

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

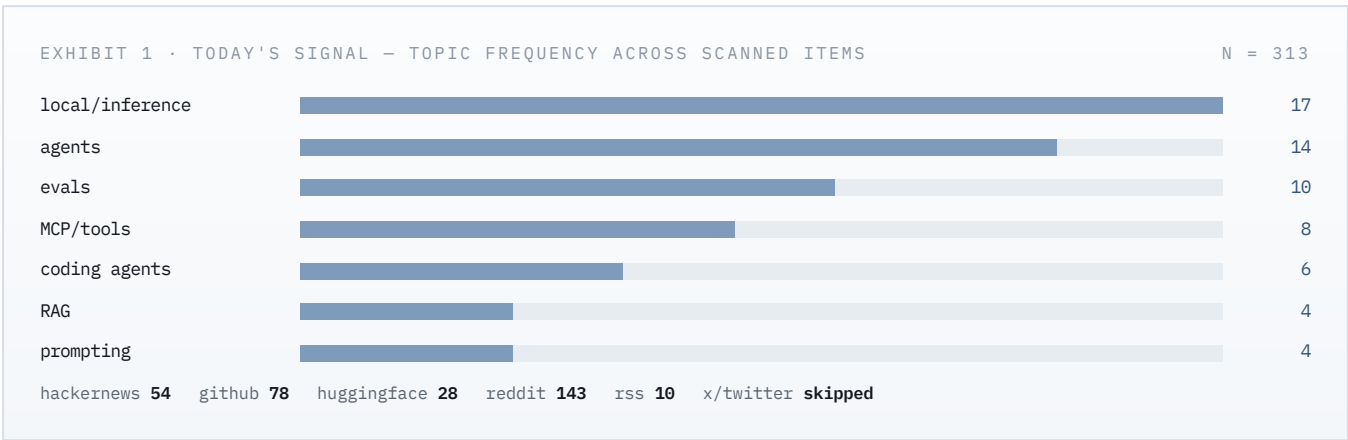
Today in AI engineering

2026-09-07

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 313 items scanned mode: **mechanical fallback**

local/inference agents evals mcp/tools coding agents rag prompting



DEEP DIVES

Show HN: Kadō – open-source habit tracker, with non-binary habit score, for iOS

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/scastiel/kado>

Show HN: Blunderbase: A Personal Chess Database

Show HN

I made a free and open source chess database. It's useful for personal game study and gathering statistics across sites. Supports import from lichess, chess.com, FICS and PGN then runs Stockfish and Maia over every game. With it you can analyze your games, take notes on positions and look up opening

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/philphilphil/blunderbase>

Show HN: Mador – Make any DOM reactive with a tiny 80-line Proxy state tuple

[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/marsbos/mador>

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>

Show HN: Dsnitch – Real-time, zero-config Docker egress inspector via eBPF

[Show HN](#)

Hi HN, I built dsnitch because I wanted a zero-config, bandwhich-style live TUI to see what my homelab Docker containers are connecting to—without running sidecar proxies, modifying container configs, or relying on reverse DNS (fails on anycast CDN IPs). Just run the binary and it automatically disc

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/infomaniac777/dsnitch>

WHAT PEOPLE ACTUALLY BUILT

Show HN: Slock – A one-bit messenger inside your capslock key

[Show HN](#)

Two computers, one bit. Hi — I built slock, a small macOS menu bar app that connects two people through their capslock lights. Hold capslock on your Mac, and the other person's capslock light turns on. Let go, and it turns off. That's all. A tiny way to interact someone without o

STEAL THIS

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

<https://github.com/jonaraphael/slock>

Show HN: I built a small arena where AI write robot and then fight each other

[Show HN](#)

(no summary)

STEAL THIS

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

<https://github.com/nigrosimone/llms-robot-arena>

Show HN: I stopped using an LLM gateway and put rate-limits/fallback in-process

[Show HN](#)

Sup HN. I built VernLLM cuz every LLM gateway I looked at, meant adding a network hop just to get rate limiting, multi provider fallback, circuit breaking and other features. Needing to take a whole seperate service just to deploy, monitor and trust with my API keys is just anno

STEAL THIS

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

<https://github.com/LakBud/vernLLM>

QUICK HITS

adtexerry-lgtm/unigit-ecosystem GITHUB

UNIGIT public brand and ecosystem hub — AI should work for everyone.

<https://github.com/adtexerry-lgtm/unigit-ecosystem>

BoardUI/boardui GITHUB

React design system for agentic interfaces. Every free BoardUI component as source, with a working AI chat app on your own model key as the

<https://github.com/BoardUI/boardui>

nahidalam/world_model_from_scratch GITHUB

Code repository for World Model From Scratch book

https://github.com/nahidalam/world_model_from_scratch

Show HN: Engrim – A universal, local-first SQLite memory engine for AI CLIs SHOW HN

<https://github.com/timgordontg/engrim>

vcruz305/vllm-exl3 GITHUB

Serve EXL3 (ExLlamaV3 trellis) quantized models on vLLM fork runtimes — any architecture, mixed per-layer bitrates, composable with source-f

<https://github.com/vcruz305/vllm-exl3>

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

Trending model on Hugging Face — image-text-to-text, transformers; 773 likes, 251,611 downloads.

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

XHToken/Spark-X2.5-4B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 645 likes, 7,216 downloads.

<https://huggingface.co/XHToken/Spark-X2.5-4B>

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 14177 likes, 6,416,358 downloads.

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

shtjww/llm-inference-capacity-handbook GITHUB

大模型推理案头手册（开源版）：给定 GPU 算力，一个模型能扛多少 QPS？三层模型 × 三面墙 × 排队论 × 开环压测

<https://github.com/shtjww/llm-inference-capacity-handbook>

google/timesfm-3.0-pytorch HF MODEL

Trending model on Hugging Face — time-series-forecasting; 534 likes, 271,713 downloads.

<https://huggingface.co/google/timesfm-3.0-pytorch>

aaravshah1311/Agent-2 GITHUB

Agent2 — An autonomous AI agent for coding, cybersecurity, terminal automation, and intelligent task execution.

<https://github.com/aaravshah1311/Agent-2>

Your intellectual fly is open when you use an LLM to author a post (2025) HN

<https://bcantrill.dtrace.org/2025/12/05/your-intellectual-fly-is-open/>

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

Trending model on Hugging Face — image-text-to-text, gguf; 486 likes, 403,292 downloads.
<https://huggingface.co/ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF>

Qwen/Qwen3.8-Flash-Next HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 4958 likes, 474,693 downloads.
<https://huggingface.co/Qwen/Qwen3.8-Flash-Next>

Lightricks/LTX-2.5 HF MODEL

Trending model on Hugging Face — image-to-video, diffusion-single-file; 3005 likes, 1,584,382 downloads.
<https://huggingface.co/Lightricks/LTX-2.5>
