

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

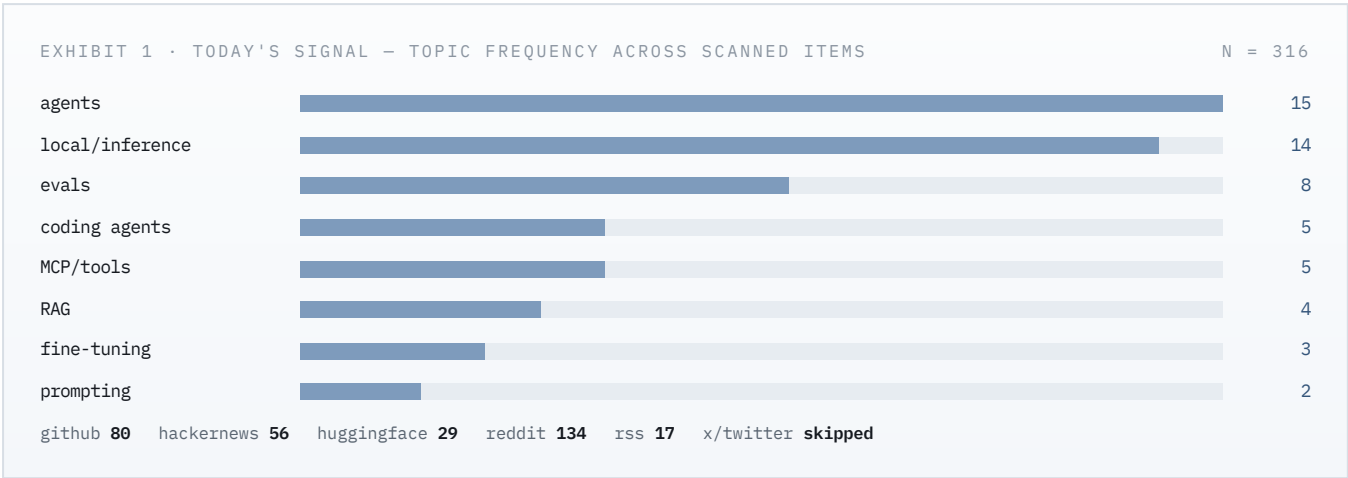
Today in AI engineering

2026-09-21

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 316 items scanned mode: **mechanical fallback**

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



DEEP DIVES

incoai/splash

GitHub

A local inference engine for Apple silicon, built around the model.

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/incoai/splash>

Show HN: Mini-AGI – Dynamic continual learning model trained on 8GB VRAM

[HN](#)

Sorry for the pretentious name, I know, I know.. It just contains all the pieces I would like to see a AGI model to have, and I can't stand the temptation. Before throwing rocks at me, please take a glance at the Readme, and I hope it will cover your mood a little bit. So, first of all it does 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/volotat/mini-AGI/>

pallavi-shekhar/ai-engineering-interview-questions-company-wise

[GitHub](#)

Your Cheat Sheet For AI Engineering Interviews at Top AI Companies - Questions and Answers.

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/pallavi-shekhar/ai-engineering-interview-questions-company-wise>

anandi1989/awesome-jev-usecases

[GitHub](#)

Evidence-backed index of real-world Jev (TypeSafe AI System One) use cases: repos, patterns, benchmarks, and measured results

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/anandi1989/awesome-jev-usecases>

prism-ml/Ternary-Bonsai-2-27B-gguf

[HF model](#)

Trending model on Hugging Face — text-generation, llama.cpp; 1624 likes, 2,227,879 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/prism-ml/Ternary-Bonsai-2-27B-gguf>

WHAT PEOPLE ACTUALLY BUILT

Show HN: Swarm, an offline family tree app for macOS

[Show HN](#)

Hi HN! My wife and I got deep into our family genealogy and wanted to map out all our relatives - but keep the data on our own Macs instead of uploading it to a genealogy service. Nothing we found did that, so we built Swarm, a native macOS app in Swift. It's built on GEDCOM 5.5.

