DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

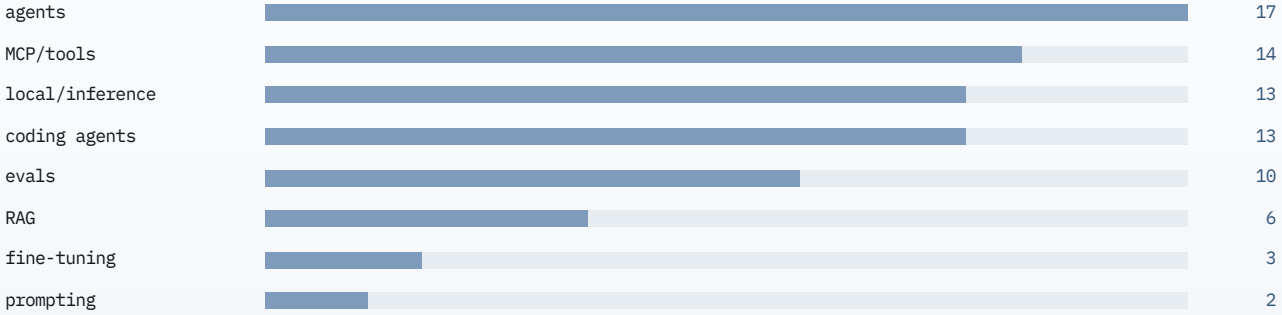
# Today in AI engineering 2026-11-09

*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 **313** items scanned mode: **mechanical fallback**

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

EXHIBIT 1 · TODAY'S SIGNAL — TOPIC FREQUENCY ACROSS SCANNED ITEMS N = 313




hackernews **56** github **79** huggingface **29** reddit **133** rss **16** x/twitter **skipped**

## DEEP DIVES

### Show HN: Flide: Make Jev-like decision models up to 9.5x faster

HN

Decision models need their own inference engine. So we developed Flide, a high-performance and exact inference engine for Jev-like decision models. Flide achieves up to a 9.5x speedup over vLLM and SGLang. It now supports 30+ popular local Jev-like decision models. Start to make your decision immedi

**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/OmniJev/Flide

### Llama-modchip: 284B DeepSeek on 64GB RAM, no GPU

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/synchopate/llama-modchip

### Show HN: Recalc – Spreadsheet where cells can be Claude Code or Codex terminals

Show HN

I built this because it was useful for me when building evals/benchmarks, and I decided to put it up for free for others. The goal was quite simple, make it easy to use a lead claude code / codex session in a panel on the side, while allowing every cell to also be a terminal session running claude/c

**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/sshah03/recalc

### shanselman/sandbox-mcp

GitHub

Fresh, throwaway Windows machines for testing your apps: an MCP server and CLI that automate Windows Sandbox

**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/shanselman/sandbox-mcp

### 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: Loop – your coding agents, from your phone

Show HN

Loop is a mobile-first companion workspace for the coding harnesses you already run. It aggregates results from your harnesses into a continuous feed that is easy to navigate and organize from your phone. I spent several months holding my newborn in one hand and trying to talk to

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

https://github.com/j-asefa/loop

## Show HN: Pgxio – a tiny PostgreSQL layer inside Nginx

Show HN

I built the first version of this stuff a decade ago as ethanolus (c2h5oh). And I've used it for most of my projects as a backend layer since then. The main motive is to eliminate unnecessary complexity from backend layers. How it works: a rust nginx module calls postgresql funct

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

https://github.com/mishabunte/pgxio

## Show HN: Harness Resource Bench – CPU, memory and disk of coding agents

Show HN

Hi HN. I built this to see what coding agents cost the machine they run on, not how well they code. Disclosure: I made smelt, which is one of the agents. The full method and code are in the repo!

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

https://leonardcser.github.io/harness-resource-bench

## QUICK HITS

### gogoymh/auto-ontology-mcp GITHUB

NVIDIA Auto Ontology's semantic layer and a personal knowledge base as an MCP server for Claude, ChatGPT and other MCP clients. No LLM API k
https://github.com/gogoymh/auto-ontology-mcp

### 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

### Show HN: My first project, an AI-powered financial news analyzer SHOW HN

I'm 17, and I started NotBot a few months ago to see whether I could turn an idea into a working project. My main interest is economics, not
https://github.com/lamarcol/NotBot

### Show HN: Nino – Fastest C# serializer, safe on untrusted input, also TS/Go/Rust SHOW HN

Nino is a source-generated binary serializer for C#, and as far as we know the fastest one. In the same BenchmarkDotNet run on GitHub runner
https://github.com/JasonXuDeveloper/Nino

### speridlabs/idol HF MODEL

Trending model on Hugging Face — image-to-3d, idol; 1514 likes, 24,747 downloads.
https://huggingface.co/speridlabs/idol

### xtools-at/DeepSeek-V4-Pro-Qwen3.5-9B-UNCENSORED-heretic HF MODEL

Trending model on Hugging Face — text-generation, transformers; 983 likes, 11,755 downloads.
https://huggingface.co/xtools-at/DeepSeek-V4-Pro-Qwen3.5-9B-UNCENSORED-heretic

### 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

### moonshotai/Kimi-K3.1-Pro HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 1768 likes, 27,926 downloads.
https://huggingface.co/moonshotai/Kimi-K3.1-Pro

### YT-Proteus-bot/MineAgent GITHUB

agent定制自由化，agent流程可视化。在浏览器里用画布搭出工业级 Agent 流程，选框架自动生成架构图， 拖节点、连线、写代码、配模型，点「保存」就把整张画布编译成一个可独立运行的 Python Agent 项目， 点「运行」直接在你的机器上真实执行并流式回传每一步。欢迎

https://github.com/YT-Proteus-bot/MineAgent

### Goldman Sachs is pulling Claude Code from its bankers and traders HN

https://www.thestateofai.com/news/goldman-sachs-claude-code-traders-bankers

### Nvidia CEO Jensen Huang's letter to staff about affair HN

https://www.axios.com/2026/11/08/nvidia-ceo-jensen-huang-aff... , https://archive.ph/Hxn5h

https://www.nytimes.com/2026/11/06/us/politics/ch

https://nvidianews.nvidia.com/news/a-message-to-our-company

### ChatGPT Desktop forwards your SSH agent to remote machines HN

The agent running on a remote machine can authenticate using your local SSH credentials. That means it can push to your private repos on Git

https://news.ycombinator.com/item?id=50422628

### LakoMoor/portcode GITHUB

Port anything to Rust, with proof. Open-source terminal agent for reverse engineering and Rust ports on local/self-hosted LLMs.

https://github.com/LakoMoor/portcode

### SHOW HN: GerGit - A Free Open-Source Git GUI Client for Linux Inspired by Fork SHOW HN

https://github.com/IcaruzSoftware/GerGit

### The Ultimate MoE Training Handbook HF SPACE

Trending Space (docker, 443 likes). A hands-on guide to training Mixture-of-Experts models

https://huggingface.co/spaces/huggingface/moe-blogpost

---

Generated 2026-11-09 · 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.