

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

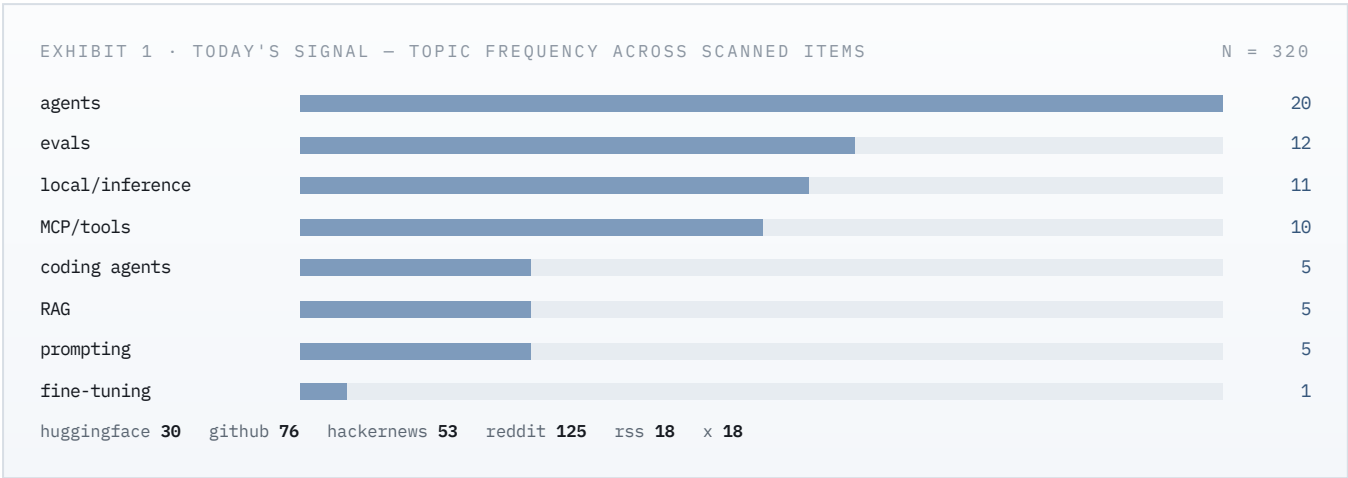
Today in AI engineering

2026-08-09

Developers build and share AI agent frameworks, LLM models, and automation tools

5 deep dives 2 builder stories 14 quick hits 320 items scanned

llm ai agents automation rust transformers



DEEP DIVES

MiniMaxAI/MiniMax-H3

huggingface

The MiniMaxAI team has released MiniMax-H3, a model that generates videos from text prompts. This model is available on the Hugging Face model hub and has gained significant attention with over 26,000 downloads. The model uses diffusion-based image synthesis to generate videos.

Why it matters. MiniMax-H3 demonstrates the potential of AI in video generation, which can be applied to various fields such as entertainment, education, and advertising.

TRY THIS

Experiment with the MiniMax-H3 model by fine-tuning it on a custom dataset to generate videos for a specific use case.

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

llm video generation diffusion

i3T4AN/KADATH

[github](#)

The KADATH project is an evolutionary multi-agent runtime that breeds, evaluates, and improves autonomous agents. This open-source project provides a framework for developing and testing AI agents in a reproducible manner.

Why it matters. KADATH offers a unique approach to AI agent development, enabling the creation of more sophisticated and adaptive agents.

TRY THIS

Explore the KADATH repository and try running the provided examples to understand how the framework works.

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

[ai agents](#)[evolutionary algorithms](#)[multi-agent systems](#)

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

[github](#)

The DeepSeek-V4-Flash-0731 project provides a reproducible recipe for training and deploying LLM models on NVIDIA RTX PRO 6000 GPUs. The repository includes pre-built images and benchmarking tools.

Why it matters. This project enables developers to easily train and deploy high-performance LLM models on specific hardware configurations.

TRY THIS

Use the provided recipe to train and deploy an LLM model on an RTX PRO 6000 GPU and experiment with different hyperparameters.

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

[llm](#)[gpu acceleration](#)[reproducibility](#)

antonellof/ferrox

[github](#)

Ferrox is a Rust-based GGUF inference engine that supports quantized CPU, Metal, and CUDA kernels. The project provides a benchmarked head-to-head comparison against llama.cpp.