STEAL THIS

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

<https://github.com/samoilev/swarm>

Show HN: Bailout – The coding agent meant to be deleted

[Show HN](#)

hey hn, i made bailout for a pretty specific problem. you boot a fresh vm and want an agent to help get it set up, except you dont have your agent set up yet. no gh, no config, no api key handy. might not even have codex / claude code / opencode curled and loaded. or you already

STEAL THIS

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

<https://github.com/storozhenko98/bailout>

Run Jev-style typed decisions locally on Apple Silicon

[r/LLMDevs](#)

I built Laya MPS, a PyTorch runtime for running Jev-style typed decisions locally on Apple Silicon GitHub repo:

<https://github.com/afshinm/laya-mps> On an M5 Pro: - Default mode: ~32 ms latency, ~2.1 GB peak process RAM - Lowest-memory mode: ~171 ms latency, ~0.74 GB peak process

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/1wm6we4/run_jevstyle_typed_decisions_locally_on_apple/

QUICK HITS

convaiinnovations/laya HF MODEL

Trending model on Hugging Face — text-classification, transformers; 1409 likes, 0 downloads.

<https://huggingface.co/convaiinnovations/laya>

Qwen/Qwen-Image-2.1 HF MODEL

Trending model on Hugging Face — text-to-image, diffusers; 1127 likes, 6,523 downloads.

<https://huggingface.co/Qwen/Qwen-Image-2.1>

ruc-datalab/EvoOntology GITHUB

EvoOntology: A Self-Evolving Ontology Layer for Data Agents 🌟 EvoOntology插件为Claude Code/Codex 建立&进化本体层

<https://github.com/ruc-datalab/EvoOntology>

zeeshan8281/slo-router GITHUB

SLO-aware LLM inference router with Jev decisions, live queue metrics, counterfactual evaluation, and reproducible latency/cost benchmarks

<https://github.com/zeeshan8281/slo-router>

Heman10x-NGU/openJev-verdict-2.0 GITHUB

Calibrated 151M Non-Autoregressive Decision Engine beating TypeSafe Jev & Laya on LocalLLaMA/typed-decisions (77.10% acc, 0.0636 Brier, 0.01

<https://github.com/Heman10x-NGU/openJev-verdict-2.0>

XingChen-AGI/Xing4.0-29B-A4B HF MODEL

Trending model on Hugging Face — text-generation, transformers; 990 likes, 18,394 downloads.

<https://huggingface.co/XingChen-AGI/Xing4.0-29B-A4B>

deepseek-ai/DeepSeek-V4.1-Flash HF MODEL

Trending model on Hugging Face — image-text-to-text, transformers; 3473 likes, 512,120 downloads.

<https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash>

Show HN: jevals – replacing LLM judges with typed Jev decisions SHOW HN

<https://github.com/openlayer-ai/jevals>

elsechord/CyberGuard GITHUB

Evidence-driven autonomous SOC team on AgentTeams: hash-bound approvals, HMAC audit chain, rollback and reproducible attack scenarios — audi

<https://github.com/elsechord/CyberGuard>

INSANE0777/Awwards-mcp GITHUB

Free, open-source MCP server that gives AI coding agents (Claude Code, Codex, Cursor, OpenCode, ZCode...) design inspiration from award-winnin

<https://github.com/INSANE0777/Awwards-mcp>

Qwen/Qwen3.8-27B HF MODEL

Trending model on Hugging Face— image-text-to-text, transformers; 15907 likes, 7,153,238 downloads.

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

erendikmenn/jev-rag-benchmark GITHUB

Reproducible benchmark for measuring Jev reranking quality, latency, and cost in RAG

<https://github.com/erendikmenn/jev-rag-benchmark>

valentynkit/awesome-jev-typesafe GITHUB

Typed decisions with TypeSafe's Jev, the first System One model

<https://github.com/valentynkit/awesome-jev-typesafe>

ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF HF MODEL

Trending model on Hugging Face— image-text-to-text, gguf; 1503 likes, 1,292,471 downloads.

<https://huggingface.co/ISTA-DASLab/Qwen3.8-27B-GSQ-RCO-GGUF>

harshatheg/Qwen-2.5-1B-RLCD HF MODEL

Trending model on Hugging Face— text-generation, mlx; 491 likes, 0 downloads.

<https://huggingface.co/harshatheg/Qwen-2.5-1B-RLCD>
