

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

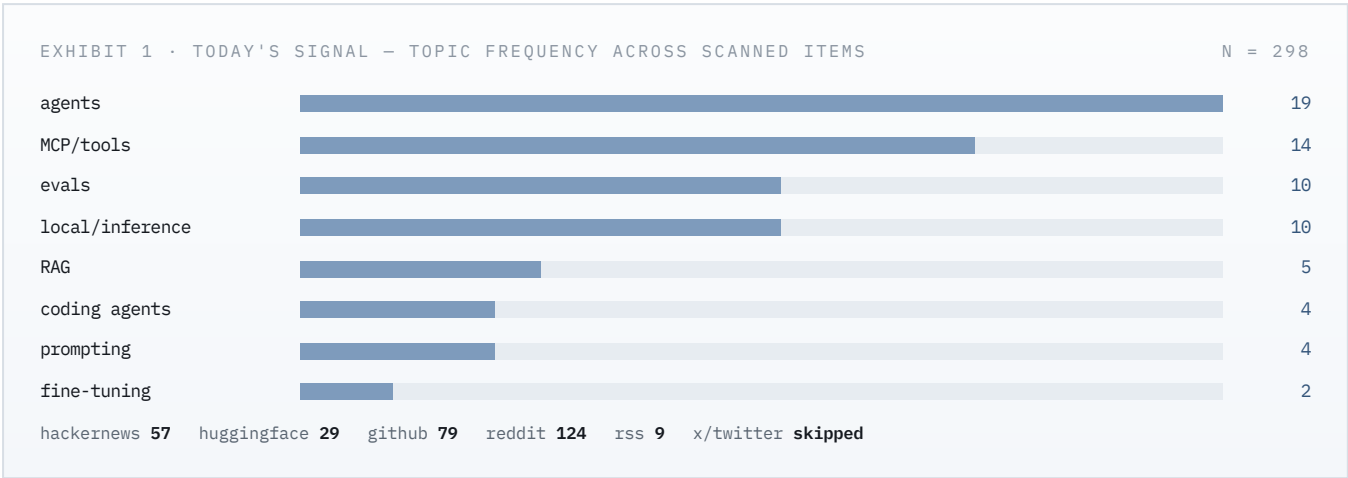
Today in AI engineering

2026-08-30

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 298 items scanned mode: **mechanical fallback**

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



DEEP DIVES

Show HN: Delete yourself from data brokers without a subscription

Show HN

I signed up for a paid service to get my phone, addresses, and vehicles off the public internet. It didn't work. So I built remove-your-data, an open-source repo that tells your agent exactly what to do to remove your personal information online.

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/k7cfo/remove-your-data>

Qwen/Qwen3.8-Flash-Next

HF model

Trending model on Hugging Face — image-text-to-text, transformers; 4326 likes, 121,976 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/Qwen/Qwen3.8-Flash-Next>

UditAkhouri/cdaf

[GitHub](#)

CDAF (Cached Descriptive Asset Files) - open sidecar format for video so AI agents stop re-analyzing the same footage. Spec, CLI, agent skill, reproducible benchmark.

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/UditAkhouri/cdaf>

Show HN: DeepSeekGUI – A Windows desktop client for DeepSeek’s coding agent

[Show HN](#)

Hi HN, I built a desktop client for DeepSeek Harness (DeepSeek's open-source coding agent). V1 wraps the official Harness Web UI in an Electron shell with some desktop additions — installer, system tray, built-in browser panel (visible Edge, so you can watch the agent browse), and a sandboxed termin

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/See-So1-Lab/DeepSeekGUI>

zorost/sparkduet

[GitHub](#)

Serve DeepSeek and Qwen, run an on-demand model library, and fine-tune with Unsloth QLoRA on 2x NVIDIA DGX Spark. TP=2 vLLM lanes behind one OpenAI-compatible endpoint for OpenCode, Cursor, and Hermes. Honest benchmarks, committed artifacts.