Why it matters. Ferrox offers a high-performance inference engine for LLM models, which can be used in various applications such as natural language processing and computer vision.

TRY THIS

Experiment with Ferrox by integrating it into a project that requires high-performance LLM inference.

<https://github.com/antonellof/ferrox>

[llm](#)[inference engine](#)[rust](#)

agent-hop

[hackernews](#)

Agent-hop is a tool that allows users to search and resume chats across multiple coding agents. The project provides a solution to the problem of managing multiple agent sessions.

Why it matters. Agent-hop demonstrates the potential of AI in improving developer productivity and workflow management.

TRY THIS

Try using agent-hop to manage multiple agent sessions and explore its features.

<https://github.com/hetpatel-11/agent-hop>

[ai agents](#)[productivity](#)[workflow management](#)

Show HN: agent-hop – reverse-engineered session formats to resume in any agent

hackernews

A tool that allows users to search and resume chats across multiple coding agents.

How it works. The tool uses reverse-engineered session formats to enable seamless switching between agents.

STEAL THIS

By leveraging AI and workflow management, developers can improve their productivity and efficiency.

<https://github.com/hetpatel-11/agent-hop>

Show HN: 49IDE – 2D Grid IDE for managing many agents, Git trees, issues

hackernews

A 2D Grid IDE for managing multiple agents, Git trees, and issues.

How it works. The IDE uses a grid-based interface to visualize and manage complex workflows and agent interactions.

STEAL THIS

Visualizing complex workflows and agent interactions can help developers better understand and manage their systems.

<https://github.com/alpbahadur/49Agents>

QUICK HITS

deepseek-ai/DeepSeek-V4-Flash-0731 HUGGINGFACE

A trending LLM model on Hugging Face with over 785,000 downloads.

<https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731>

moonshotai/Kimi-K3 HUGGINGFACE

A trending model on HuggingFace with over 1.3 million downloads.

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

TOPDEV99999/AI-Knowledge-Management-Platform GITHUB

An open-source AI knowledge management platform that enhances RAG capabilities.

<https://github.com/TOPDEV99999/AI-Knowledge-Management-Platform>

rexblade58/agenteval GITHUB

An open-source agent evaluation framework for scoring AI agents across providers.

<https://github.com/rexblade58/agenteval>

SergioB/intel-arc-pro-b70-inference-cookbook GITHUB

A repository providing open recipes and benchmark harnesses for LLM inference on Intel Arc Pro B60/B70.

<https://github.com/SergioB/intel-arc-pro-b70-inference-cookbook>

YINGLINGH/limioryn GITHUB

A high-level edge-cloud AI multi-agent framework for real devices and verifiable actuation.

<https://github.com/YINGLINGH/limioryn>

malwarejake/CUSTODY-framework GITHUB

The CUSTODY framework for AI agent containment.

<https://github.com/malwarejake/CUSTODY-framework>

Ed-Marcavage/awesome-security-agent-harnesses GITHUB

A collection of AI agents for pentesting, code audit, fuzzing, vulnerability discovery, and reverse engineering.

<https://github.com/Ed-Marcavage/awesome-security-agent-harnesses>

drmikecrypto/WebSearchFree GITHUB

A free open-source alternative to Tavily for keyless web search and extract for AI agents.

<https://github.com/drmikecrypto/WebSearchFree>

davidahmann/fde-guide GITHUB

A field guide for FDEs and applied AI teams, covering value engineering and production architecture.

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

Sparkfetch/sparkfetch GITHUB

An open-source web fetching and extraction API for turning any URL into clean, structured, LLM-ready content.

<https://github.com/Sparkfetch/sparkfetch>

Gen-Verse/Skill-Entropy-RL GITHUB

A project focused on skill-native LLMs, using skill entropy for benchmarking and training long-horizon reasoning.

<https://github.com/Gen-Verse/Skill-Entropy-RL>

samurdhilbk/gaas GITHUB

An API for AI spinner verbs, providing a service for generating text based on given prompts.

<https://github.com/samurdhilbk/gaas>

wanmol/goal-flow GITHUB

A production-grade framework for combining workflow graphs and agent loops, supporting LangGraph and Dify DSL.

<https://github.com/wanmol/goal-flow>
