

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

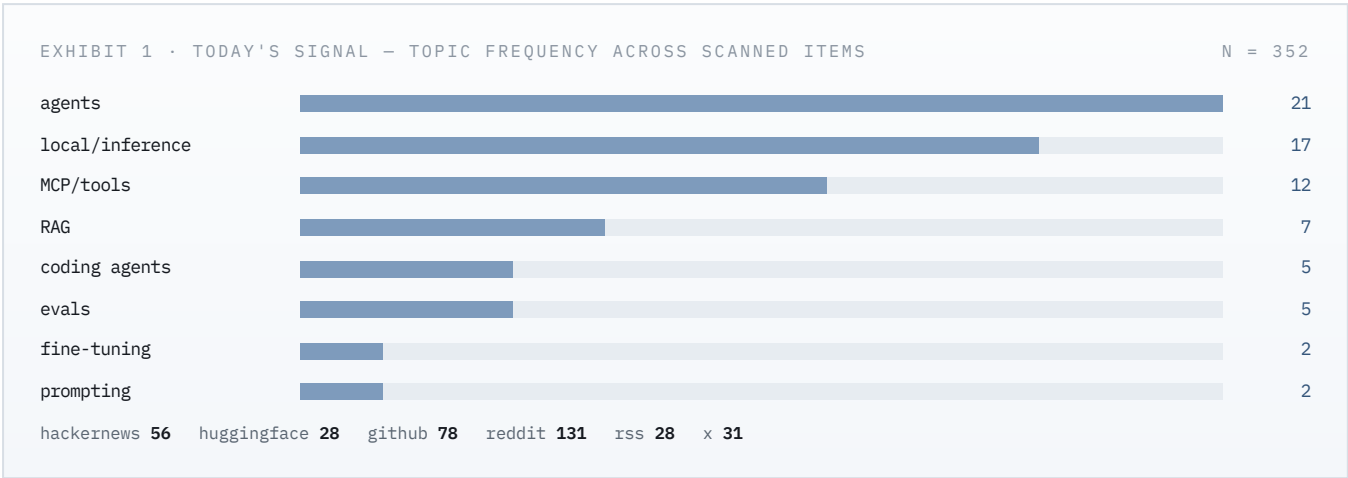
Today in AI engineering

2026-07-30

Developers are building and sharing open-source tools for AI agents, including specialized inference engines, merge queues, and snapshot testing libraries.

3 deep dives 2 builder stories 13 quick hits 352 items scanned

- ai agents
- inference engines
- merge queues
- snapshot testing
- open-source tools



DEEP DIVES

Show HN: Open-source engine running Gemma 4 26B in 2 GB RAM on any M-series Mac

hackernews

A developer built an open-source inference engine called TurboFieldfare, which can run Gemma 4 26B on any M-series Mac using about 2 GB of RAM. The engine is written in Swift and Metal. The developer's goal was to make it possible to run powerful AI models on-device, without relying on cloud services.

Why it matters. This project demonstrates the potential for running complex AI models on consumer-grade hardware, which could enable new use cases for AI-powered applications. It also highlights the importance of optimizing AI models for low-resource devices.

TRY THIS

Try running the TurboFieldfare engine on your own M-series Mac to see how it performs with different AI models.

<https://github.com/drumih/turbo-fieldfare>

- ai inference
- on-device ai
- m-series mac

moonshotai/Kimi-K3

huggingface

The Kimi-K3 model is a trending model on Hugging Face, with 8776 likes and 99,214 downloads. It is an image-text-to-text model, built using transformers.

Why it matters. The Kimi-K3 model is a popular choice for image-text-to-text tasks, and its availability on Hugging Face makes it easy for developers to integrate into their own projects. Its high download count suggests that it is widely used in the AI community.

TRY THIS

Try using the Kimi-K3 model in your own project to see how it performs on image-text-to-text tasks.

<https://huggingface.co/moonshotai/Kimi-K3>

image-text-to-text

transformers

hugging face

mikehasa/agentacct

github

A developer built a local-first dashboard for tracking the activity of AI agents, including Claude Code, Codex, and OpenCode. The dashboard provides a detailed breakdown of each task, including the tools used, files changed, tests run, and time and tokens spent.

