

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

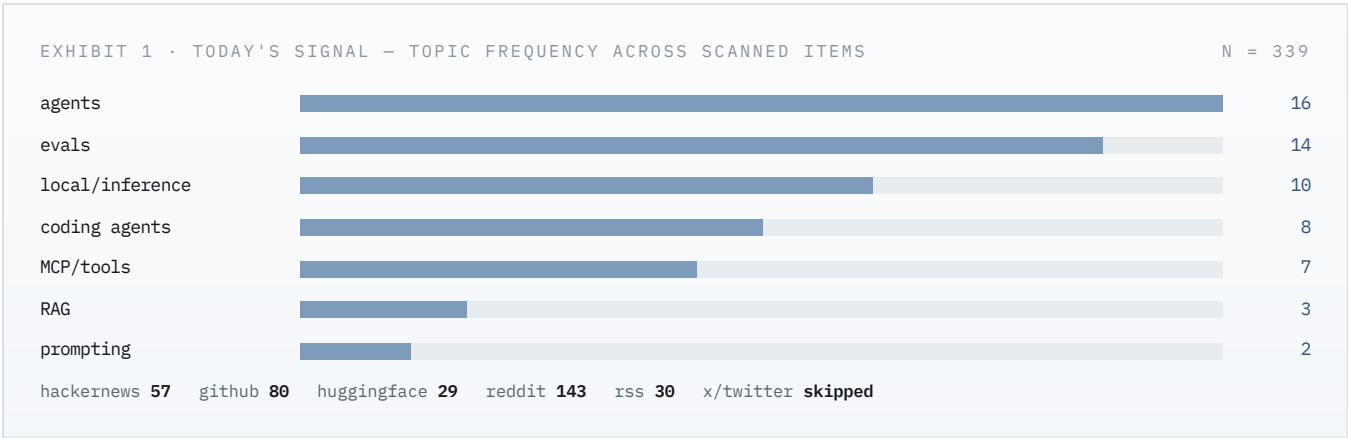
Today in AI engineering

2026-09-23

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 339 items scanned mode: **mechanical fallback**

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



DEEP DIVES

incoai/splash

GitHub

A local inference engine for Apple silicon, built around the model.

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/incoai/splash>

pallavi-shekhar/ai-engineering-interview-questions-company-wise

GitHub

Your Cheat Sheet For AI Engineering Interviews at Top AI Companies - Questions and Answers.

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/pallavi-shekhar/ai-engineering-interview-questions-company-wise>

convaiinnovations/laya

HF model

Trending model on Hugging Face — text-classification, transformers; 2809 likes, 0 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/convaiinnovations/laya>

anandi1989/awesome-jev-usecases

GitHub

Evidence-backed index of real-world Jev (TypeSafe AI System One) use cases: repos, patterns, benchmarks, and measured results

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/anandi1989/awesome-jev-usecases>

Qwen/Qwen-Image-2.1

HF model

Trending model on Hugging Face — text-to-image, diffusers; 1896 likes, 28,407 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/Qwen-Image-2.1>

WHAT PEOPLE ACTUALLY BUILT

Show HN: JevBench, a reproducible benchmark for typed decision models

HN

Hi HN! I built JevBench because Jev kicks ass, and the world deserves to know how the serious open source and fake lookalike projects really perform in comparison. Jev-class models return bounded choices and probabilities instead of text, and are disruptively faster and cheaper t

STEAL THIS

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

<https://benchmarkheaven.com/jev-models>

Show HN: A fast, offline Markdown viewer for macOS built in Swift

Show HN

Hi HN! I made Tim's Markdown Reader because I wanted a free, open-source app just for reading Markdown. I spend a lot of time working with AI coding agents, and sometimes I just want to open the files they produce and read them without opening an editor. It's written in Swift and

STEAL THIS

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

<https://github.com/temir-dev/tims-markdown-reader>

Show HN: What's Next – end a Claude Code turn with a choice, not prose

[Show HN](#)

Hi HN, I built `What's Next` (whats-next@kvit) – a Claude Code plugin that ends a turn with a short report + a card of 2-4 options instead of wall of Claude's prose. Think of it as reverse Jev: Jev's thesis is when the answer is a choice, return a typed choice instead of generati

STEAL THIS

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

<https://github.com/kvitapp/kvit-plugins>

QUICK HITS

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

Trending model on Hugging Face — text-generation, llama.cpp; 1909 likes, 2,815,979 downloads.

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

XingChen-AGI/Xing4.0-29B-A4B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 1444 likes, 39,009 downloads.

<https://huggingface.co/XingChen-AGI/Xing4.0-29B-A4B>

Heman10x-NGU/openJev-verdict-2.0 GITHUB

Calibrated 151M Non-Autoregressive Decision Engine beating TypeSafe Jev & Laya on LocalLLaMA/typed-decisions (77.10% acc, 0.0636 Brier, 0.01

<https://github.com/Heman10x-NGU/openJev-verdict-2.0>

Jev Decision Index HF SPACE

Trending Space (static, 145 likes). Benchmarks and news on various repros of TypeSafe's Jev

<https://huggingface.co/spaces/multimodalart/jev-decision-index>

abenzers/Qwen-Image-2.1-Uncensored-GGUF HF MODEL

Trending model on Hugging Face — text-to-image, gguf; 1269 likes, 350,678 downloads.

<https://huggingface.co/abenzers/Qwen-Image-2.1-Uncensored-GGUF>

zeeshan8281/slo-router GITHUB

SLO-aware LLM inference router with Jev decisions, live queue metrics, counterfactual evaluation, and reproducible latency/cost benchmarks

<https://github.com/zeeshan8281/slo-router>

sno-ai/sno-station GITHUB

Sno Station — your Claude Code and Codex working as one squad on your own machine. Shared encrypted memory, agent-to-agent messaging (Reach)

<https://github.com/sno-ai/sno-station>

Claude Opus 5.5 HN

<https://www.anthropic.com/claude-opus-5-5>

GPT-6 Sol and Luna HN

<https://openai.com/index/introducing-gpt-6-sol-and-luna/>

Show HN: graf (1000x faster graphify in Rust) SHOW HN

I liked the idea of graphify, but for large repos the search queries were visibly very slow. I dug into it and saw an opportunity to make it

<https://github.com/ctxrs/graf>

Show HN: InstinctFlash – High-Performance Serving Runtime for Robotics Models SHOW HN

Hey HN, Guanming here, cofounder of General Instinct. We just released InstinctFlash, a high-performance serving framework for robotics mode

<https://github.com/General-Instinct/InstinctFlash>

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>

Show HN: Npunlock – Run custom C kernels for Intel NPUs SHOW HN

<https://github.com/hsfzxjy/npunlock>

INSANE0777/Awwards-mcp GITHUB

Free, open-source MCP server that gives AI coding agents (Claude Code, Codex, Cursor, OpenCode, ZCode...) design inspiration from award-winnin

<https://github.com/INSANE0777/Awwards-mcp>

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

Trending model on Hugging Face— image-text-to-text, transformers; 3636 likes, 570,909 downloads.

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