

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

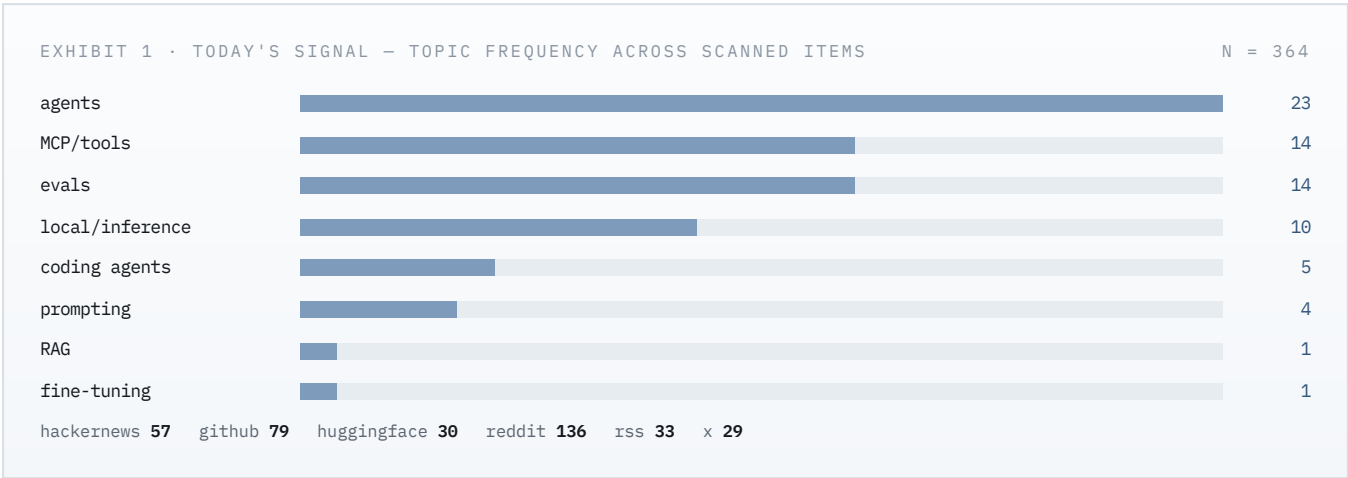
Today in AI engineering

2026-07-23

Developers release open-source AI tools and models for various applications

5 deep dives 2 builder stories 16 quick hits 364 items scanned

llm ai agents open-source transformers text-generation



DEEP DIVES

Millwright – Rust-based, self-hosted LLM router

hackernews

Northwood-Systems built Millwright, a self-hosted LLM router, to provide a cost-effective and transparent solution. Millwright is an open-source router that can be used with various LLM models. The developer was motivated by the potential acquisition of OpenRouter and the proliferation of hosted LLM routers. Millwright aims to provide a self-hosted alternative with better performance.

Why it matters. Millwright offers a self-hosted solution for LLM routing, providing users with more control over their data and costs. This is particularly important for applications where data privacy and security are crucial.

TRY THIS

Explore the Millwright repository and experiment with integrating it into your own projects to improve LLM routing efficiency.

<https://github.com/Northwood-Systems/millwright>

llm rust self-hosted

Cactus Hybrid: We taught Gemma 4 to know when it's wrong

hackernews

Cactus developed Cactus Hybrid, a system that teaches Gemma 4 to recognize when it's wrong. This is achieved through post-training Gemma 4 to provide confidence scores for its responses. The goal is to improve the reliability of small, on-device models by making them aware of their limitations.

Why it matters. Cactus Hybrid enhances the trustworthiness of AI models by providing a mechanism for them to indicate uncertainty, which is critical for real-world applications where accuracy and reliability are paramount.

TRY THIS

Investigate how confidence scoring can be applied to your own AI projects to improve their reliability and trustworthiness.

<https://github.com/cactus-compute/cactus-hybrid>

ai

confidence scoring

gemma 4

Agent Execution Partnership

github

Eli-labz created the Agent Execution Partnership, an open-source control plane designed to manage AI agent actions. It ensures that every action is authorized, observable, and verifiable, providing a robust framework for AI agent execution.

Why it matters. The Agent Execution Partnership offers a comprehensive solution for managing AI agents, ensuring that their actions are secure, transparent, and accountable, which is essential for complex and sensitive applications.

TRY THIS

Explore the Agent Execution Partnership repository to understand how it can be integrated into your AI projects for enhanced security and control.

<https://github.com/eli-labz/Agent-Execution-Partnership>

ai agents

control plane

security

Klaatcode

github

KlaatAI developed Klaatcode, an open-source AI coding agent for the terminal. It supports smart model routing, allowing it to use the right AI model for each task, thereby reducing costs. Klaatcode is compatible with several models, including Claude, GPT, and Gemini.

Why it matters. Klaatcode provides an efficient and cost-effective solution for coding tasks by leveraging the strengths of different AI models, making it a valuable tool for developers.

TRY THIS

Experiment with Klaatcode to see how it can streamline your coding workflow and reduce costs.

