

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

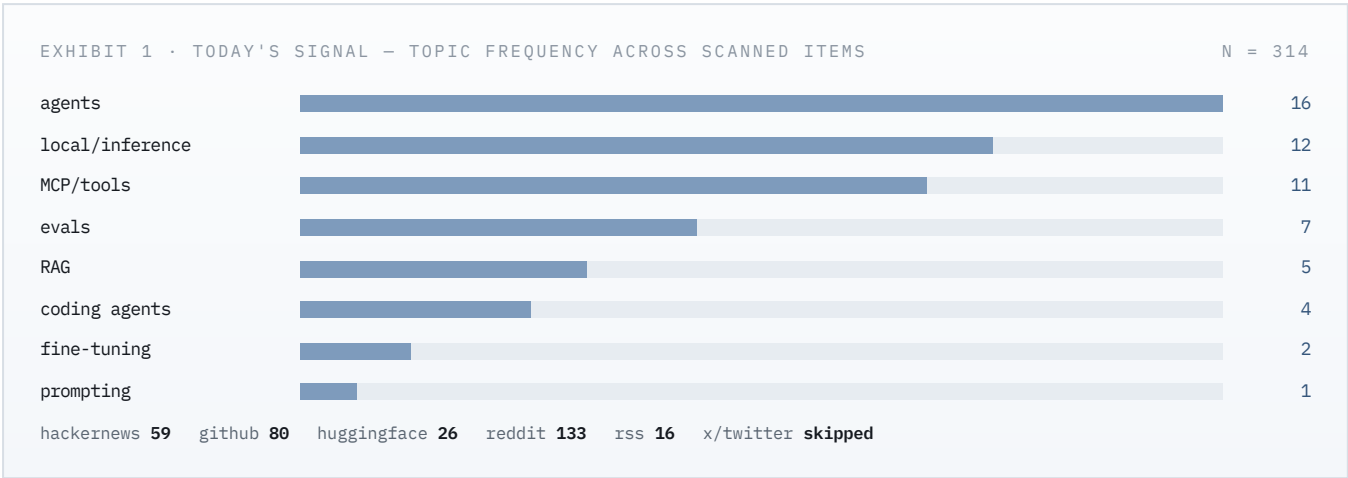
Today in AI engineering

2026-09-06

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 314 items scanned mode: **mechanical fallback**

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



DEEP DIVES

Show HN: Created a WhatsApp <-> discord bridge (no business API)

Show HN

built a self-hosted WhatsApp/Discord bridge that keeps conversations synced between both platforms. It supports messages, media, replies, edits, reactions, pins, and more. The Meta is not triggered for human paced convos and hence a normal number can be used, have been using it successfully for a de

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/Karvy-Singh/OSdc-wa>

2akouwu/reverify

[GitHub](#)

Stop your AI from making things up — it proposes, deterministic tools decide, every claim checked against ground truth with evidence. Grounded facts and context survive resets. Reverse engineering is the proving ground. MCP server + CLI.

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/2akouwu/reverify>

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: 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 opening a chat window

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/jonaraphael/slock>

Show HN: HyperCard to HTML Converter

[Show HN](#)

I created a converter that turns HyperCard 2.x stacks into standalone HTML so they can run in a modern browser. I built it to share my own stacks from the early '90s, which have been sitting on old floppy images for 30 years. It's MIT-licensed and I'd welcome contributions - especially stacks that b

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/stachon/hc2html>

WHAT PEOPLE ACTUALLY BUILT

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: Fast Cut Video tool for cutting video for Agents

[Show HN](#)

Hi HN, Build this tool in a couple of hours with OpenAI Codex to scratch an itch. As part of my videos workflows I needed to cut the videos myself since the agent is pretty bad cutting and timing using only the transcription. But using a big app like Premiere or DaVinci Resolve w

STEAL THIS

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

<https://github.com/modecir/fast-cutvid>

Show HN: Vibes – A Mac app for friends to see who's online and coding

[Show HN](#)

Hi HN, I built Vibes because I wanted a small sense of company while working on my own projects. It's a native Mac app that shows which friends are online, what they've chosen to put in their status, and a summary of their day's Git activity. You pick local repositories and conne

STEAL THIS

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

<https://vibes.opentangle.com/>

QUICK HITS

nahidalam/world_model_from_scratch [GITHUB](#)

Code repository for World Model From Scratch book

https://github.com/nahidalam/world_model_from_scratch

LLMs as a Cognitive Virus [HN](#)

<https://arxiv.org/abs/2609.03344>

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>

Show HN: Myjs An accidental pure-Python JavaScript interpreter [SHOW HN](#)

Hi HN! byteface here. myjs was a recent accident I'd like to share. Several years ago I began writing DOM API's in Python as a curiosity: ht

<https://pypi.org/project/myjs/>

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>

adtexerry-lgtm/unigit-ecosystem [GITHUB](#)

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

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

OKF Agent Memory – Git-native persistent memory for AI coding agents [HN](#)

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

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

Trending model on Hugging Face — image-text-to-text, transformers; 699 likes, 209,191 downloads.

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

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

Trending model on Hugging Face — image-text-to-text, transformers; 14063 likes, 6,190,807 downloads.

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

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

Trending model on Hugging Face — text-generation, transformers; 570 likes, 5,477 downloads.

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

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>

Qwen/Qwen3.8-Flash-Next HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 4926 likes, 432,966 downloads.

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

ARahim3/kaggle-tpu-lab GITHUB

Qwen3.8-27B (bf16) on a free Kaggle TPU: OpenAI-compatible endpoint, 262k context, ~130 tok/s, works with Claude Code, Codex, Opencode and P

<https://github.com/ARahim3/kaggle-tpu-lab>

google/timesfm-3.0-pytorch HF MODEL

Trending model on Hugging Face — time-series-forecasting; 473 likes, 144,455 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>
