

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

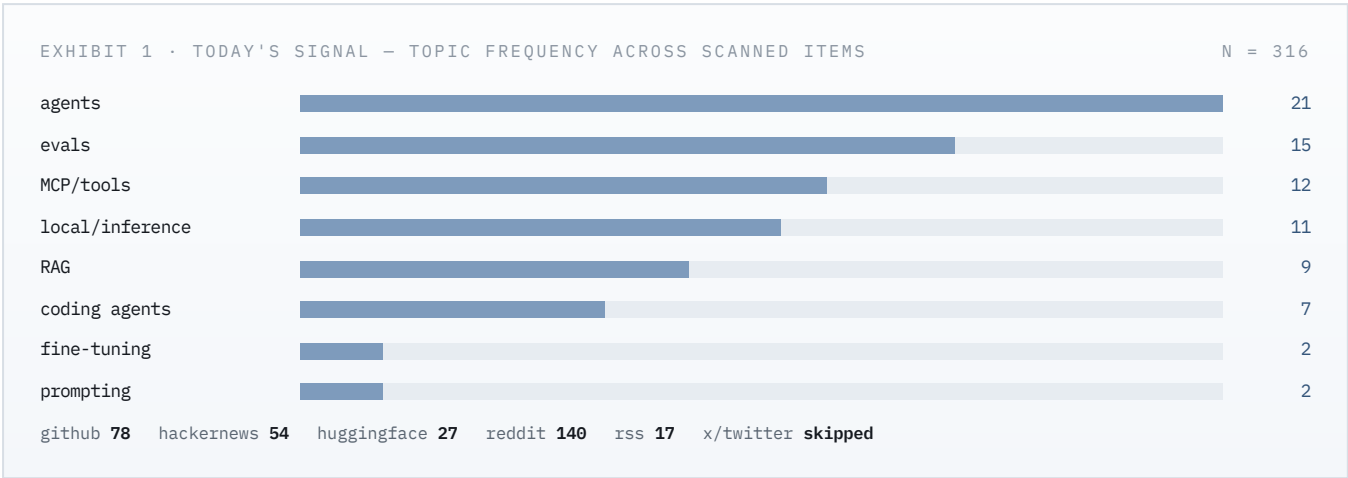
Today in AI engineering

2026-09-14

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 316 items scanned mode: **mechanical fallback**

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



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>

Show HN: Plainoldanalytics: Analytics Middleware for Go

[Show HN](#)

I built a middleware that gives you posthog-style analytics in go. It is all first-party. It logs every HTTP request, lets you track users, and lets you record screen sessions. The thing uses DuckDB as the underlying data store. All of my side projects are now Golang + SQLite. I wanted an easy way to

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/poundifdef/plainoldanalytics>

Show HN: I Built a CLI for Proton Mail, Drive, Calendar, Pass and Contacts

[Show HN](#)

I made a command line tool for Proton. It covers Mail, Drive, Calendar, Pass and Contacts. It is one static binary for Linux, macOS and Windows. Proton doesn't have a CLI for all their web interfaces. There is one for Drive and one for Pass, but the other ones are not really there. There's no Mail C

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/roman-16/proton-cli>

deepseek-ai/DeepSeek-V4.1-Flash

[HF model](#)

Trending model on Hugging Face — image-text-to-text, transformers; 2324 likes, 288,414 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: Is It Greg?

[Show HN](#)

My friend Greg often posts incredible things he just made, just for fun. Often, I don't realize that it's something HE made and then I see the URL, AND IT'S MY FRIEND, GREG! A few things he made that I think are amazing: - ikea complexity index: <https://ikea.greg.technology/> - Boing: <https://boing.g>

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/antoineleclair/is-it-greg>

WHAT PEOPLE ACTUALLY BUILT

Show HN: I built my own knowledge graph from the code AI writes

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/aeroscissorz/learnlance/>

Show HN: My coding agent apologized for being right, so I built Frank

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/HimanshuJ16/frank>

I didn't want to give AI agents my whole private repo, so I built GHOSTBOUND

r/LLMDevs

I use AI quite a lot when working on private coding projects, but not always in the same role. A workflow I like is using ChatGPT as a planner/reviewer first. It gets the relevant context, looks at the architecture and helps work out what actually needs to change. Then I use a se

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/1wf5ag2/i_didnt_want_to_give_ai_agents_my_whole_private/

QUICK HITS

Show HN: Deeplearning from Scratch in 1400 Lines In my own Programming Language

SHOW HN

https://github.com/NoSavedDATA/Neve_benchmarks

Edge0/Edge0-35B-A3B-preview

HF MODEL

Trending model on Hugging Face — text-generation, mlx; 1437 likes, 8,109 downloads.

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

openbmb/MiniCPM5-2B

HF MODEL

Trending model on Hugging Face — text-generation, transformers; 1367 likes, 206,774 downloads.

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

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>

nex-agi/Nex-N2.5-mini

HF MODEL

Trending model on Hugging Face — text-generation, transformers; 774 likes, 4,543 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; 15067 likes, 7,703,400 downloads.

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

nex-agi/Nex-N2.5-Pro

HF MODEL

Trending model on Hugging Face — text-generation, transformers; 630 likes, 30,489 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; 989 likes, 819,784 downloads.
<https://huggingface.co/ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF>

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

Trending model on Hugging Face — text-generation, transformers; 1163 likes, 24,084 downloads.
<https://huggingface.co/XHToken/Spark-X2.5-4B>

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>

Astra and Fable still hack on simple variants of alignment evals from 2025 HN

<https://www.lesswrong.com/posts/munJKF7iWMsWJLAH2/astra-and-fable-still-hack-on-simple-variants-of-alignment>

ivyfan-toowell/IvyClaw GITHUB

A production-oriented multi-agent AI Agent system for software engineering. | 面向软件研发任务的多智能体 AI Agent 工程系统
<https://github.com/ivyfan-toowell/IvyClaw>

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>

Show HN: Makefaster.dev SHOW HN

I had a bunch of extra Fable credits, so I spent around \$10k in api costs doing autoresearch loops on the top 200 github repos with frontend
<https://makefaster.dev>
