

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

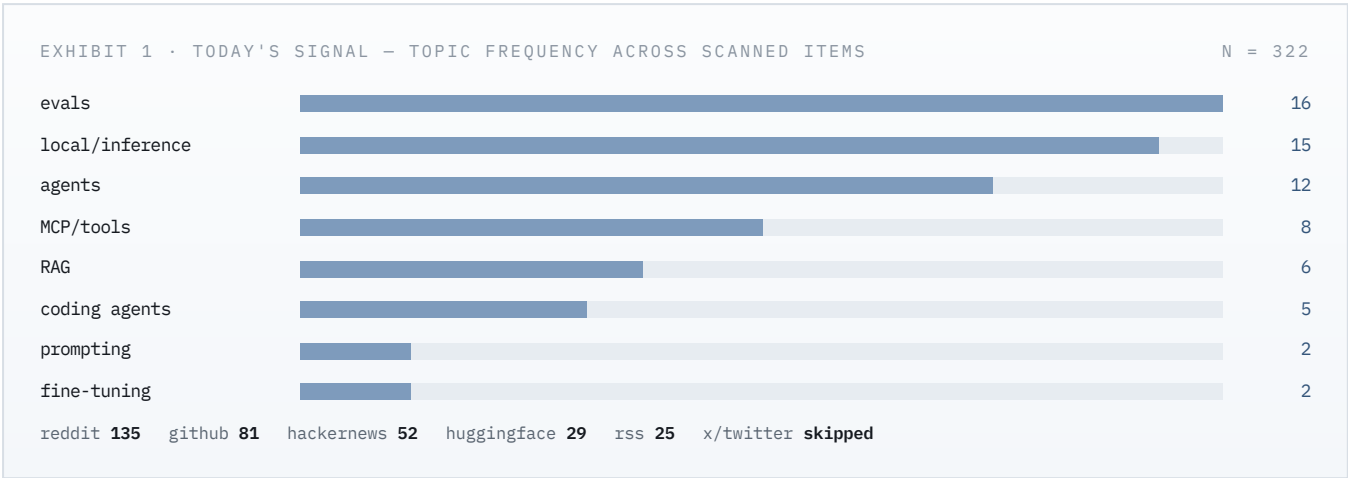
Today in AI engineering

2026-09-20

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 322 items scanned mode: **mechanical fallback**

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



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>

Show HN: MNIST64 – 99.5% accuracy MNIST BNN on a stock Commodore 64

[Show HN](#)

I built an 8-bit binarized convolutional neural network that recognizes handwritten digits on C64 with state-of-the-art accuracy. I also designed a cascade model to get interactive response, and implemented everything with self-modifying cycle optimized 6502 assembly - resulting 0.653s average infer

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/jarnoh/mnist64>

Show HN: CUA-S1 – A System One Model for Computer Use

[HN](#)

Hello HN! We're Dillon and Francesco from Cua. We were wondering how many computer use tasks actually need a full general purpose LLM (e.g. gpt-6-astra, claude-opus-5 etc.) to think through all their decisions and steps. Some tasks require thinking about a plan, exploring different paths, recovering

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/trycua/cua>

Show HN: I created an open source locally usable full fledged AI platform

[Show HN](#)

hi to all the readers this post is for my recent opensource project called ENZO now answering what is enzo so enzo is an opensource platform where i clubbed all the free available api for anyone use under one hood with more than 2000 models available to use for chatting coding researching and much m

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/theguysudo/ENZO>

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>

WHAT PEOPLE ACTUALLY BUILT

I built a paged LLM inference engine for Apple Silicon and cut attention memory by 128x

r/LLMDevs

I've been building PagedServe, an experimental open-source LLM inference runtime for Apple Silicon. The problem I wanted to explore was KV-cache fragmentation. Most simple inference implementations allocate contiguous memory or reconstruct scattered KV blocks before attention. At a

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/1w1bx42/i_built_a_paged_llm_inference_engine_for_apple/

Show HN: Seal – Letters and passwords that open for your family after you die

Show HN

I'm a CS undergrad who has developed an interest in hardware keys ever since I used them for a school project a few years ago. I made a little chat app that only allows users to chat if they registered a key, nothing else. The backstory of how I came to work on Seal (this project

STEAL THIS

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

<https://github.com/jasonepage/Seal>

Show HN: Reader – I made a Mac workspace for books, browser tabs, notes, and AI

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/Marvy101/reader-source>

QUICK HITS

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

Trending model on Hugging Face — text-generation, llama.cpp; 1326 likes, 1,908,396 downloads.

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

anandi1989/awesome-jev-usecases GITHUB

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

<https://github.com/anandi1989/awesome-jev-usecases>

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

Trending model on Hugging Face — image-text-to-text, transformers; 3360 likes, 496,684 downloads.

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

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>

convaiinnovations/laya HF MODEL

Trending model on Hugging Face — text-classification, transformers; 766 likes, 0 downloads.

<https://huggingface.co/convaiinnovations/laya>

Btrfs/ZFS/bcachefs under workloads classic benchmarks skip HN

<https://bartosz.fenski.pl/modern-fs-benchmark/>

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>

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

Trending model on Hugging Face — text-generation, transformers; 762 likes, 12,617 downloads.

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

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 15808 likes, 7,331,932 downloads.

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

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>

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>

agentscope-ai-java/agentscope-java GITHUB

Build distributed, production-grade, long-running agents.

<https://github.com/agentscope-ai-java/agentscope-java>

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

Trending model on Hugging Face — text-to-audio; 891 likes, 17,403 downloads.

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

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>

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

Trending model on Hugging Face — image-text-to-text, gguf; 1445 likes, 1,217,204 downloads.

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