenrichment>

Show HN: ToolJet – Claude Code and Codex build internal tools, no codegen SHOW HN

Hey HN, Co-founder here. We spent 11 months to build a multi-agent system for generating applications. Then we scrapped it and rebuilt aroun
<https://github.com/ToolJet/ToolJet>

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>
