DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

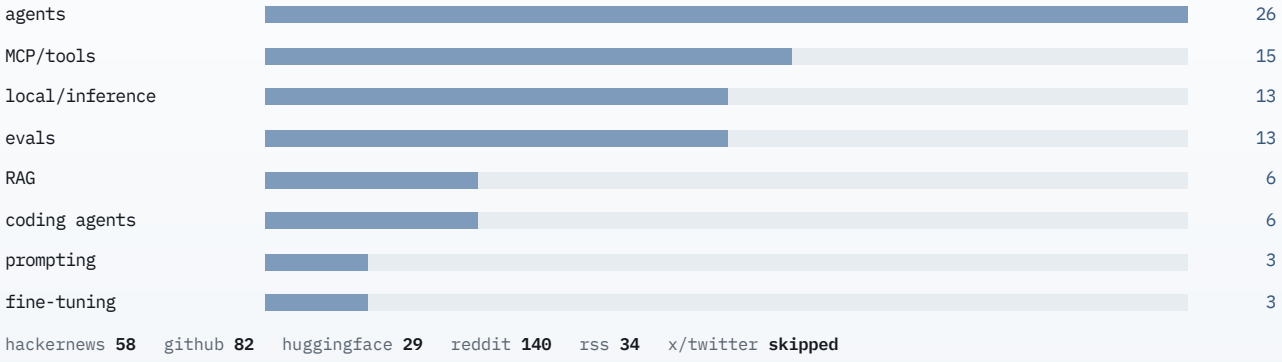
Today in AI engineering
2026-11-05

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 **343** items scanned mode: **mechanical fallback**

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

EXHIBIT 1 · TODAY'S SIGNAL — TOPIC FREQUENCY ACROSS SCANNED ITEMS N = 343



hackernews **58** github **82** huggingface **29** reddit **140** rss **34** x/twitter **skipped**

DEEP DIVES

Show HN: Codemap – one-page interactive map of any Go repository

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/peterretief/codemap

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: Gauge – Open-Source CLI to recursively improve your docs

Show HN

Hi HN, I'm Evan from Gauge. We're open-sourcing our CLI to recursively improve your docs for coding agents. The CLI includes a local execution backend, so you don't need an account to try it. Explainer video here: https://youtu.be/zhLUlIKB1Mg The CLI supports a hillclimbing loop: - set up tasks for

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/gauge-sh/gauge-cli

Show HN: AutoCua, one prompt drives computer, browser, phones, code at same time

Show HN

Hey HN! I am Ashish from AutoCua (automatic computer use agent). I am building super computer use agent. Imagine a computer use that is not limited to computer automation. If the task needs it, it write code, control browser through extension that it loads runtime to use same profile, control iphone

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/ComputerUseAgent/AutoCua

Show HN: Phonon TTS on WebGPU, 44MB, 115x real time, with live word timings

Show HN

I tried Gradium's Phonon 31M text-to-speech model and was blown away by its quality to size ratio. I needed timings for every word to be able to highlight them, so I wrote a second runtime for it with a different design: the whole generation loop stays on the GPU in WebGPU kernels written specifical

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/blixt/phonon

WHAT PEOPLE ACTUALLY BUILT

Show HN: I built a tool that prices a pre-built PC part by part

Show HN

I've missed out on an awesome computer that I could use for a project because figuring out value took too long. Users paste a listing and receive a value grade, dated component estimates and a shareable report. AI reads the listing but my custom built price engine determines the

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

https://pcdealcheck.com/

Can a 1-trillion-parameter model run on a 64 GB desktop? I built an open-source engine to find out, and it already beats llama.cpp on models bigger than RAM.

r/LLMDevs

Kimi K2 has 1 trillion parameters but uses only about 32 billion per word. Banach keeps the right experts in RAM and streams the rest from NVMe, so the model doesn't have to fit in memory. Measured: a 22 GB model on 16 GB of RAM. llama.cpp runs out of memory; Banach is 2.2x faste

> **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/1xo8uzn/can_a_1trillionparameter_model_run_on_a_64_gb/

recalc: a free spreadsheet app where each cell can be an agent that you can monitor and steer

r/LLMDevs

https://github.com/sshah03/recalc I was using Claude to help me build benchmark sets, and found pouring through CSVs / MD files of the results a bit tiresome, so I build a spreadsheet app for myself that's driven by Claude Code, and where each cell can be its own Agent if you wan

> **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/1xo3u3s/recalc_a_free_spreadsheet_app_where_each_cell_can/

QUICK HITS

ruvnet/ruvllm GITHUB

Governed neural weight editing and coding model adaptation. Native Rust patches, exact rollback, LoRA, Claude Code routing, and reproducible

https://github.com/ruvnet/ruvllm

MarkIvor/predictorlab GITHUB

Self-hosted tabular ML that doesn't lie. Leakage Guard, conformal intervals, profit-optimized thresholds, counterfactuals & instant REST/MCP

https://github.com/MarkIvor/predictorlab

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

exzade/EXZA-Remote GITHUB

Secure AI-powered remote computer access for ChatGPT, Claude and compatible MCP clients.

https://github.com/exzade/EXZA-Remote

Show HN: GPUWorks – A 40 MB GPU cluster manager for vLLM/SGLang, no K8s needed SHOW HN

https://github.com/flitai-io/gpuworks

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

Trending model on Hugging Face — text-to-image, gguf; 5656 likes, 233,059 downloads.

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

moonshotai/Kimi-K3.1-Pro HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 765 likes, 11,658 downloads.

https://huggingface.co/moonshotai/Kimi-K3.1-Pro

beipo-ai/NorthCUA GITHUB

Turn intent into work. Open-source Desktop/Web agent harness with long-horizon memory, live evaluation and recoverable batches.
https://github.com/beipo-ai/NorthCUA

speridlabs/idol HF MODEL

Trending model on Hugging Face — image-to-3d, idol; 677 likes, 17,180 downloads.
https://huggingface.co/speridlabs/idol

wendersoon/radar-cnpj-mcp GITHUB

Servidor MCP e CLI para localizar empresas brasileiras por CNAE e região, utilizando dados abertos de CNPJ da Receita Federal.
https://github.com/wendersoon/radar-cnpj-mcp

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

Trending model on Hugging Face — image-text-to-text, transformers; 1603 likes, 9,592 downloads.
https://huggingface.co/mistralai/Mistral-Large-4-1T-A52B

ai-gateways/awesome-ai-gateways GITHUB

Awesome AI gateways: open source and commercial LLM gateways, MCP gateways, agent (A2A) gateways, model routers, specs, and guides.
https://github.com/ai-gateways/awesome-ai-gateways

Show HN: An Open Source Superhuman Email Client Clone SHOW HN

Hi HN, I'm a long-time Superhuman user, and a few weeks ago I cloned it as a native Mac app. It's keyboard-first, with split inboxes, remind
https://github.com/michaelmagan/OpenHuman

Tell HN: Gemini trains on chat history, even for subscribers (personal accounts) HN

If you have a personal Google AI subscription, all your saved chat history in Gemini can be used for training and there is no way to opt out
https://news.ycombinator.com/item?id=50363461

sepehr071/iran-jobs-mcp GITHUB

Unofficial read-only MCP server for Iranian job sites (استخدام) IranTalent, Jobinja and Jobvision: job search, companies, reviews, interview
https://github.com/sepehr071/iran-jobs-mcp

Generated 2026-11-05 · 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.