DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

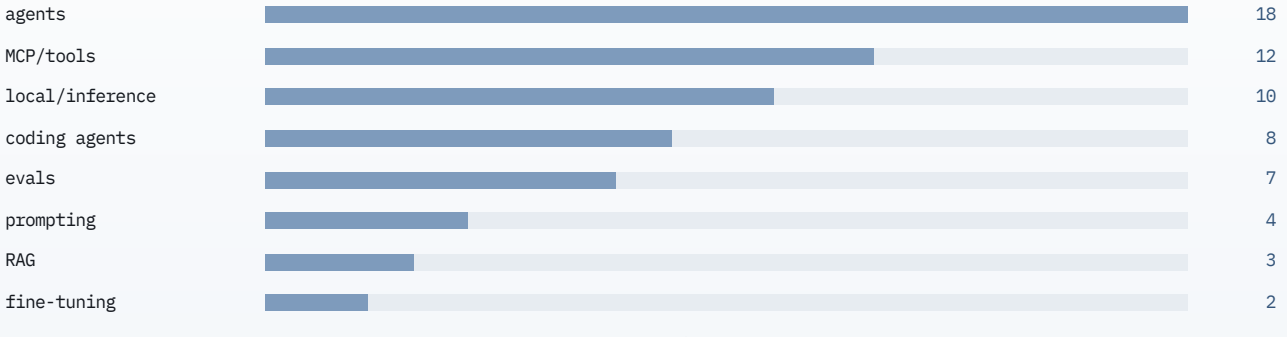
Today in AI engineering 2026-10-31

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 **323** items scanned mode: **mechanical fallback**

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

EXHIBIT 1 · TODAY'S SIGNAL — TOPIC FREQUENCY ACROSS SCANNED ITEMS N = 323



hackernews **58** github **82** huggingface **28** reddit **124** rss **31** x/twitter **skipped**

DEEP DIVES

Show HN: A working Nintendo in ~1GB of Qwen weights

Show HN

(no summary)

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/dbreunig/nesformer

qybaihe/bla

GitHub

🛡️ BLA (Blind LLM Agents) — Adversarial defense shield that protects binaries, Electron apps, JS bundles, Python, and WebAssembly against AI reverse-engineering agents like REA. Poison -> Exhaust -> Hallucinate.

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/qybaihe/bla

Show HN: Automatic deduction of FlashAttention kernel from ~25 lines of SQL

Show HN

In this demo, we automatically deduce a FlashAttention kernel with ~25 lines of SQL, which just expresses vanilla attention. This is possible due to three portable SQL extensions that we're building: - xarray-sql: a library that efficiently represents petabyte scale array datasets in the cloud as ta

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/xqlsystems/xarray-sql/blob/6debe2495dcb4e8fac20ab97df5b89baae2a31a6/benchmarks/ml/01_attention.py

codician-team/diffinery

GitHub

Open-source, agent-first GitHub and GitHub Actions alternative that runs on your own Cloudflare account: git hosting, pull requests, code review and drop-in GitHub Actions YAML on Workers, Durable Objects, R2 and Containers.

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/codician-team/diffinery

emanuele-em/kanzo

GitHub

Self-hosted kanban for humans and agents. One 4 MB binary, starts in 2 ms, zero setup.

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/emanuele-em/kanzo

WHAT PEOPLE ACTUALLY BUILT

Show HN: I built a multiplayer game where you're an aerial fighter cat

Show HN

Hi everyone! I had the idea for catcopter on my mind for a long time (got the domain ~13 years ago). It's a web multiplayer game where you can fly a dangerous cat and engage in competition with other players and critters (pterodactyls, chickens and angry penguins). I made it init

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

https://catcopter.com

I made a Claude Code skill that gives essays a research, revision, and fact-checking workflow

r/LLMDevs

I made JP Essay, an open-source skill for Claude Code that takes a question through research, outlining, drafting, revision, and fact-checking to produce one final essay. One part of the workflow is a fresh AI reader: a separate agent reads the draft and reconstructs its argument

> **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/1xjqp3b/i_made_a_claude_code_skill_that_gives_essays_a/

I built a free tool that finds cost spikes in cloud billing CSVs (code does the math, Gemini only explains). Looking for feedback

r/LLMDevs

Hey all, I made a small project called AI Cloud Cost Detective and I'd love honest feedback. What it does: you upload a billing CSV (date, service, category, region, cost, currency) and it shows daily cost trends, breakdowns by service and region, anomalies grouped into incidents

> **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/1xk4ql9/i_built_a_free_tool_that_finds_cost_spikes_in/

QUICK HITS

Show HN: PicoPlayer – A tiny, open-source website for playing quality webgames SHOW HN

Hi HN! I just released a new minimalist website that lets you play over 150 PICO 8 games on the web that I curated myself. It's fully open s

https://picoplayer.net/

larchanka/skills GITHUB

This repository is a collection of independent skills and tools for software development and automation. Each skill is a self-contained proj

https://github.com/larchanka/skills

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 week

https://github.com/glincker/awesome-selfhosted-ai

abenzerps/Qwen-Image-2.1-Uncensored-Quantized HF MODEL

Trending model on Hugging Face — text-to-image, gguf; 5390 likes, 2,558,497 downloads.

https://huggingface.co/abenzerps/Qwen-Image-2.1-Uncensored-Quantized

Nanbeige/Nanbeige4.3-33B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 818 likes, 13,832 downloads.

https://huggingface.co/Nanbeige/Nanbeige4.3-33B

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

Trending model on Hugging Face — image-text-to-text, transformers; 4052 likes, 156,373 downloads.

https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017

Show HN: Spinf-serve – inference server for bulk scoring and decisions SHOW HN

https://github.com/spinfinc/spinf-serve

Backdoor in Claude Code for VS Code extension on the Cursor Marketplace HN

https://cursor.com/blog/claude-code-extension-security-notic...
https://phorion.io/blog/offical-claude-code-vsx-compromise/

LiquidAI/LFM3-1.7B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 594 likes, 32,933 downloads.
https://huggingface.co/LiquidAI/LFM3-1.7B

boyuchengz/canvas-mcp GITHUB

Read-only MCP server that connects Claude, ChatGPT and other AI assistants to Canvas LMS.
https://github.com/boyuchengz/canvas-mcp

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

Trending model on Hugging Face — automatic-speech-recognition, nemo; 516 likes, 24,357 downloads.
https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Transcribe-30B-A4B

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

baidu/Unlimited-OCR-2 HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 466 likes, 69,231 downloads.
https://huggingface.co/baidu/Unlimited-OCR-2

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

Trending model on Hugging Face — image-text-to-text, transformers; 1367 likes, 2,945 downloads.
https://huggingface.co/mistralai/Mistral-Large-4-1T-A52B

Backdoors in official Claude Code extensions HN

https://cyberplace.social/@GossiTheDog/115728831681803563

Generated 2026-10-31 · 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.