

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

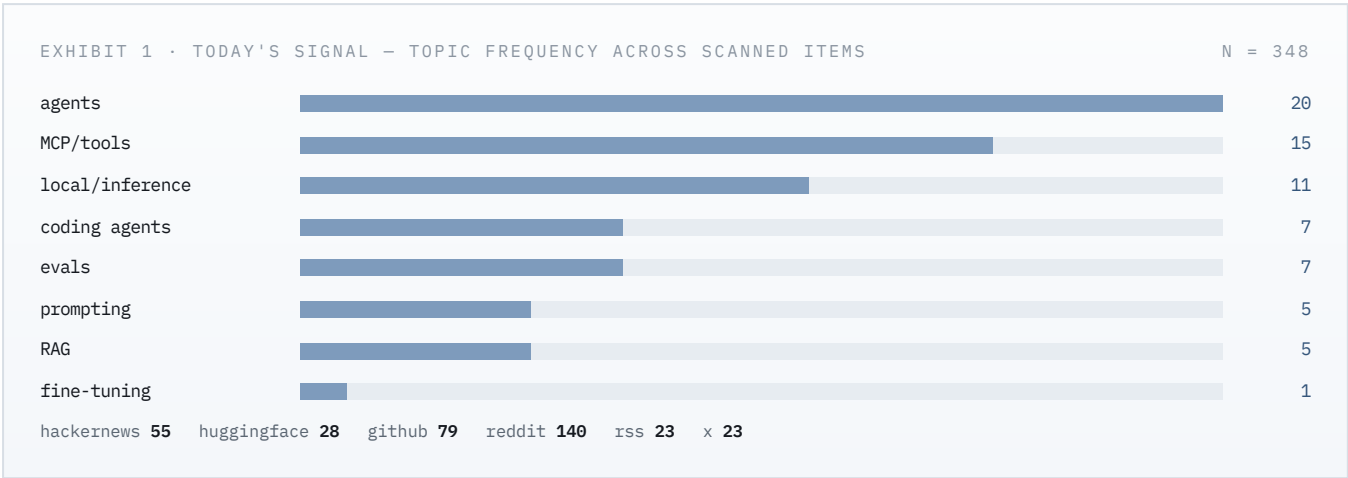
Today in AI engineering

2026-07-29

Developers are building and sharing open-source tools for AI agents, LLMs, and automation

5 deep dives 2 builder stories 17 quick hits 348 items scanned

- llms
- ai agents
- open-source
- automation
- transformers



DEEP DIVES

moonshotai/Kimi-K3

huggingface

The Kimi-K3 model is a trending image-text-to-text model on Hugging Face. It has 8293 likes and 99,214 downloads. The model uses transformers and is suitable for various applications. A developer can use this model as a starting point for their own projects.

Why it matters. The Kimi-K3 model is a popular and widely-used model that can be used for various applications. It is a good example of how developers can use pre-trained models and fine-tune them for their own use cases.

TRY THIS

Try fine-tuning the Kimi-K3 model for a specific task, such as image captioning or text generation.

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

- llms
- transformers
- image-text-to-text

Show HN: I was tired of opening 2 tabs for every HN link, so I made a userscript

hackernews

A developer created a userscript to solve the problem of having to open two tabs for every Hacker News link. The script allows users to view the discussion and the article in the same tab. The developer shared the script on GitHub and received positive feedback from the community.

Why it matters. This project shows how developers can use their skills to solve real-world problems and share their solutions with others. It also highlights the importance of community feedback and collaboration.

TRY THIS

Try creating a userscript to solve a problem that you face regularly, and share it with the community.

<https://github.com/twalichiewicz/HNewhere>

userscript

hacker news

productivity

Cynative – Read-only CLI in Go that explains your live infrastructure

hackernews

The Cynative project is a read-only CLI tool written in Go that helps users understand their live infrastructure. It connects to various cloud services and provides a simple way to answer security questions. The project is open-source and available on GitHub.

Why it matters. The Cynative project is a useful tool for developers who need to understand their infrastructure and answer security questions. It is also a good example of how developers can use Go to build CLI tools.

TRY THIS

Try using the Cynative tool to understand your own infrastructure and answer security questions.

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

cli

go

infrastructure

Oxwilliamortiz/openclaude-improved

github

The OpenClaude project is an open-source implementation of the Claude AI model. The improved version is available on GitHub and includes various enhancements and bug fixes. The project is a good example of how developers can collaborate and improve existing models.

Why it matters. The OpenClaude project is a useful resource for developers who want to use the Claude AI model in their own projects. It is also a good example of how open-source collaboration can lead to improved models and tools.

TRY THIS

Try using the OpenClaude model in your own project and contribute to the open-source community.

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

llms

open-source

collaboration

XYZ-AI-Lab/axrl

github

The AxRL project is an agentic RL post-training framework built on SGLang rollout, Megatron training, and real-world agent workflows. The project is available on GitHub and includes various tools and resources for developers. The framework is designed to help developers train and deploy AI agents in real-world environments.