Why it matters. This project highlights the importance of transparency and accountability in AI development, particularly when working with multiple agents. It also demonstrates the potential for local-first solutions to provide more control and flexibility than cloud-based alternatives.

TRY THIS

Try using the agentacct dashboard to track the activity of your own AI agents and see how it can help you optimize their performance.

<https://github.com/mikehasa/agentacct>

ai agents

local-first

dashboard

WHAT PEOPLE ACTUALLY BUILT

Show HN: Qwen Scribe – local transcription and dictation for Apple Silicon

hackernews

A developer built Qwen Scribe, a local transcription and dictation tool for Apple Silicon devices.

How it works. Qwen Scribe uses on-device AI to transcribe audio and provide dictation capabilities, without relying on cloud services.

STEAL THIS

The developer's use of on-device AI for transcription and dictation highlights the potential for local-first solutions to provide more control and flexibility than cloud-based alternatives.

<https://github.com/VladUZH/qwen-scribe>

Show HN: A local merge queue for parallel Claude Code agents

hackernews

A developer built a local merge queue for parallel Claude Code agents, which allows for more efficient and controlled development workflows.

How it works. The merge queue enables developers to manage multiple agents and their respective tasks, reducing the risk of conflicts and errors.

STEAL THIS

The developer's use of a local merge queue highlights the importance of workflow management and control when working with multiple AI agents.

<https://github.com/funador/claude-code-merge-queue>

Show HN: Parse a PDF from your terminal with PaddleOCR-VL-1.6 HACKERNEWS

A developer built OpenParser, a CLI tool for parsing PDFs using PaddleOCR-VL-1.6.

<https://github.com/eigenpal/openparser>

Oxwilliamortiz/openclaude-improved GITHUB

A developer improved OpenClaude, an open-source alternative to Claude AI.

<https://github.com/Oxwilliamortiz/openclaude-improved>

Handbook.md shows that long policy documents do not reliably govern agents HACKERNEWS

A research paper found that long policy documents are not effective in governing AI agents.

<https://arxiv.org/abs/2607.25398>

bybit-exchange/kaas GITHUB

A developer built Kaas, a tool for turning scattered notes and docs into a queryable Markdown wiki.

<https://github.com/bybit-exchange/kaas>

OneInterface/stormy-cookbook GITHUB

A developer built Stormy Cookbook, an open-source cookbook for the Stormy Social Data API and MCP server.

<https://github.com/OneInterface/stormy-cookbook>

Zazza/svoi-oss GITHUB

A developer built Svoi OSS, a self-hosted, open-source personal AI assistant with streaming chat and long-term memory.

<https://github.com/Zazza/svoi-oss>

QwenAudio/qwen-audio-agent GITHUB

A developer built Qwen Audio Agent, a realtime voice runtime for AI agents.

<https://github.com/QwenAudio/qwen-audio-agent>

Anatomy of a Frontier Lab Agent Intrusion: A Timeline of the July 2026 Incident HACKERNEWS

A blog post detailed the timeline of a recent agent intrusion incident at Frontier Lab.

<https://huggingface.co/blog/agent-intrusion-technical-timeline>

2001z/llm-engineering-platform GITHUB

A developer built an LLM engineering platform with OpenAI-compatible serving, streaming, and evaluation capabilities.

<https://github.com/2001z/llm-engineering-platform>

citrolabs/ego-browser-benchmark-framework GITHUB

A developer built a benchmark framework for comparing model browser-use performance under controlled settings.

<https://github.com/citrolabs/ego-browser-benchmark-framework>

casio review20-glitch/forged-os GITHUB

A developer built Forge OS, an open-source control plane for AI agents with skill routing, context governance, and multi-agent orchestration.

<https://github.com/casio review20-glitch/forged-os>

HezaoHezao/poirot GITHUB

A developer built Poirot, a deep research agent kernel for those who care about agent architecture.

<https://github.com/HezaoHezao/poirot>

enowdev/succubus GITHUB

A developer built Succubus, a cross-agent coordination tool for AI coding agents.

<https://github.com/enowdev/succubus>
