

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

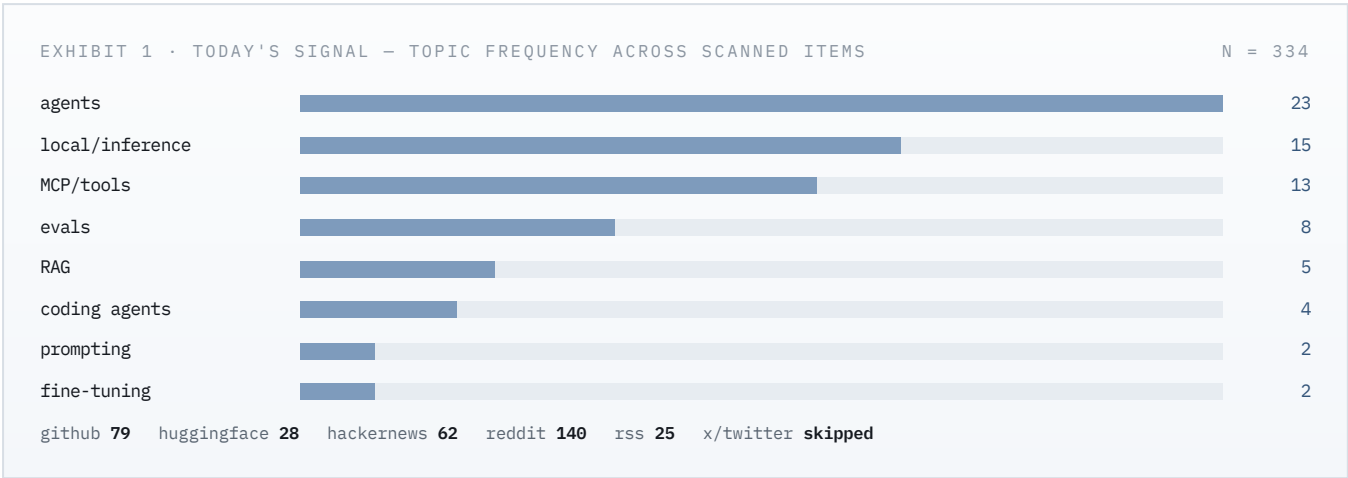
Today in AI engineering

2026-09-18

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 334 items scanned mode: **mechanical fallback**

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



DEEP DIVES

youngyangyang04/llm-master

GitHub

大模型（LLM）全栈学习路线与中文教程🔥：覆盖 Prompt Engineering、RAG、AI Agent、MCP、微调、模型部署、Transformer、AI 编程与大厂面试，从入门到生产实践。

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/youngyangyang04/llm-master>

shinthink/blitzstrike

[GitHub](#)

✂ Blitz Strike — a universal MCP penetration-testing toolbelt. Structured methodology: reconnaissance & attack-surface mapping, source-to-sink analysis, and live validation. 57 escalation chains, 130-tool catalog, intelligence data layer. One server, every agent.

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/shinthink/blitzstrike>

Edge0/Edge0-35B-A3B-preview

[HF model](#)

Trending model on Hugging Face — text-generation, mlx; 3353 likes, 37,131 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/Edge0/Edge0-35B-A3B-preview>

Show HN: AutoBot – live voice control for long-running AI work

[Show HN](#)

I wanted to manage long horizon agentic workstreams via voice, then put my phone down, and have a harness manage completion - extending into full computer use. I was trying to build a personal Jarvis, so I benchmarked AutoBot to see how close I could get: - OSWorld: 32.41% (moved Sol Max from 4th 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/demeyer1/Autobot>

AetherLabsAI/RSIAgent

[GitHub](#)

A training-free multi-agent framework for recursive self-improvement in new environments through broad-then-deep autonomous exploration and reusable memory.

