

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

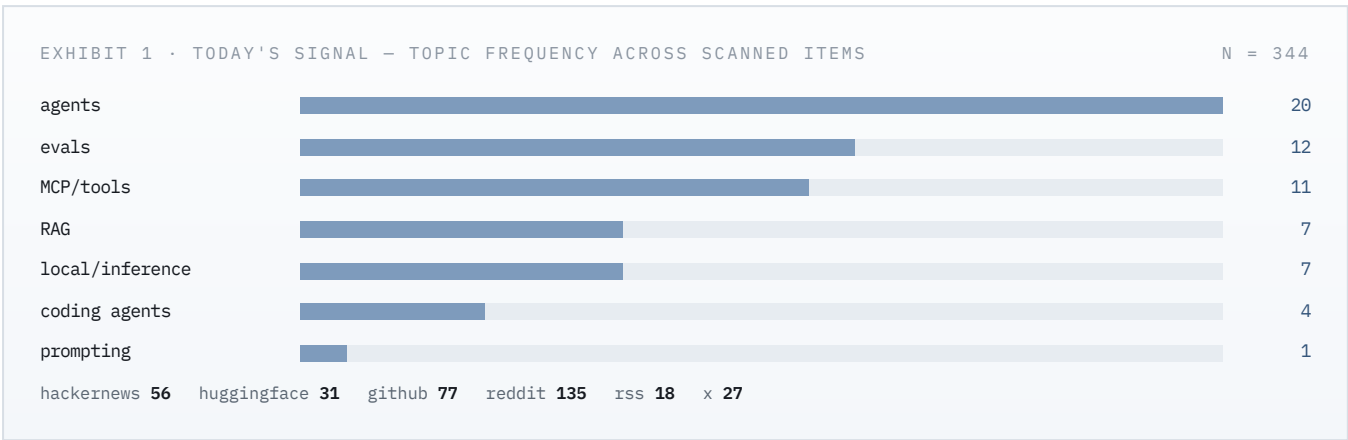
# Today in AI engineering

## 2026-08-18

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
- rag
- local/inference
- coding agents
- prompting



### DEEP DIVES

#### Show HN: Visimer – open-source visual editor for Mermaid diagrams

Show HN

Hi HN, Nick here, founder at OpenKnowledge. We built an open source library, Visimer, for editing Mermaid diagrams with a what-you-see-is-what-you-get visual editor. Try the playground at <https://visimer.com/playground>. You can rename labels, change shapes, add new nodes, etc. by clicking and dragg

**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/inkeep/visimer>

#### Qwen/Qwen3.8-27B

HF model

Trending model on Hugging Face — image-text-to-text, transformers; 10810 likes, 415,039 downloads.

**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://huggingface.co/Qwen/Qwen3.8-27B>

---

## decionis/agent-safe-pipeline

[GitHub](#)

Reference architecture for AI agents that propose actions but cannot authorize them — immutable intent capture, an independent Decionis policy verdict (ALLOW/ESCALATE/BLOCK), verified human approval, and a SafeExecutor that consumes a single-use intent-bound grant.

**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/decionis/agent-safe-pipeline>

---

## Leutenegger/book-to-skill

[GitHub](#)

Turn any technical book PDF into a Claude Code skill — ready to study, reference, and use while you work.

**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/Leutenegger/book-to-skill>

---

## syv-ai/qwen38-27b-rtx3090

[GitHub](#)

Qwen3.8-27B on a single RTX 3090 with vLLM: ~1,000 tok/s at 64 concurrent (int8 tensor-core GEMMs, fp16 DeltaNet state), 84-90 tok/s single-user (MTP with cheap drafts), 150k context — 262k with the ported KVarN 4/2-bit KV cache; patches, requant scripts, 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/syv-ai/qwen38-27b-rtx3090>

---

## WHAT PEOPLE ACTUALLY BUILT

---

### Show HN: Desktopcolors.com – A museum for solid background colors of classic OS

[Show HN](#)

Hi HN, in my last vacation, I built desktopcolors.com as a place to collect iconic solid background colors. Let me know what you think.