<https://github.com/KlaatAI/klaatcode>

ai coding

terminal

cost-effective

thinkingmachines/Inkling

huggingface

ThinkingMachines released Inkling, a trending model on Hugging Face. Inkling is an image-text-to-text model that utilizes transformers. It has gained significant attention with over 1465 likes and 16,441 downloads.

Why it matters. Inkling demonstrates the community's interest in multimodal models that can process and generate text based on images, showcasing the potential for applications in image description, generation, and understanding.

TRY THIS

Explore the Inkling model on Hugging Face to learn more about its capabilities and potential applications in your projects.

<https://huggingface.co/thinkingmachines/Inkling>

image-text

transformers

hugging face

Show HN: Agent in 9 Lines Python

hackernews

A minimal implementation of an AI agent in just 9 lines of Python, capable of tool calling and basic functionality.

How it works. The agent is designed to be simple and lightweight, focusing on essential features without unnecessary dependencies, making it easy to understand and modify.

STEAL THIS

Simplicity can be a powerful approach in AI development, allowing for quick prototyping and iteration.

<https://gist.github.com/tosh/6e91a9dbf08dd630c535e7345ac7f0b5>

Show HN: RunKit – a browser based tmux manager

hackernews

RunKit, a browser-based tmux manager designed to manage and run multiple AI agents in parallel, especially useful for remote access over Tailscale.

How it works. RunKit operates without a database, reading state live from tmux, and is the first tool the developer installs on new machines for managing AI agents.

STEAL THIS

Streamlining the management of AI agents can significantly improve productivity, especially when working with multiple agents or in distributed environments.

<https://github.com/sahil87/run-kit>

QUICK HITS

Bento - An entire PowerPoint in one HTML file HACKERNEWS

Bento allows for the creation of interactive slides with web frontend technologies, simplifying the editing process.

<https://bento.page/slides/>

livetennisapi-mcp GITHUB

An MCP server for providing real-time tennis scores and odds to AI agents like Claude and Cursor.

<https://github.com/livetennisapi/livetennisapi-mcp>

drumih/turbo-fieldfare GITHUB

Achieves Gemma 4 26B-A4B inference in approximately 2 GB of RAM on any M-series MacBook, showcasing efficient model deployment.

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

Octen-Team/search-eval GITHUB

A re-runnable web-search evaluation framework for assessing AI agent performance with in-house benchmarks and open-benchmark scoring.

<https://github.com/Octen-Team/search-eval>

ToolJet/ActionRail GITHUB

An open-source runtime action grounding framework for verifying proposed actions against live systems of record before execution, enhancing AI agent reliability.

<https://github.com/ToolJet/ActionRail>

drpwchen/textbook-to-note GITHUB

A tool for converting PDF textbooks into AI-searchable knowledge bases with structured, fully-cited notes, emphasizing local-first and token-frugal approaches.

<https://github.com/drpwchen/textbook-to-note>

Keksuccino/Better-Codex-App-Custom-Provider-Support GITHUB

Adds support for selecting custom model providers in the Codex app without requiring logout, enhancing flexibility and model choice for users.

<https://github.com/Keksuccino/Better-Codex-App-Custom-Provider-Support>

aka-kika/hig-mcp GITHUB

An MCP server serving Apple Human Interface Guidelines as structured design tokens for AI coding agents, facilitating consistent design practices.
<https://github.com/aka-kika/hig-mcp>

AliAkrami1375/Li-Translate GITHUB

An open-source AI-powered subtitle generation and natural language translation platform for video and audio content, supporting diverse language needs.
<https://github.com/AliAkrami1375/Li-Translate>

vakovalskii/qwen-native-agents GITHUB

A repository for building and prompting Qwen3.x agents in their native dialect, extracted from qwen-code and Qwen-Agent, focusing on native integration and performance.
<https://github.com/vakovalskii/qwen-native-agents>

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

A trending model on Hugging Face for text generation, leveraging llama.cpp and demonstrating community interest in efficient and powerful language models.
<https://huggingface.co/prism-ml/Ternary-Bonsai-27B-gguf>

upstage/Solar-Open2-250B HUGGINGFACE

A large-scale text generation model available on Hugging Face, indicating the ongoing development and sharing of significant AI models within the community.
<https://huggingface.co/upstage/Solar-Open2-250B>

zai-org/GLM-5.2 HUGGINGFACE

A widely-used text generation model on Hugging Face, with over 4345 likes and 545,109 downloads, showcasing its popularity and applicability in various projects.
<https://huggingface.co/zai-org/GLM-5.2>

sudoingX/dgx-spark-laguna GITHUB

A working stack for serving Poolside Laguna S 2.1 on the NVIDIA DGX Spark without system hangs, including crash-safe configurations and benchmarks.
<https://github.com/sudoingX/dgx-spark-laguna>

EndymionLee/PilotBrowseMCP GITHUB

A browser runtime allowing AI agents to control a real Chrome browser via MCP, enabling agents to explore websites, generate manuals, and reuse them efficiently.
<https://github.com/EndymionLee/PilotBrowseMCP>

Realynx/PolySIEM GITHUB

An open-source homelab SIEM solution for mapping, logging, and monitoring lab environments, integrating Proxmox, OPNsense, Elastic, and Suricata for comprehensive security.
<https://github.com/Realynx/PolySIEM>
