

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

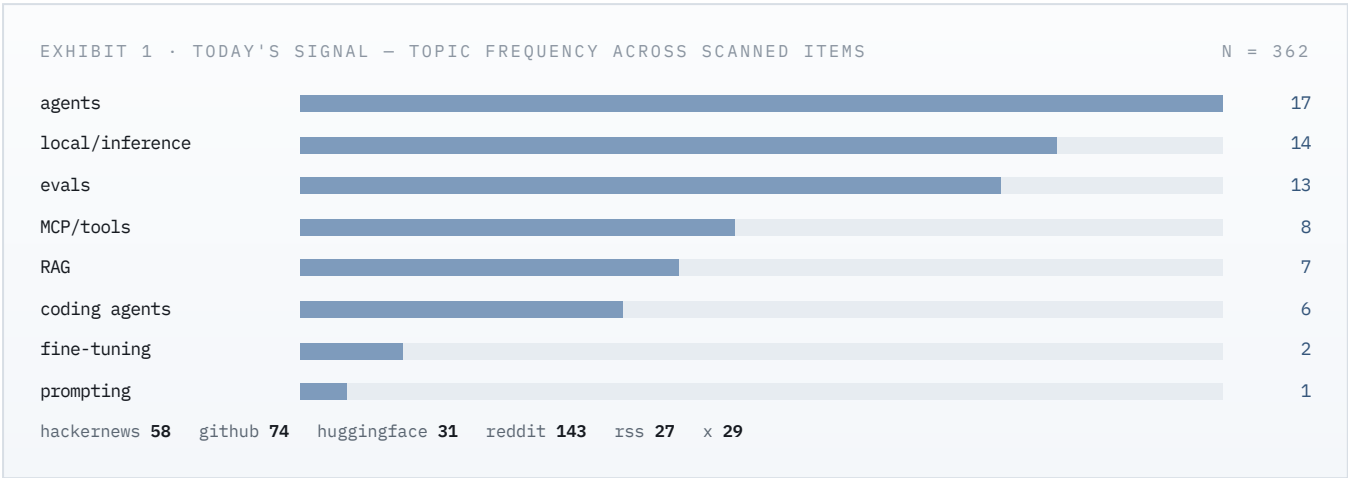
Today in AI engineering

2026-08-12

Developers are actively building and sharing tools for AI agent automation, LLM inference, and multi-agent systems

3 deep dives 2 builder stories 16 quick hits 362 items scanned

- llm
- ai agents
- inference optimization
- multi-agent systems



DEEP DIVES

H3-metal – Native MiniMax-H3 inference for Apple Silicon

hackernews

The developer antirez has created a native implementation of MiniMax-H3 inference for Apple Silicon, focusing on performance optimization. This project aims to provide a high-performance solution for AI inference on Apple devices. By leveraging the capabilities of Apple Silicon, the developer has achieved significant speed improvements.

Why it matters. This project matters because it demonstrates the potential for optimized AI inference on specific hardware platforms, which can lead to more efficient and powerful AI applications.

TRY THIS

Explore the h3.c repository and try compiling and running the inference code on an Apple Silicon device to see the performance benefits.

<https://github.com/antirez/h3.c>

- apple silicon
- llm inference
- optimization

sv-number/mcp-server

[github](#)

The developer sv-number has created an MCP server for AI agents that need a phone number, allowing for the ordering of private numbers in over 200 countries and the reading of SMS verification codes. This server aims to provide a convenient and automated solution for AI agents requiring phone number functionality.

Why it matters. This project is important because it addresses a common need in AI agent development, providing a straightforward way to integrate phone number capabilities into AI applications.

TRY THIS

Visit the sv-number/mcp-server repository and explore the API documentation to understand how to integrate the MCP server into your own AI agent projects.

<https://github.com/sv-number/mcp-server>

[mcp server](#)[ai agents](#)[phone number integration](#)

yangzc/ainote

[github](#)

The developer yangzc has created an AI agent workflow platform called ainote, which features visual flow orchestration, drag-and-drop forms, knowledge base RAG, multi-model LLM, and digital workers. This platform aims to provide a comprehensive toolset for building and managing AI workflows.

Why it matters. This project is significant because it offers a unified platform for designing, executing, and managing AI workflows, making it easier for developers to create complex AI applications.

