

DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

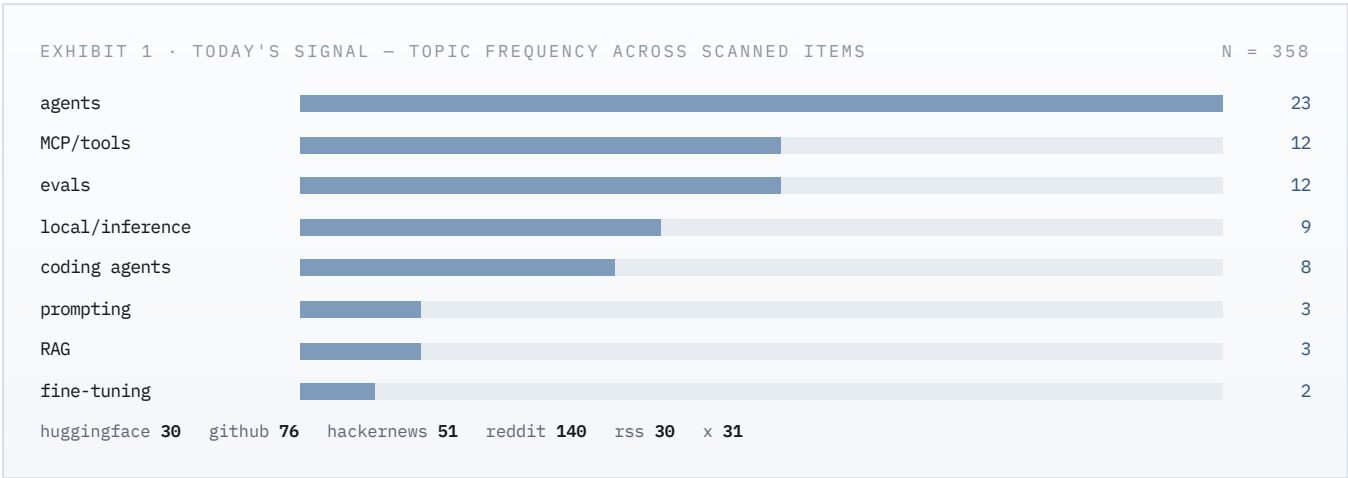
Today in AI engineering

2026-08-15

AI engineers are building and sharing tools for agent safety, LLM optimization, and automation workflows

3 deep dives 2 builder stories 16 quick hits 358 items scanned

- agent safety
- llm optimization
- automation workflows
- ai engineering
- open-source tools
- hugging face models



DEEP DIVES

decionis/agent-safe-pipeline

github

Decionis built an open-source pipeline for AI agents that propose actions but cannot authorize them. The pipeline includes immutable intent capture, an independent Decionis policy verdict, verified human approval, and a SafeExecutor that consumes a single-use intent-bound grant. This pipeline is designed to ensure safe and secure interactions between AI agents and humans. The developer's goal is to provide a reference architecture for building safe AI systems.

Why it matters. This pipeline is important because it provides a framework for building safe AI systems that can interact with humans without compromising security or safety. By using this pipeline, developers can ensure that their AI agents are designed with safety and security in mind from the outset.

TRY THIS

Try implementing the agent-safe-pipeline in your own AI project to see how it can improve safety and security

<https://github.com/decionis/agent-safe-pipeline>

- agent safety
- ai engineering
- open-source tools

Leutenegger/book-to-skill

github

Leutenegger built an open-source tool that turns any technical book PDF into a Claude Code skill. This tool is designed to make it easy for developers to create new skills for Claude Code based on existing technical books. The tool is ready to study, reference, and use while working.

Why it matters. This tool is important because it provides a way for developers to create new skills for Claude Code based on existing technical books. By using this tool, developers can expand the capabilities of Claude Code and make it more useful for a wider range of tasks.

TRY THIS

Try using the book-to-skill tool to create a new skill for Claude Code based on a technical book you've been wanting to learn from

<https://github.com/Leutenegger/book-to-skill>

llm optimization

ai engineering

open-source tools

Qwen/Qwen3.8-27B

huggingface

Qwen built a trending model on Hugging Face called Qwen3.8-27B. This model is an image-text-to-text transformer that has gained popularity among developers. The model has 9238 likes and 2 downloads, indicating its potential for use in various applications.

Why it matters. This model is important because it provides a pre-trained model that developers can use for image-text-to-text tasks. By using this model, developers can save time and resources that would be required to train a model from scratch.

TRY THIS

Try using the Qwen3.8-27B model in your own project to see how it can improve performance on image-text-to-text tasks

<https://huggingface.co/Qwen/Qwen3.8-27B>

hugging face models

ai engineering

image-text-to-text

WHAT PEOPLE ACTUALLY BUILT

Show HN: Mole – Deep research agent for your terminal

hackernews

Lajosdeme built a deep research agent called Mole that can be used in the terminal. Mole is designed to help with research tasks and can provide more accurate results than traditional search methods.