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/zorost/sparkduet>

WHAT PEOPLE ACTUALLY BUILT

Show HN: PhpEZ – A tiny PHP framework for shared LAMP hosting

[Show HN](#)

I needed to make a tiny webapp backend to upload to shared LAMP hosting (they still exist, and I think they're fun despite us being well beyond 2005), but I couldn't face waiting for FTP to sync down however thousands files Laravel or similar frameworks need. So I made a tiny PHP

STEAL THIS

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

<https://github.com/QcFe/phpEZ>

I analyzed 31,352 hourly LLM benchmark scores: within-day variation was 2.8 points, while between-day variation was 8.4

r/LLMDevs

AI Stupid Level dashboard Disclosure: I built AIStupidLevel. The frontend and backend are open source under the MIT license. One of the clearest findings from continuously benchmarking LLMs was how differently their performance behaves within a single day compared with longer per

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/1w1jm6h/i_analyzed_31352_hourly_llm_benchmark_scores/

Open-source tool to detect unauthorized document retrieval in RAG apps

r/LLMDevs

Hey Guys, I built a small open-source tool that checks whether a RAG application retrieves documents a user shouldn't have access to. It supports offline test cases and live HTTP API testing with bearer token/API-key auth. I'm looking for a few engineers to try it on a test or no

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/1w1zj3p/opensource_tool_to_detect_unauthorized_document/

QUICK HITS

zai-org/GLM-5.3-Flash HF MODEL

Trending model on Hugging Face — text-generation, transformers; 1651 likes, 346,516 downloads.
<https://huggingface.co/zai-org/GLM-5.3-Flash>

zai-org/GLM-5.3 HF MODEL

Trending model on Hugging Face — text-generation, transformers; 1304 likes, 50,116 downloads.
<https://huggingface.co/zai-org/GLM-5.3>

vLLM v0.28.0 HN

<https://github.com/vllm-project/vllm/releases/tag/v0.28.0>

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 13290 likes, 4,511,348 downloads.
<https://huggingface.co/Qwen/Qwen3.8-27B>

JordyZomer/lemmalog GITHUB

A Datalog engine for LLM agent memory: stratified rules, provenance-tracked facts, incremental derivation, and an MCP server that lets your
<https://github.com/JordyZomer/lemmalog>

Show HN: Tokensift, an open-sourced token-efficiency linter for LLM prompts SHOW HN

<https://github.com/ritenv/tokensift>

unsloth/Qwen3.8-Flash-Next-GGUF HF MODEL

Trending model on Hugging Face — image-text-to-text; 582 likes, 328,195 downloads.
<https://huggingface.co/unsloth/Qwen3.8-Flash-Next-GGUF>

unsloth/Qwen3.8-27B-GGUF HF MODEL

Trending model on Hugging Face — model; 3201 likes, 8,839,153 downloads.
<https://huggingface.co/unsloth/Qwen3.8-27B-GGUF>

Stupidoodle/swissdevjobs-cli GITHUB

Search & apply to ~4,700 salary-transparent tech jobs across 7 countries (🇨🇭🇩🇪🇬🇧🇺🇸🇨🇦🇳🇱🇫🇷) from your terminal or AI agent — zero-dependency P
<https://github.com/Stupidoodle/swissdevjobs-cli>

Gen-Verse/Recuris GITHUB

Recursive Experiential–Working Memory Evolution for Long-Horizon Agent Harnesses
<https://github.com/Gen-Verse/Recuris>

omaekumiko2-create/kru GITHUB

Local-first MCP password and credential manager for AI agents. Use passwords, API keys, SSH identities, and TOTP without exposing hidden pla
<https://github.com/omaekumiko2-create/kru>

Show HN: Flint – A minimal C/C++ package manager and build tool written in C SHOW HN

<https://github.com/mainak55512/flint>

Show HN: 3D Embodiment of your Git repo SHOW HN

<https://github.com/michaelaz774/RepoWorld>

Lightricks/LTX-2.5 HF MODEL

Trending model on Hugging Face— image-to-video, diffusion-single-file; 2177 likes, 1,137,181 downloads.
<https://huggingface.co/Lightricks/LTX-2.5>

tencent/Hy4-preview HF MODEL

Trending model on Hugging Face— text-generation, transformers; 295 likes, 2,123 downloads.
<https://huggingface.co/tencent/Hy4-preview>