TRY THIS

Explore the yangzc/ainote repository and try setting up a simple AI workflow using the platform's visual orchestration tools.

<https://github.com/yangzc/ainote>

[ai workflow](#)[visual orchestration](#)[rag](#)

WHAT PEOPLE ACTUALLY BUILT

Show HN: Browser Agent – cost efficient browser automation

[hackernews](#)

The developer visnia-ai built a browser agent harness that outperforms Browser Code on their own benchmark, achieving a higher success rate, speed, and cost efficiency.

How it works. The browser agent harness utilizes a custom-built framework to automate browser interactions, allowing for more efficient and reliable browser automation.

STEAL THIS

Developers can apply similar principles to their own browser automation projects by focusing on optimizing the automation framework and leveraging custom-built solutions.

<https://github.com/visnia-ai/browser-agent>

Show HN: TermDOM – HTML, CSS and JavaScript (With a Real DOM) for TUIs and CLIs

hackernews

The developer bikeshaving built TermDOM, a library that renders a real DOM with HTML, CSS, and JavaScript for terminal user interfaces (TUIs) and command-line interfaces (CLIs).

How it works. TermDOM uses a custom rendering engine to translate web technologies into terminal-friendly output, enabling the creation of complex and interactive TUIs and CLIs.

STEAL THIS

Developers can use TermDOM to create more engaging and interactive terminal applications by leveraging web technologies and the library's rendering capabilities.

<https://github.com/bikeshaving/termdom>

QUICK HITS

Show HN: Git-knife – Edit commit messages, authors, and dates like a spreadsheet HACKERNEWS

A tool for easily editing Git commit history.

<https://github.com/TheRealYT/git-knife>

H3-metal – Native MiniMax-H3 inference for Apple Silicon HACKERNEWS

A native implementation of MiniMax-H3 inference for Apple Silicon.

<https://github.com/antirez/h3.c>

sv-number/mcp-server GITHUB

An MCP server for AI agents that need a phone number.

<https://github.com/sv-number/mcp-server>

Apple Silicon and macOS VMs: Faster LLM Inference with llama.cpp HACKERNEWS

A blog post about achieving faster LLM inference on Apple Silicon using llama.cpp.

<https://github.com/trycua/cua/blob/main/blog/gpu-passthrough-macos-vm.md>

i3T4AN/KADATH GITHUB

An evolutionary multi-agent runtime for autonomous agents.

<https://github.com/i3T4AN/KADATH>

Show HN: Mcptoon – Token-efficient MCP CLI client HACKERNEWS

A token-efficient MCP CLI client.

<https://github.com/activeing123/mcptoon>

Show HN: ExtractBench, an open-source schema extraction benchmark HACKERNEWS

An open-source benchmark for schema extraction.

<https://github.com/run-llama/ExtractBench>

ombori/deepseek-v4-flash-0731-sglang-4x-rtx-pro-6000 GITHUB

A reproducible recipe for DeepSeek-V4-Flash-0731 on 4x RTX PRO 6000.

<https://github.com/ombori/deepseek-v4-flash-0731-sglang-4x-rtx-pro-6000>

MiniMaxAI/MiniMax-H3 HUGGINGFACE

A trending model on Hugging Face for image-text-to-video tasks.

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

Show HN: Browser Agent – cost efficient browser automation HACKERNEWS

A cost-efficient browser automation tool.

<https://github.com/visnia-ai/browser-agent>

dauidahmann/fde-guide GITHUB

A forward deployed engineering guide for production AI systems.

<https://github.com/dauidahmann/fde-guide>

matonhp5108/Opencodex GITHUB

An open-source coding agent inside VS Code.

<https://github.com/matonhp5108/Opencodex>

Juror-AI/juror GITHUB

A cheaper and better alternative to Greptile that runs on GitHub Actions.

<https://github.com/Juror-AI/juror>

moonshotai/Kimi-K3 HUGGINGFACE

A trending model on HuggingFace for image-text-to-text tasks.

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

Stealing Reasoning Traces from Proprietary LLM APIs HACKERNEWS

A website discussing the theft of reasoning traces from proprietary LLM APIs.

<https://stolen-thoughts.com/>

vpromise/inside-coding-agents GITHUB

A visual, evidence-backed handbook and lab for coding agent architecture.

<https://github.com/vpromise/inside-coding-agents>
