

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

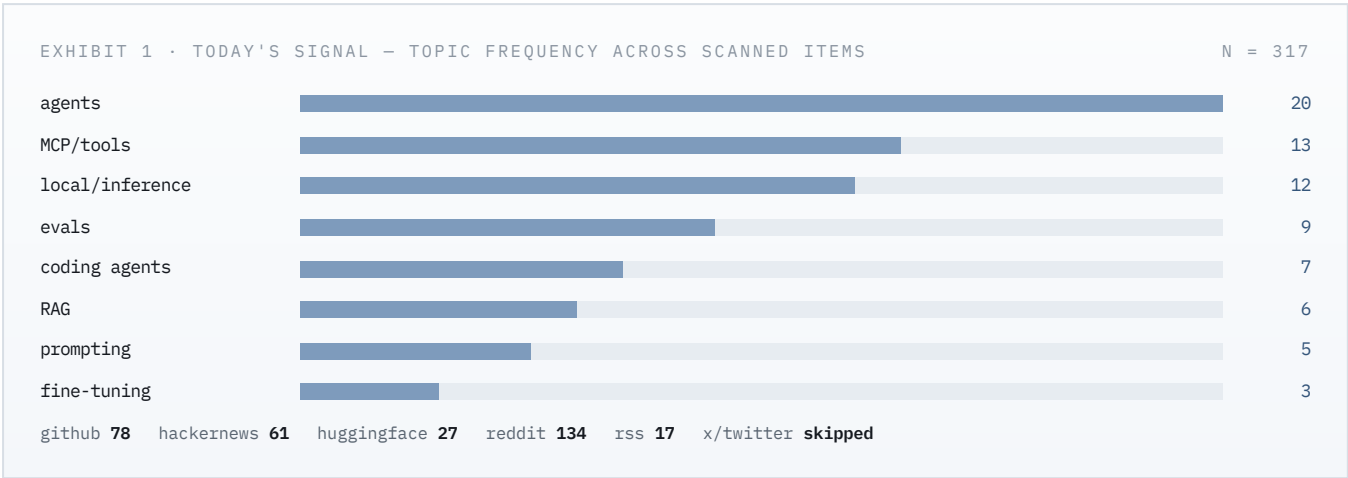
Today in AI engineering

2026-09-13

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 317 items scanned mode: **mechanical fallback**

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



DEEP DIVES

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>

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>

Show HN: Open-sourced solution to AI manage Playwright automations

[Show HN](#)

We built a comprehensive system that connects directly to the AI agents through MCP for writing, managing most type of browser automations. Purpose is for automated testing or automating browser based workflows. It provides notifications, scheduling, pre/post hooks, IAM management, cookies support,

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/rakutentech/sorify>

deepseek-ai/DeepSeek-V4.1-Flash

[HF model](#)

Trending model on Hugging Face — image-text-to-text, transformers; 2082 likes, 244,457 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>

openbmb/MiniCPM5-2B

[HF model](#)

Trending model on Hugging Face — text-generation, transformers; 1289 likes, 150,110 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/openbmb/MiniCPM5-2B>

WHAT PEOPLE ACTUALLY BUILT

Show HN: EarthToPixels – Turning real cities into explorable pixel art

[Show HN](#)

Hi HN! A few friends and I built EarthToPixels, a project that turns real-world cities into explorable pixel-art maps. We fine-tuned an open-source Qwen image model and built a pipeline that takes aerial imagery, generates pixel-art tiles, and keeps neighboring generations visual

STEAL THIS

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

<https://www.earthtopixels.com>

I built a benchmark to test whether AI agents break architecture rules. It found none. (null result, with the harness and data)

r/AI_Agents

The common claim: your CLAUDE.md / AGENTS.md is guidance, a lint rule is enforcement and AI coding agents drift from prose, so you need the deterministic rule. I wanted to measure that gap. I mined a real import-boundary rule from a near-census of 24,888 public TypeScript repos (

STEAL THIS

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

https://www.reddit.com/r/AI_Agents/comments/1wen0hh/i_built_a_benchmark_to_test_whether_ai_agents/

Show HN: I made an alternative to the archived filebrowser

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/skidoodle/filebrowser>

QUICK HITS

Real-SWE: Benchmarking AI models on private, real-world, enterprise codebases HN

<https://withspecific.com/benchmarks/real-swe>

panda-sandeep/dearagent GITHUB

Email inbox API for AI agents, self-hosted on Cloudflare Workers. Open source, with an MCP server. AgentMail alternative.

<https://github.com/panda-sandeep/dearagent>

I refuse to let SPICE die HN

https://github.com/nefarius/vd_agent/

Show HN: Open-Sourced Chrome Extension to Save, Organize and Switch Bookmarks SHOW HN

<https://github.com/ringlochid/leotabs>

Edge0/Edge0-35B-A3B-preview HF MODEL

Trending model on Hugging Face — text-generation, mlx; 644 likes, 3,552 downloads.

<https://huggingface.co/Edge0/Edge0-35B-A3B-preview>

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

Trending model on Hugging Face — text-generation, transformers; 748 likes, 3,970 downloads.

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

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 14898 likes, 7,768,964 downloads.

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

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

Trending model on Hugging Face — text-generation, transformers; 1147 likes, 21,336 downloads.

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

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

Trending model on Hugging Face — text-generation, transformers; 619 likes, 30,289 downloads.

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

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

Trending model on Hugging Face — image-text-to-text, gguf; 926 likes, 769,557 downloads.
<https://huggingface.co/ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF>

Show HN: A C Compiler written with help of Claude SHOW HN

For fun, I wrote a c compiler (non-optimizing) in Java targeting modern C (C23) using claude. I developed a c shell similar to jshell in jav
<https://github.com/Javier-Barrio/myCC/tree/main>

cadamcat/llms-on-apple-neural-engine GITHUB

Apple Neural Engine (ANE/NPU) vs GPU for local LLM inference on Apple Silicon / macOS. Core ML (CoreML), Core AI (CoreAI), MLX and Metal ben
<https://github.com/cadamcat/llms-on-apple-neural-engine>

krmisystems/fantasy-football-manager GITHUB

ESPN draft and lineup tools for Codex and other MCP clients, with browser sign-in, approval modes, per-team limits, and server monitoring.
<https://github.com/krmisystems/fantasy-football-manager>

forefy/reburp GITHUB

A Burp Suite extension that exposes the full Montoya API as a local REST API, with Swagger UI
<https://github.com/forefy/reburp>

Show HN: Forma ONNX Graph Viewer and Editor SHOW HN

Hey all! Most ONNX tooling stops at just "looking". For those of you who know about Netron, it allows you to read the graph but the moment y
<https://github.com/Hussain004/forma>