How it works. Mole uses a combination of natural language processing and machine learning algorithms to provide relevant results for research queries. The agent can be used to search for information on a wide range of topics and can provide suggestions for further research.

STEAL THIS

The key takeaway from this project is the importance of using AI agents to improve research workflows and provide more accurate results.

<https://github.com/lajosdeme/mole>

Show HN: Shoehorn, a library to quantize an LLM to fit your Mac's VRAM

hackernews

Notactuallytreyanastasio built a library called Shoehorn that can be used to quantize a large language model (LLM) to fit on a Mac's VRAM. This library is designed to make it possible to run LLMs on devices with limited memory.

How it works. Shoehorn uses a combination of quantization and pruning techniques to reduce the size of the LLM while maintaining its performance. The library provides a simple API for quantizing LLMs and can be used with a variety of models.

STEAL THIS

The key takeaway from this project is the importance of optimizing LLMs for use on devices with limited memory, and the potential for quantization and pruning techniques to achieve this goal.

<https://github.com/notactuallytreyanastasio/shoehorn>

QUICK HITS

OpenLabs-so/openanalytics GITHUB

OpenLabs-so built an open-source, privacy-first web analytics tool with revenue attribution and an MCP server.
<https://github.com/OpenLabs-so/openanalytics>

meta-models/Muse-Glimmer-30B HUGGINGFACE

Meta-models built a trending model on Hugging Face called Muse-Glimmer-30B, an image-text-to-text transformer.
<https://huggingface.co/meta-models/Muse-Glimmer-30B>

yangzc/ainote GITHUB

Yangzc built an AI agent workflow platform with visual flow orchestration, drag-and-drop forms, and knowledge base RAG.
<https://github.com/yangzc/ainote>

Lightricks/LTX-2.5 HUGGINGFACE

Lightricks built a trending model on Hugging Face called LTX-2.5, an image-to-video diffusion model.
<https://huggingface.co/Lightricks/LTX-2.5>

MiniMaxAI/MiniMax-H3 HUGGINGFACE

MiniMaxAI built a trending model on Hugging Face called MiniMax-H3, an image-text-to-video minimax-h3 model.
<https://huggingface.co/MiniMaxAI/MiniMax-H3>

shixinnt/codex-image-context-runtime GITHUB

Shixinnt built a Codex plugin and local MCP runtime for context-bounded image generation and inspection.
<https://github.com/shixinnt/codex-image-context-runtime>

smshahbaj/crucible GITHUB

Smshahbaj built a production-ready Claude Code decision intelligence marketplace with adversarial routing and automated validation.
<https://github.com/smeshahbaj/crucible>

z15114664687-dot/hr-onboarding-agent GITHUB

Z15114664687-dot built an open-source AI-assisted HR onboarding tool with configurable workflows and document OCR.
<https://github.com/z15114664687-dot/hr-onboarding-agent>

vpromise/inside-coding-agents GITHUB

Vpromise built a visual, evidence-backed handbook and lab for coding agent architecture and design.
<https://github.com/vpromise/inside-coding-agents>

raullenchai/twinspace GITHUB

Raullenchai built a production-grade recipe for DeepSeek-V4-Flash-0731 on 2x NVIDIA DGX Spark with self-healing and reboot-verified capabilities.
<https://github.com/raullenchai/twinspace>

Riccardo8888/agent-link GITHUB

Riccardo8888 built an end-to-end encrypted channel between coding agents, anywhere.

<https://github.com/Riccardo8888/agent-link>

memoket/memoket-kite GITHUB

Memoket built a memory engine for AI agents with token-efficient, explainable retrieval beyond vector similarity.

<https://github.com/memoket/memoket-kite>

bar181/bar-observatory GITHUB

Bar181 built a deterministic, local-only audit report tool for Claude Code AI agent sessions with Rust and SQLite.

<https://github.com/bar181/bar-observatory>

dnv2003/v100-skinny GITHUB

Dnv2003 built hand-written NVFP4 W4A16 CUDA kernels and chain-MTP speculative serving for Qwen3.6-27B on four Tesla V100s.

<https://github.com/dnv2003/v100-skinny>

EmilResearch/RAGless HACKERNEWS

EmilResearch built a tool similar to RAG, but with \$0 LLM API costs at runtime.

<https://github.com/EmilResearch/RAGless>

choiyounggi/cliclaw HACKERNEWS

Choiyounggi built a tool to control Claude Code, Codex, Pi, and Gemini CLI from Telegram.

<https://github.com/choiyounggi/cliclaw>
