

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

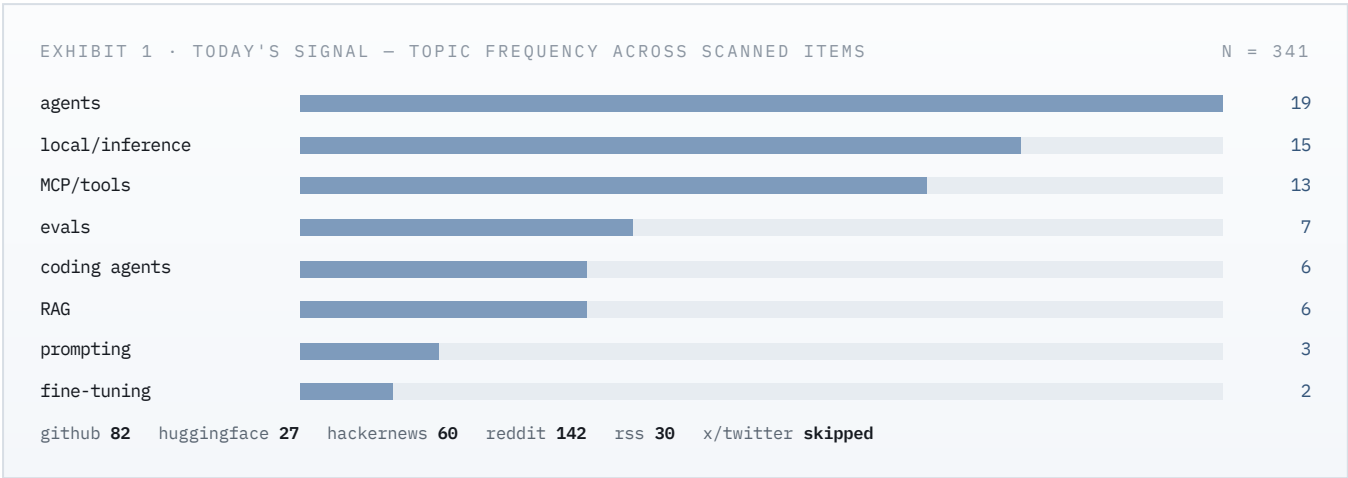
Today in AI engineering

2026-09-11

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 341 items scanned mode: **mechanical fallback**

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



DEEP DIVES

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>

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 stable OpenAI-compatible API, and automatic rollback.

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/hsj576/virtual-ai-infra-team>

deepseek-ai/DeepSeek-V4.1-Flash

[HF model](#)

Trending model on Hugging Face — image-text-to-text, transformers; 1592 likes, 6 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/deepseek-ai/DeepSeek-V4.1-Flash>

Show HN: Open-source simulation testing infra for voice agents

[Show HN](#)

Hey HN, we're Nischal & Naman. We're brothers, and together we're building an open-source platform for simulation based testing of voice agents (try it out in 5 mins - <https://docs.egma.ai/docs/get-started/quickstart> , 2 min demo video - <https://youtu.be/wgDWE5UAUY>) Platforms that help you do simu

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/egma-ai/egma>

WHAT PEOPLE ACTUALLY BUILT

Tell HN: OpenAI keeps re-enabling the 'allow training' setting

[HN](#)

I've reset this more than once and the last time I made a careful note of when I did it and to my surprise I found it re-enabled when I checked just now. Make sure you check this thing to see if it hasn't been re-enabled if you believe it to be off right now.

STEAL THIS

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

<https://news.ycombinator.com/item?id=49643556>

Show HN: I couldn't afford interview prep, so I built a free alternative

[Show HN](#)

(no summary)

STEAL THIS

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

<https://learningto.co/>

Show HN: Rocky Surf – Open-source control plane for coding-agent cloud VMs

[Show HN](#)

Rocky Surf started late last year after Claude Code crashed my laptop. The alpha version was designed as a commercial product, but I ditched the idea because it would have required significant capital. Basically, the business would have been a cloud reseller with a bit of software

STEAL THIS

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

<https://www.rockysurf.com/>

QUICK HITS

openbmb/MiniCPM5-2B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 1146 likes, 42,289 downloads.

<https://huggingface.co/openbmb/MiniCPM5-2B>

Show HN: MultiMatte, a Promptable Image Background Removal Model SHOW HN

Hey HN, I'm Shreyash from Feyn. We help companies build custom models from their data. Today we're releasing MultiMatte, a background removal tool.

Show HN: We served 5k dynamic websites from a 2-vCPU, 4 GB VPS SHOW HN

<https://github.com/Kooboo/Kooboo/blob/main/Docs/5000-sites-benchmark.md>

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

Trending model on Hugging Face — text-generation, transformers; 1077 likes, 15,930 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; 14687 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; 811 likes, 614,850 downloads.

<https://huggingface.co/ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF>

Show HN: Dbmask – Discover, mask, and verify sensitive data in SQL databases SHOW HN

Hi HN I developed *dbmask*, an open-source Python tool designed to discover sensitive columns in SQL databases, masking them with deterministic values.

<https://github.com/sealandseacat/dbmask>

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

Trending model on Hugging Face — text-generation, transformers; 668 likes, 2,444 downloads.

<https://huggingface.co/nex-agi/Nex-N2.5-mini>

More questions about whether researchers can trust OpenAI with unpublished math HN

<https://mathstodon.xyz/@andreasthom/117240536885387540> <https://mathstodon.xyz/@andreasthom/117240537520615623>

<https://x.com/ValerioCapraro/s>

<https://mathstodon.xyz/@andreasthom/117240535270608201>

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>

Atomburstofficial/geiger GITHUB

A Geiger counter for AI agents — one read-only command that inventories every agent, MCP server, plugin, and AI extension on a machine
<https://github.com/Atomburstofficial/geiger>

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

Trending model on Hugging Face — text-generation, transformers; 590 likes, 3 downloads.
<https://huggingface.co/nex-agi/Nex-N2.5-Pro>

Show HN: Charter – Operate production-safe agents that run on your own infra SHOW HN

Even though everyone is talking about AI agents doing everything for them at work, many companies still aren't using autonomous agents to au
<https://github.com/boundflow/charter>

Show HN: Ridge - Connect coding agents to local, SSH, Docker, and S3 resources SHOW HN

I am working on an open source project called Ridge that gives coding agents a common interface to resources such as local projects, Docker
<https://github.com/vasinov/ridge-core>

Show HN: A local, open-source version of Claude Code SHOW HN

<https://github.com/just-not-google/BiNeuron>