### STEAL THIS

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

<https://desktopcolors.com>

---

### Show HN: LinXiv v0.4.0; paper manager now using n0's iroh for P2P collaboration

[Show HN](#)

I made a post here about linXiv about two months ago but since have made a ton of QOL updates and mini, non-obtrusive features (RSS feeds, full-text search, to name a few) and finally added iroh for p2p sharing of research projects, notes, etc... YT Demo:

<https://www.youtube.com/>

### STEAL THIS

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

<https://github.com/linxiv-dev/linXiv>

---

## Show HN: Saggar, a Mac terminal that keeps sessions and your attention organized

[Show HN](#)

Hi HN, like many of you talented folk, I've been building more, faster than ever. I built Saggar after my terminal stopped feeling like a tool and started feeling like a tab-management problem. Let me know you think!

### STEAL THIS

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

<https://saggar.marginalutility.dev/>

---

### QUICK HITS

---

#### unsloth/Qwen3.8-27B-GGUF HF MODEL

Trending model on Hugging Face — model; 1671 likes, 2,727,609 downloads.

<https://huggingface.co/unsloth/Qwen3.8-27B-GGUF>

---

#### memoket/memoket-kite GITHUB

An open-source, vector-free long-term memory engine for AI agents, achieving SOTA on LoCoMo and LongMemEval with significantly less context.

<https://github.com/memoket/memoket-kite>

---

#### Oxblewalker/memory-recall-rerank GITHUB

Reproducible benchmark and reference pipeline for reranked long-term memory retrieval

<https://github.com/Oxblewalker/memory-recall-rerank>

---

#### Show HN: UL-SMF – Open-source linear-complexity ~300x KV-cache compression SHOW HN

<https://github.com/livetruth/UL-SMF-Cache-Compression>

---

#### Qwen/Qwen3.8-2.4T-A95B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 1043 likes, 9,465 downloads.

<https://huggingface.co/Qwen/Qwen3.8-2.4T-A95B>

---

#### Lightricks/LTX-2.5 HF MODEL

Trending model on Hugging Face — image-to-video, diffusion-single-file; 1135 likes, 465,529 downloads.

<https://huggingface.co/Lightricks/LTX-2.5>

---

#### MiniMaxAI/MiniMax-Music3 HF MODEL

Trending model on Hugging Face — text-to-audio, diffusers; 915 likes, 10,375 downloads.

<https://huggingface.co/MiniMaxAI/MiniMax-Music3>

---

#### Show HN: HarnessRouter: Unified interface for agent harnesses SHOW HN

Hey HN! We are building HarnessRouter, a canonical API for running Codex, Claude Code, Hermes, and other managed agent harnesses as your pro

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

---

#### soumatheusgomes/vibe-coding-toolkit GITHUB

A curated, battle-tested AI-coding toolkit: Claude Code plugins, subagent orchestration, quality gates, and ready-to-copy prompts, extracted

<https://github.com/soumatheusgomes/vibe-coding-toolkit>

---

#### meta-models/Muse-Glimmer-30B HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 1667 likes, 334,099 downloads.

<https://huggingface.co/meta-models/Muse-Glimmer-30B>

---

#### OpenVuln HF SPACE

Trending Space (docker, 86 likes). Find bugs in your repository with GLM

<https://huggingface.co/spaces/zai-org/OpenVuln>

---

#### MiniMaxAI/MiniMax-H3 HF MODEL

Trending model on Hugging Face — image-text-to-video, minimax-h3; 4098 likes, 2,403,238 downloads.

<https://huggingface.co/MiniMaxAI/MiniMax-H3>

---

**[Open Source] Native DeepSeek-V4-Flash-0731 284B-A13B inference on one laptop CPU in pure C** R/LLMDEVs

Affiliation up front: this is an open-source project I maintain. Disclosure: I used an AI assistant to help refine the English wording of th  
[https://www.reddit.com/r/LLMDevs/comments/1vqxs2/open\\_source\\_native\\_deepseekv4flash0731\\_284ba13b/](https://www.reddit.com/r/LLMDevs/comments/1vqxs2/open_source_native_deepseekv4flash0731_284ba13b/)

---

**smshahbaj/crucible** GITHUB

Production-ready Claude Code decision intelligence marketplace with adversarial routing, evidence-driven analysis, specialized agents, bench  
<https://github.com/smshahbaj/crucible>

---

**Show HN: Eve Software Factory** SHOW HN

Build a software factory in minutes with this free OSS template  
<https://github.com/vercel-labs/eve-software-factory-template/tree/main>