DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

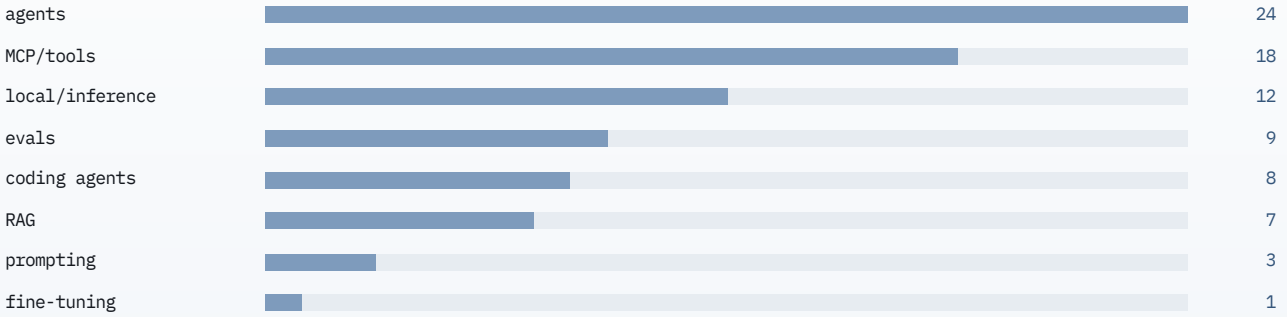
Today in AI engineering
2026-09-28

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 **309** items scanned mode: **mechanical fallback**

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

EXHIBIT 1 · TODAY'S SIGNAL — TOPIC FREQUENCY ACROSS SCANNED ITEMS N = 309



huggingface **30** hackernews **57** github **75** reddit **135** rss **12** x/twitter **skipped**

DEEP DIVES

convaiinnovations/laya

HF model

Trending model on Hugging Face — text-classification, transformers; 4205 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

Show HN: Gat – Version large files with Git, without an LFS server

Show HN

Hi HN, I built Gat because I wanted one thing from data versioning: check out an old Git commit and recover the datasets, model weights, or assets that belong to it. It's basically a lock file for large files. I'd appreciate any feedback!

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/getgat-dev/gat

Jev Decision Index

HF Space

Trending Space (static, 287 likes). Benchmarks and news on various repros of TypeSafe's Jev

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/spaces/multimodalart/jev-decision-index

breakstageaxe61/genspark-claw

GitHub

🦀⚡ Community skill pack & CLI for Genspark Claw, OpenClaw & Hermes agents — 5 production-ready SKILL.md skills (deep research, reports, slides, browser automation, AI call prep), zero-dependency Node.js installer, MCP server, one-line install

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/breakstageaxe61/genspark-claw

abenzerps/Qwen-Image-2.1-Uncensored-GGUF

HF model

Trending model on Hugging Face — text-to-image, gguf; 2173 likes, 1,062,921 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/abenzerps/Qwen-Image-2.1-Uncensored-GGUF

WHAT PEOPLE ACTUALLY BUILT

Show HN: EthosLM - Turn a sentence into a Minecraft city

Show HN

I built EthosLM to turn a description of a place into something you can walk around in Minecraft. The demo starts with “Build me Ba Sing Se.” It produces a city with districts, roads, walls and furnished interiors. Here's the one-minute video: https://www.youtube.com/watch?v=lgix

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

https://github.com/chaitbuilds/EthosLM

Show HN: swe-mux – A terminal multiplexer for coding agents optimized for mobile

Show HN

I run a lot of coding agents at once, mostly Claude Code, Codex and pi, across several projects. I got tired of having so many windows open, of agents leaving hung processes behind, of tracking which one needed input, and of having to remote into my desktop or SSH into a single s

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

https://github.com/jatoran/swe-mux

Qwen3.8-Flash-Next 177B NVFP4(119GiB): SSD streaming at 9-10 tok/s on one 16 GB RTX 5060 Ti + 32 GB RAM

r/LocalLLaMA

We built an inference engine InferredThoughts for MoE models that don't fit in VRAM + RAM. Most of the model stays on the SSD, and experts are read as tokens need them. This started as a proof of concept, and poc worked, we are getting 9-10 tok/s decode on Qwen3.8-Flash-Next NVFP

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

https://www.reddit.com/r/LocalLLaMA/comments/1wrxap8/qwen38flashnext_177b_nvfp4119gib_ssd_streaming_at/

QUICK HITS

Show HN: TinyAIArena watch AI agents battle it out HN

Did you ever click on an “AI Arena” expecting glorious battle and instead get a boring benchmark? If so, this project is for you: proper lif
https://tinyaiarena.com/

Show HN: Aldine. Self-hosted Overleaf alternative. Every paper = 1 Git Repo SHOW HN
https://github.com/trahloff/Aldine

Qwen/Qwen-Image-2.1 HF MODEL

Trending model on Hugging Face — text-to-image, diffusers; 2542 likes, 58,693 downloads.
https://huggingface.co/Qwen/Qwen-Image-2.1

Edge0/Audio8-ASR-Infinite HF MODEL

Trending model on Hugging Face — automatic-speech-recognition, transformers; 1320 likes, 19,963 downloads.
https://huggingface.co/Edge0/Audio8-ASR-Infinite

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

Trending model on Hugging Face — text-generation, transformers; 1788 likes, 45,834 downloads.
https://huggingface.co/XingChen-AGI/Xing4.0-29B-A4B

ThinkFlowLab/system1-agents GITHUB

System 1 decision models (Jev, Laya, Cua-S1) as brain for agents: Browser use, computer use, games and robotics
https://github.com/ThinkFlowLab/system1-agents

XiaomiMiMo/MiMo-V2.6-Pro-RL HF MODEL

Trending model on Hugging Face — text-generation, transformers; 569 likes, 76,518 downloads.
https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Pro-RL

graygnatconsole/mcp-audit-tool GITHUB

🛡️ Security audit CLI for Model Context Protocol (MCP) servers — scan AI agent configs for tool poisoning, rug pulls, hardcoded secrets, com
https://github.com/graygnatconsole/mcp-audit-tool

XingChen-AGI/TeleOCR HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 645 likes, 27,904 downloads.
https://huggingface.co/XingChen-AGI/TeleOCR

Show HN: Lofi Cities – Pixel-art city nights with browser-generated lofi SHOW HN
https://loficities.com/

architectds/collabosm GITHUB

A local app that rents a GPU in your own Colab account and serves Qwen3.8 on it: 27B on an A100-40G, or Flash-Next (125B-A6B MoE) on an A100
https://github.com/architectds/collabosm

Unsealed Briefs in Authors' Case v. Microsoft/OpenAI HN

https://authorsguild.org/news/ag-v-openai-top-execs-knew-mass-book-piracy-was-illegal/

"As a Language Model": Chat Template Switches LLM Self-Referential Voice HN

https://arxiv.org/abs/2609.25021

Show HN: Asimov's 4 Laws for AI Contexts SHOW HN

For community and agent comments, thoughts, refinements, critiques, pull requests, and use.

https://github.com/davidsonff/asimov4laws

Show HN: AstraBox – an open-source alternative to Claude Managed Agents SHOW HN

https://github.com/Colton-z/AstraBox

Generated 2026-09-28 · Sources: Hacker News, Reddit, GitHub, Hugging Face, arXiv, engineering RSS. Filtered for real, implementable work; marketing and cash-grabs removed. Built by the ai-daily-brief automation · LUCA CONARROE — AI Automation.