

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

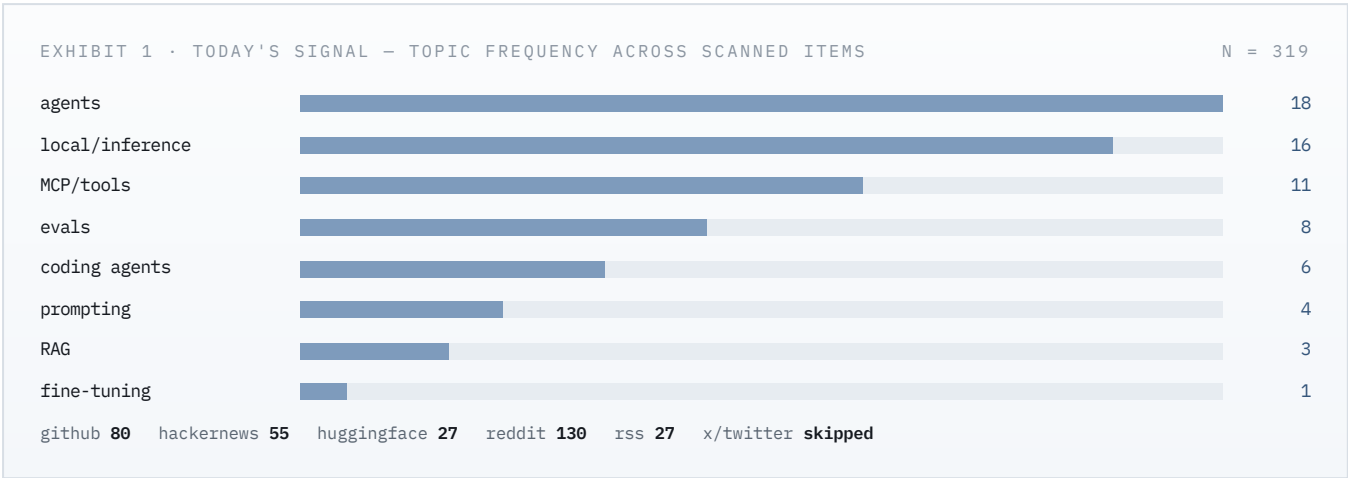
Today in AI engineering

2026-09-12

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 319 items scanned mode: **mechanical fallback**

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



DEEP DIVES

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: Godot and Rust based multiplexer (terminal panes and more)

[HN](#)

I wanted to share my side-project: gPTY. Started off as an idea to combine Godot and Rust in a project (two stacks I wanted to use more to learn more). The base inspiration was tmux - simply allow spawning multiple PTYs and then let the user grid/tile them how they see fit. But since we have the God

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/godot-pty/gpty>

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: Toast, a beautiful by default in terminal IDE

[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/paradise-runner/toast>

deepseek-ai/DeepSeek-V4.1-Flash

[HF model](#)

Trending model on Hugging Face — image-text-to-text, transformers; 1870 likes, 75,774 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>

WHAT PEOPLE ACTUALLY BUILT

Show HN: Minimal Static Website Hosting Tool

[Show HN](#)

I built this because I self host a lot of stuff and wanted to simplify tossing up a simple static site with no backend, no NGINX/Caddy/Traffik configs, instantly deploy with a single file copy. Powered by Golang's built-in reverse proxy package. Another awesome freebie is this is

STEAL THIS

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

<https://github.com/VictorioBerra/mistatic>

Show HN: Claude Read Aloud – Hear Claude’s Replies Instead of Reading

[Show HN](#)

When I'm working all day sometimes my eyes get tired and I don't read Claude's replies fully and understand them. So I made Claude Read Aloud to read Claude's replies for me. It's very helpful for when I'm running on low energy and I think it can also be helpful for people who ar

STEAL THIS

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

<https://github.com/MichaelPGifford/claude-read-aloud>

I built a multilingual RAG document assistant with FastAPI, React, FAISS and Ollama

[r/LLMDevs](#)

I've been learning more about RAG systems recently, so I decided to build one instead of just following tutorials. The result is CloudRAG : <https://github.com/abderaoufsec/CloudRAG> The basic flow is: 📄 Documents → ✂️ Chunking → 🧠 Embeddings → 🔍 FAISS → 📖 Retrieved chunks → 🗣️ Olla

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/1wdw3ig/i_built_a_multilingual_rag_document_assistant/

QUICK HITS

RTK reports token savings, but our cost benchmarks disagree [HN](#)

<https://quesma.com/blog/does-rtk-make-ai-coding-cheaper/>

Show HN: ResolveHQ – A Helpdesk Built on Cloudflare Workers, D1, R2 and Queues [SHOW](#) [HN](#)

<https://github.com/mirza-rizvi/ResolveHQ>

Litelm: LiteLLM Without the Bloat [HN](#)

<https://github.com/kennethwolters/litelm>

Show HN: Graphify C# – Compiler-accurate Find Usages for coding agents [SHOW](#) [HN](#)

<https://github.com/zachsaw/graphify-csharp>

openbmb/MiniCPM5-2B [HF](#) [MODEL](#)

Trending model on Hugging Face — text-generation, transformers; 1218 likes, 67,550 downloads.

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

Show HN: Keychron macOS Battery Life and Controls [SHOW](#) [HN](#)

I was struggling with getting this working as there's no official support for reporting battery life via bluetooth and via cable. It took a

<https://github.com/bkovacev/keychron-macos>

Show HN: Hazzel – tiny, open-source, Git-native coding agent [SHOW](#) [HN](#)

<https://github.com/mukundzha/hazzel>

A misalignment of AI in mathematics [HN](#)

<https://terrytao.wordpress.com/2026/09/11/a-severe-misalignm...> <https://www.economist.com/science-and-technology/2026/09/11/...> ,

<https://un>

<https://mathandai.org/>

XHToken/Spark-X2.5-4B [HF](#) [MODEL](#)

Trending model on Hugging Face — text-generation, transformers; 1124 likes, 17,712 downloads.

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

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

Trending model on Hugging Face — text-generation, transformers; 704 likes, 3,121 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; 14797 likes, 7,563,763 downloads.

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

Show HN: ClaudeStatsBar: your session is 486k deep and nothing told you SHOW HN

<https://github.com/Field-Logic-Ltd/ClaudeStatsBar>

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>

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>

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

Trending model on Hugging Face — image-text-to-text, gguf; 851 likes, 682,187 downloads.

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