Why it matters. The AxRL project is a useful resource for developers who want to train and deploy AI agents in real-world environments. It is also a good example of how developers can use frameworks and tools to simplify the development process.

TRY THIS

Try using the AxRL framework to train and deploy your own AI agents.

<https://github.com/XYZ-AI-Lab/axrl>

ai agents

rl

post-training

Show HN: I was tired of opening 2 tabs for every HN link, so I made a userscript

hackernews

A userscript to view Hacker News discussions and articles in the same tab.

How it works. The script uses JavaScript to modify the Hacker News webpage and display the discussion and article in the same tab.

STEAL THIS

Developers can use their skills to solve real-world problems and share their solutions with others.

<https://github.com/twalichiewicz/HNewhere>

Show HN: Minute – Offline meeting notes on macOS with Whisper and llama.cpp

hackernews

An offline meeting notes app for macOS that uses Whisper and llama.cpp for transcription and summarization.

How it works. The app uses Whisper for transcription and llama.cpp for summarization, and stores the notes locally on the user's device.

STEAL THIS

Developers can use open-source libraries and frameworks to build useful tools and apps.

<https://github.com/mraza007/minute>

QUICK HITS

Codex Security HACKERNEWS

OpenAI's Codex Security project provides a set of tools and resources for securing AI models and data.

<https://github.com/openai/codex-security>

Show HN: XY – A Fast, composable, GPU-accelerated interactive plotting library HACKERNEWS

The XY library is a fast and composable plotting library that uses GPU acceleration for interactive visualizations.

<https://github.com/reflex-dev/xy>

Formally verified 3D CSG: Trust 93 lines spec, not 1000 lines AI code HACKERNEWS

The verified 3D CSG project provides a formally verified implementation of 3D constructive solid geometry operations.

<https://github.com/schildep/verified-3d-mesh-intersection>

baidu/Unlimited-OCR HUGGINGFACE

The Unlimited-OCR model is a trending model on Hugging Face that uses transformers for image-text-to-text tasks.

<https://huggingface.co/baidu/Unlimited-OCR>

upstage/Solar-Open2-250B HUGGINGFACE

The Solar-Open2-250B model is a trending model on Hugging Face that uses transformers for text generation tasks.

<https://huggingface.co/upstage/Solar-Open2-250B>

QwenAudio/qwen-audio-agent GITHUB

The QwenAudio agent is a real-time voice runtime that keeps agents talking, working, and present.

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

BrowserAct: Browser Layer for Your AI Agent HACKERNEWS

The BrowserAct project provides a browser layer for AI agents, allowing them to interact with web pages and perform tasks.

<https://github.com/browser-act/skills>

Lean4 Datalog DSL Based on Google Zanzibar for AI Projects HACKERNEWS

The Lean4 Datalog DSL project provides a domain-specific language for AI projects, based on Google's Zanzibar datalog language.

<https://github.com/jagg-ix/zil-lean>

Nanbeige/Nanbeige4.2-3B HUGGINGFACE

The Nanbeige4.2-3B model is a trending model on Hugging Face that uses transformers for text generation tasks.
<https://huggingface.co/Nanbeige/Nanbeige4.2-3B>

microsoft/Mage-Flow HUGGINGFACE

The Mage-Flow model is a trending model on Hugging Face that uses diffusers for text-to-image tasks.
<https://huggingface.co/microsoft/Mage-Flow>

Kwaipilot/KAT-Coder-V2.5-Dev HUGGINGFACE

The KAT-Coder-V2.5-Dev model is a trending model on Hugging Face that uses transformers for text generation tasks.
<https://huggingface.co/Kwaipilot/KAT-Coder-V2.5-Dev>

Zazza/svoi-oss GITHUB

The Svoi-OSS project provides a self-hosted, open-source personal AI assistant with streaming chat and agentic tools.
<https://github.com/Zazza/svoi-oss>

200lz/llm-engineering-platform GITHUB

The LLM engineering platform project provides a production-oriented platform for building and deploying LLMs.
<https://github.com/200lz/llm-engineering-platform>

Vladimir-Human/ru-marketplace-mcp GITHUB

The Ru-Marketplace-MCP project provides a set of MCP servers for Russian marketplaces, allowing for easy integration and comparison of prices.
<https://github.com/Vladimir-Human/ru-marketplace-mcp>

casioreview20-glitch/forgo-os GITHUB

The Forge-OS project provides an open-source control plane for AI agents, including skill routing, context governance, and trustworthy execution.
<https://github.com/casioreview20-glitch/forgo-os>

enowdev/succubus GITHUB

The Succubus project provides a cross-agent coordination system for AI coding agents, allowing them to share plans, tasks, and file claims.
<https://github.com/enowdev/succubus>

cadamcat/dual-radeon-vllm GITHUB

The Dual-Radeon-VLLM project provides a multi-GPU vLLM implementation on consumer Radeon hardware, including a fix for the RCCL crash and 292 benchmark measurements.
<https://github.com/cadamcat/dual-radeon-vllm>
