DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

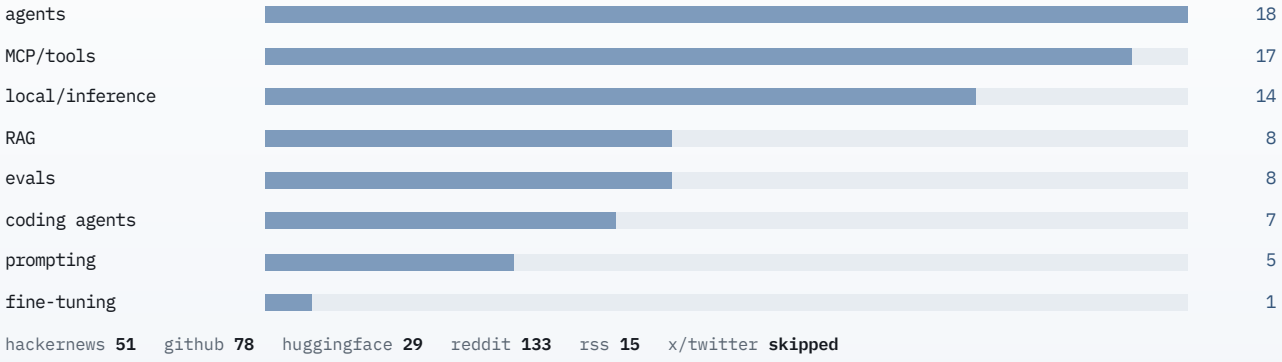
Today in AI engineering 2026-11-02

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 **306** items scanned mode: **mechanical fallback**

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

EXHIBIT 1 · TODAY'S SIGNAL — TOPIC FREQUENCY ACROSS SCANNED ITEMS N = 306



hackernews **51** github **78** huggingface **29** reddit **133** rss **15** x/twitter **skipped**

DEEP DIVES

Show HN: OmniDiff – Opensource, super fast, content aware Git difftool, in Rust

Show HN

I have been researching what the ultimate diff tool would look like, based on actual diffs in open source repositories. OmniDiff can diff: code (25 languages), config, images, PDFs, archives, cursors, fonts... It's fast. It can diff C++ files of tens of thousands of lines in under 100ms, on a 2023 T

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/ivankovic/omnidiff

glincker/awesome-selfhosted-ai

GitHub

Curated list of self-hosted open-source AI: local LLMs, chat UIs, agents, MCP servers, RAG, speech and image generation. Health-checked weekly, dead projects removed automatically.

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/glincker/awesome-selfhosted-ai

Show HN: CheckOwners to find your repo's hidden single points of failure

Show HN

Hey HN Community! I created checkOwners to solve a problem by many engineering teams: who knows your codebase, and what will happen when they go away? This is an open-source, local-first tool which performs analysis based on Git history in order to find hidden knowledge risk in repositories. It is a

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/smusali/checkOwners

muaaz9329/cheap-ag

GitHub

50x your Claude usage with cheap-ag. Opus plans, a dirt-cheap model writes every line, and tests decide what ships. Your $20 plan stops hitting limits. One MCP server for Claude Code.

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/muaaz9329/cheap-ag

abenzerps/Qwen-Image-2.1-Uncensored-Quantized

HF model

Trending model on Hugging Face — text-to-image, gguf; 5517 likes, 2,516,622 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-Quantized

WHAT PEOPLE ACTUALLY BUILT

Show HN: Sweet-search – local code search for agents (5–33% cheaper runs)

Show HN

Hi everyone :) After almost a year of fighting windmills, I am proud to say I built sweet-search, a local code search tool for Claude Code, Codex and Opencode. It delivers 5-33% cheaper agentic runs with Sol 6.1 and Opus 5.5 with Claude Code, Codex and Opencode, with pooled solve

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

https://github.com/mrsladoje/sweet-search

Show HN: HackerTube for Apple TV. Videos HN Recommends, on Your TV

Show HN

Andrej Karpathy built HackerTube last week. It’s a site of YouTube videos recommended in HN comments. I wanted it on my TV, and I wanted it to stay current on its own. So I built an Apple TV app and a small backend that keeps the data fresh. The app is open source. You build it i

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

https://news.ycombinator.com/item?id=50335401

Show HN: People you know, an end-to-end encrypted private Instagram replacement

Show HN

I built an alternative to popular social networks because I was sick of the unending enshittification of instagram. Every UX change in recent years intended to make their app more addictive to teenagers has made me want to use it less, but I don't want to lose contact with the ac

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

https://peopleyouk.now/

QUICK HITS

Show HN: HackerBooks – HackerTube, but for books mentioned on HN HN

Saw Karpathy's HackerTube and wanted the same thing for books. So I had Claude build one (Opus 5.5). It's ~16.5k books that at least two dif
https://hackerbooks.org

deepseek-ai/DeepSeek-V4.1-Flash-1017 HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 4604 likes, 177,009 downloads.
https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017

LiquidAI/LFM3-1.7B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 779 likes, 42,384 downloads.
https://huggingface.co/LiquidAI/LFM3-1.7B

Show HN: Jik – a self-hosted language using region-based memory management SHOW HN
https://github.com/jik-lang/jik

nvidia/NVIDIA-Nemotron-3.5-Transcribe-30B-A4B HF MODEL

Trending model on Hugging Face — automatic-speech-recognition, nemo; 643 likes, 32,439 downloads.
https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Transcribe-30B-A4B

Show HN: Wikihack – HackerTube, but for Wikipedia articles mentioned on HN SHOW HN

After seeing Karpathy's Hackertube and Hackerbooks, I just had to make Wikihack. I scraped all mentions of wikipedia on hackernews and ranke
https://wikihack.netlify.app

peterpoal134/memory-brain GITHUB

给 AstrBot 的可落地 AI 记忆与生活层：五类记忆、混合检索、遗忘治理、情绪边界、自主生活循环。不覆盖人格。

https://github.com/peterpoal134/memory-brain

mistralai/Mistral-Large-4-1T-A52B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 1475 likes, 5,500 downloads.

https://huggingface.co/mistralai/Mistral-Large-4-1T-A52B

baidu/Unlimited-OCR-2 HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 550 likes, 147,588 downloads.

https://huggingface.co/baidu/Unlimited-OCR-2

thefahien/elpis-on-7900xtx GITHUB

Qwen3.8-27B EXL3 + DFlash2 speculative decoding on an AMD RX 7900 XTX (exllamav3-rocm): recipe, patches, benchmarks

https://github.com/thefahien/elpis-on-7900xtx

Lightricks/LTX-2.5 HF MODEL

Trending model on Hugging Face — image-to-video, diffusion-single-file; 10295 likes, 2,467,613 downloads.

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

ccore/Qwen3.5-4B-3D-Coder HF MODEL

Trending model on Hugging Face — model; 526 likes, 63,499 downloads.

https://huggingface.co/ccore/Qwen3.5-4B-3D-Coder

Show HN: I've updated my Linux System Call website and API SHOW HN

I first released those in back in 2020 and now I finally added more architectures, Kernel selector, syscall descriptions and a few other nic

https://syscall.sh/

PeaceOfElf37336/resilient-telegram-rpocket-bot GITHUB

A Telegram bot built around what happens when the API doesn't answer — retry with jittered backoff, structured logging, systemd deployment.

https://github.com/PeaceOfElf37336/resilient-telegram-rpocket-bot

rmr2024/finextract GITHUB

Benchmark and fix LLM hallucinations when extracting figures from SEC 10-K filings

https://github.com/rmr2024/finextract