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/AetherLabsAI/RSIAgent>

WHAT PEOPLE ACTUALLY BUILT

Show HN: I built a new version of my fun spatial 3D online meeting app

[Show HN](#)

Hi HN! flat.social is a fun spatial online meeting app for remote teams and communities. It's largely built like a web multiplayer game, and the 3D virtual spaces are highly customisable with elements and activities (I recently added surfing!). Live demo to try it out: <https://ap>

STEAL THIS

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

<https://flat.social>

Show HN: Share your AI Setup, Learn from others

[HN](#)

I kept seeing engineers share what they were building with AI; however, I was always more curious about how they worked. Which agents did they use? What skills and tools had stuck or been thrown out the window? How did they manage longer-running tasks? So I built this with the ho

STEAL THIS

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

<https://mysetup.ai/>

Show HN: Linting 216 public Claude Code skills – 69% won't reliably trigger

[Show HN](#)

I built a linter for Claude Code artifacts (skills, subagents, slash commands, .mcp.json configs, and plugins) and then pointed it at a pile of public repos to see what people actually ship. A few of the numbers surprised me enough that I wrote them up. The one that stuck: across

STEAL THIS

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

<https://skillcrossroads.com>

QUICK HITS

deepseek-ai/DeepSeek-V4.1-Flash HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 3073 likes, 390,657 downloads.

<https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash>

NativePHP/superstack GITHUB

Open-source Laravel starter with Filament, NativePHP SuperNative, and MCP

<https://github.com/NativePHP/superstack>

A heap overflow and SSO misconfiguration to compromise OpenAI internal repos HN

<https://www.hacktron.ai/blog/hacking-openai>

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 15583 likes, 7,456,257 downloads.

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

Tokha233/deepseek-v4.1-flash-a100-turbo GITHUB

DeepSeek-V4.1-Flash inference on NVIDIA Ampere with vLLM, SM80 patches, reproducible throughput and coding-agent benchmarks. Tested on 8x A8

<https://github.com/Tokha233/deepseek-v4.1-flash-a100-turbo>

elsechord/CyberGuard GITHUB

Evidence-driven autonomous SOC team on AgentTeams: hash-bound approvals, HMAC audit chain, rollback and reproducible attack scenarios — audi

<https://github.com/elsechord/CyberGuard>

m-a-p/YuE2-3B HF MODEL

Trending model on Hugging Face — text-to-audio; 755 likes, 11,626 downloads.

<https://huggingface.co/m-a-p/YuE2-3B>

Show HN: Aclif – Agent CLI framework: one grammar, canonical names across SaaS SHOW HN

Simple discovery of the provider's default schemas is possible without any credentials (since they are built into the binary). If you want i

<https://www.aclif.ai/>

prism-ml/Ternary-Bonsai-2-27B-gguf HF MODEL

Trending model on Hugging Face — text-generation, llama.cpp; 531 likes, 0 downloads.

<https://huggingface.co/prism-ml/Ternary-Bonsai-2-27B-gguf>

ruc-datalab/EvoOntology GITHUB

EvoOntology: A Self-Evolving Ontology Layer for Data Agents  EvoOntology插件为Claude Code/Codex 建立&进化本体层

<https://github.com/ruc-datalab/EvoOntology>

Infinite-Parameter LLMs: Generating and Adapting Weights from Live Data HN

<https://arxiv.org/abs/2609.18842>

TokenRhythm/NeoHorse-1-4B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 2355 likes, 19,789 downloads.

<https://huggingface.co/TokenRhythm/NeoHorse-1-4B>

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

Trending model on Hugging Face — image-text-to-text, gguf; 1282 likes, 1,027,602 downloads.

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

Show HN: Concat: the open-source CapCut replacement. SHOW HN

I started this project about ~3 weeks ago. Since CapCut is putting everything behind a paywall after gaining millions of installs and users,

<https://github.com/jub0t/Concat>

Lightricks/LTX-2.5 HF MODEL

Trending model on Hugging Face — image-to-video, diffusion-single-file; 4267 likes, 1,602,865 downloads.

<https://huggingface.co/Lightricks/LTX-2.5>
