

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

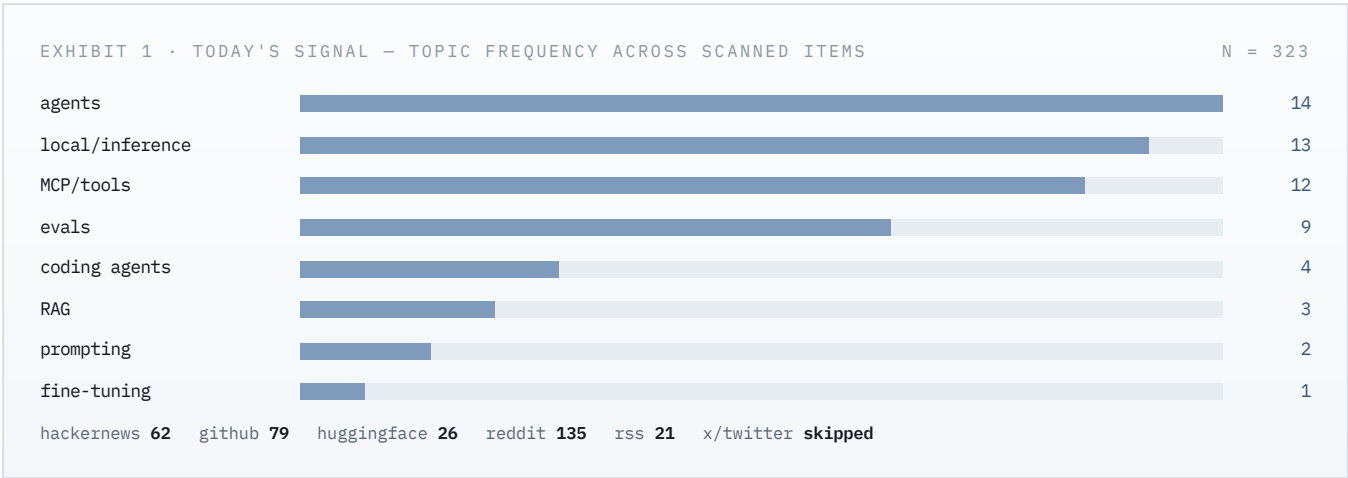
Today in AI engineering

2026-09-05

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 323 items scanned mode: **mechanical fallback**

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



DEEP DIVES

2akouwu/reverify

GitHub

Stop your AI from making things up — it proposes, deterministic tools decide, every claim checked against ground truth with evidence. Grounded facts and context survive resets. Reverse engineering is the proving ground. MCP server + CLI.

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/2akouwu/reverify>

Show HN: TERMy – A fast terminal assistant that does not use LLMs

[HN](#)

I love research and development, you may have heard of me because of PJON (Padded Jittering Operative Network). It is a network protocol I started developing in 2010, which was recently implemented in silicon by the ETH Zurich university thanks to the research of Pius Sieber. I am excited to share w

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/gioblu/NPC-Forge/blob/main/docs/development.md>

PillCrew/deedchain

[GitHub](#)

Benchmark of report fidelity for browser agents: does the agent's final report tell the truth about what it actually did?

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/PillCrew/deedchain>

Fermat's Last Theorem in Lean 4

[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/anthropics/fermats-last-theorem>

nahidalam/world_model_from_scratch

[GitHub](#)

Code repository for World Model From Scratch book

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/nahidalam/world_model_from_scratch

WHAT PEOPLE ACTUALLY BUILT

Show HN: Run open-weight OCR, VLM and vision models behind one API

[Show HN](#)

Hey HN. We built an openai-compatible API for running open-weight VLMs, OCR VLMs and ViT-based vision models. The motivation was mostly frustration when running these models in production and discovering all the details around serving VLMs, especially around visual accuracy. A fe

STEAL THIS

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

<https://www.vlmrun.com/gateway>

Show HN: Moadim.io – A scheduler for agents

[Show HN](#)

