DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

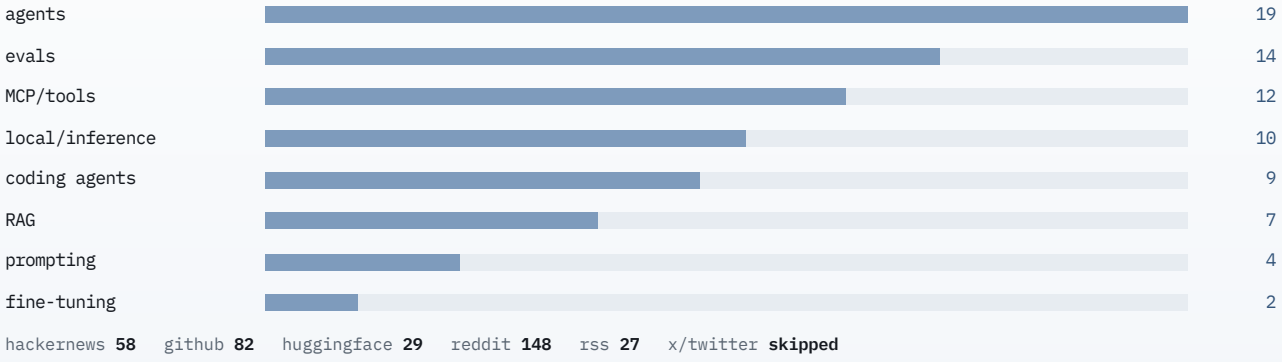
Today in AI engineering 2026-11-07

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 **344** items scanned mode: **mechanical fallback**

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

EXHIBIT 1 · TODAY'S SIGNAL — TOPIC FREQUENCY ACROSS SCANNED ITEMS N = 344



hackernews **58** github **82** huggingface **29** reddit **148** rss **27** x/twitter **skipped**

DEEP DIVES

Show HN: Agentic Identity Broker, OAuth Delegation for AI Agents

Show HN

Hi HN, I'm Jan. I am co-author of the Agentic Identity Broker (AIB), an MIT-licensed OAuth2 broker in the Zalando Incubator, that lets a user give an AI agent scoped, revocable access to MCP services. The broker solves three major challenges that are especially relevant in corporate environments: (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/zalando-incubator/agentic-identity-broker

Show HN: An open source webbluetooth and NFC browser for iOS

Show HN

I am developing hardware products with NFC and BLE. I built this browser because I lack support of WebNFC and WebBluetooth in existing browsers. Using a native app or a spare Android phone is not the fastest way to test a new feature / integration in my hardware. How it works: it uses WebKit and sta

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/hito-dev/wand-browser

Show HN: CloneGuard – Scan a repo for credential-stealing tricks before it runs

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/tharun208/cloneguard

Show HN: Agents infer Lego mechanisms from sources, then animate them

Show HN

Hi there :) Around a month ago, I released LDraw Nova, an agentic toolset for assembling LEGO CAD models by generating LDraw source code files. For more details, here's the Show HN post: https://news.ycombinator.com/item?id=49937916 From the comments, I got the impression that there was quite an int

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/anteloc/ldraw-mecha/tree/v0.5.0

ruvnet/ruvllm

GitHub

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

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/ruvnet/ruvllm

WHAT PEOPLE ACTUALLY BUILT

Show HN: Overhead – the planes overhead, painted on an e-ink frame

Show HN

Hi HN, I've been fascinated by aviation for some time now. Inspired by Fugleramme, I wanted a framed print of a cluster of aircraft over my house or a busy airport. So I built Overhead. (Photos in the README) The hero photo shows planes departing and landing in Changi, Singapore.

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

https://github.com/dmellok/tesserae-widget-overhead

I got tired of free AI models repeating the same broken fix 5 times, so I built an open-source memory & verification harness in Python (Open for contributors!)

r/LLMDevs

When using smaller or free-tier models (like Gemini 3.8 Flash or Qwen 3.8 27B) on real multi-file projects, two things usually break: installing too many skills bloats the startup prompt, and when a test fails, the model loops on the exact same broken fix. To fix this, I built Ap

> **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/1xpc2pv/i_got_tired_of_free_ai_models_repeating_the_same/

I built a native macOS Postgres client because they all suck

r/LLMDevs

DBJoy is a lightweight, native macOS Postgres client. It’s AI-first, with a built-in agent for working with your data and an MCP server so Claude Code can read (or write) to it. It also has all the features you’d expect from any other database client: diff and migration tools, tu

> **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/1xq1ht8/i_built_a_native_macos_postgres_client_because/

QUICK HITS

zaxiu88888/issue-rag GITHUB

RAG over GitHub issues, with a hand-reviewed eval set and an audit of the judge itself. Reports the interval, not just the best number.

https://github.com/zaxiu88888/issue-rag

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

Show HN: Ryph – Self-Hosted, Open-Source Coding Agents for Incident Response SHOW HN

https://github.com/ryph-hq/ryph

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

Glen-Web-App/best-agent-memory-mcp GITHUB

Best Agent Memory MCP servers, ranked and compared: long-term memory for Claude Code, Codex, Cursor and any MCP client. Glen, Mem0, Supermem

https://github.com/Glen-Web-App/best-agent-memory-mcp

phonny40rowle/rag-search GITHUB

A self-hosted retrieval-augmented search API: web search, content extraction, embeddings, and answer generation.
https://github.com/phonny40rowle/rag-search

speridlabs/idol HF MODEL

Trending model on Hugging Face — image-to-3d, idol; 1007 likes, 20,689 downloads.
https://huggingface.co/speridlabs/idol

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

Trending model on Hugging Face — text-to-image, gguf; 5697 likes, 65,099 downloads.
https://huggingface.co/abenzerps/Qwen-Image-2.1-Turbo-Quantized

Show HN: Goethe – The infrastructure for self-hosted AI SHOW HN

Hey, we launched Goethe, because deploying models on hardware you control end-to-end is pain in the ass. Goethe is like Vercel, but for AI m
https://github.com/meshapplied/Goethe

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

didilili/mianstack GITHUB

🚀面栈 MianStack | 程序员面试题库天花板！Java、Python、数据库、算法、大模型、AI Agent、RAG 一栈备齐。真实面经关联高频考点，按公司、岗位定向备考；图文题解、精华专题、项目源码拆解，全系标配。面试官出题，面栈出招！免费开源，持续更新。
https://github.com/didilili/mianstack

liangjianshu/multimodal-image-retrieval GITHUB

多模态检索实战：意图识别、Prompt 工程、Qwen LoRA、EmbeddingGemma 2 与 InfoNCE 对比学习。含真实演示、自有数据接入、训练评测及 RAG 检索组件学习指南。
https://github.com/liangjianshu/multimodal-image-retrieval

shanselman/sandbox-mcp GITHUB

Fresh, throwaway Windows machines for testing your apps: an MCP server and CLI that automate Windows Sandbox
https://github.com/shanselman/sandbox-mcp

StarTrail-org/DRPixel GITHUB

DR Pixel: Page-Native Deep Research with Vision-Language Models — benchmark, evaluation harness and training code
https://github.com/StarTrail-org/DRPixel

Show HN: BeepGrid – Lightweight open source agent comms SHOW HN

https://github.com/beepgrid/beepgrid