Why can't we get an agent scheduler that supports all of the following: - git compatible - agent agnostic - 100% open source - os and system-agnostic - multi-runner support - support mcp/ui/http - unlimited routines/crons So I built one, moadim.io is a local Rust daemon you insta

STEAL THIS

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

<https://moadim.io/>

Show HN: Free macOS App to import photos from memory card with file naming rules

[Show HN](#)

I love photography and I've been struggling with finding a tool to import photos from memory card with customizable file naming rules and copy verification. So I built the tool I wanted: Photo Batch Importer, renames with EXIF properties and auto-incremental sequence number, then

STEAL THIS

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

<https://github.com/jianfenglin-git/photo-batch-importer>

QUICK HITS

Show HN: Open-Source elnk Bike Computer [SHOW HN](#)

Hey all, i just launched my Eink Bike computer project and think it is cool. Another tidbit, in the crazy things that AI has done... It has <https://opentrailpaper.com>

Discovery of a new OpenAI agent message board [HN](#)

<https://www.reuters.com/world/europe/openai-agents-hijacked-...>
<https://collusion.wiki/>

adtexterry-lgtm/unigit-ecosystem [GITHUB](#)

UNIGIT public brand and ecosystem hub — AI should work for everyone.
<https://github.com/adtexterry-lgtm/unigit-ecosystem>

vcruz305/vllm-exl3 [GITHUB](#)

Serve EXL3 (ExLlamaV3 trellis) quantized models on vLLM fork runtimes — any architecture, mixed per-layer bitrates, composable with source-f
<https://github.com/vcruz305/vllm-exl3>

deepseek-ai/DeepSeek-V4-Flash-Vision-Exp [HF](#) [MODEL](#)

Trending model on Hugging Face — image-text-to-text, transformers; 628 likes, 133,024 downloads.
<https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-Vision-Exp>

BoardUI/boardui [GITHUB](#)

React design system for agentic interfaces. Every free BoardUI component as source, with a working AI chat app on your own model key as the
<https://github.com/BoardUI/boardui>

Corporate America is getting hooked on open-source AI [HN](#)

<https://archive.is/kmOqm>
<https://www.nytimes.com/2026/09/04/technology/open-source-ai-anthropic-openai.html>

Qwen/Qwen3.8-27B [HF](#) [MODEL](#)

Trending model on Hugging Face — image-text-to-text, transformers; 13980 likes, 6,024,467 downloads.
<https://huggingface.co/Qwen/Qwen3.8-27B>

Qwen/Qwen3.8-Flash-Next HF MODEL

Trending model on Hugging Face— image-text-to-text, transformers; 4889 likes, 351,374 downloads.
<https://huggingface.co/Qwen/Qwen3.8-Flash-Next>

Show HN: Aurict – An open-source Terminal AI SHOW HN

<https://github.com/aurict/aurict>

Show HN: Rubato – Retro-Mac desk device mirrors AI coding state-ESP8266 SHOW HN

Hi everyone! I have built an AI health device (ESP8266 + 240*240 screen). It can show AI Agents' status in real time with breathing bubble.
https://github.com/lovaxi/Rubato_Device

zai-org/GLM-5.3-Flash HF MODEL

Trending model on Hugging Face— image-text-to-text, transformers; 2059 likes, 654,957 downloads.
<https://huggingface.co/zai-org/GLM-5.3-Flash>

zai-org/GLM-5.3 HF MODEL

Trending model on Hugging Face— text-generation, transformers; 1710 likes, 303,534 downloads.
<https://huggingface.co/zai-org/GLM-5.3>

XHToken/Spark-X2.5-4B HF MODEL

Trending model on Hugging Face— text-generation, transformers; 493 likes, 3,524 downloads.
<https://huggingface.co/XHToken/Spark-X2.5-4B>

shtjww/llm-inference-capacity-handbook GITHUB

大模型推理案头手册（开源版）：给定 GPU 算力，一个模型能扛多少 QPS？三层模型 × 三面墙 × 排队论 × 开环压测
<https://github.com/shtjww/llm-inference-capacity-handbook>
