DAILY BRIEF · WHAT AI ENGINEERS ARE BUILDING

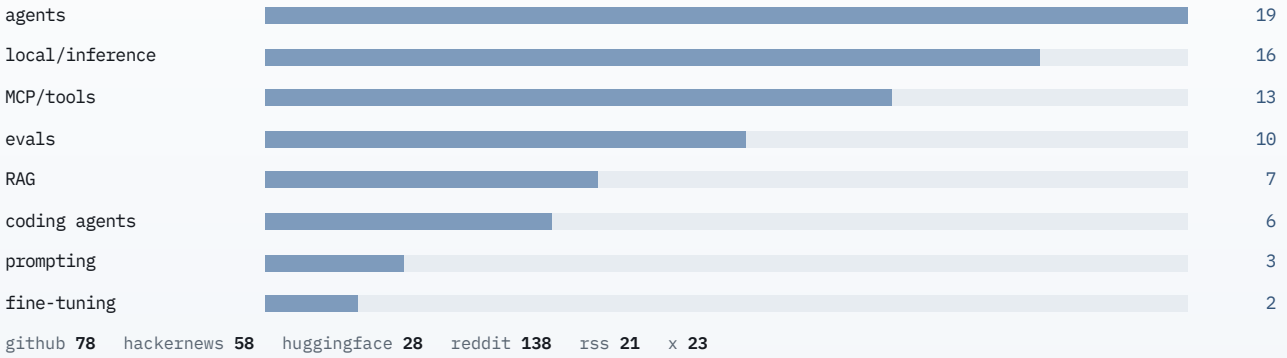
Today in AI engineering 2026-08-04

Advances in LLM inference, optimization, and automation continue to drive innovation in AI engineering

3 deep dives **2** builder stories **10** quick hits **346** items scanned

llm inference | ai optimization | model context protocol (mcp) | automation | graphrag | rag

EXHIBIT 1 · TODAY'S SIGNAL — TOPIC FREQUENCY ACROSS SCANNED ITEMS N = 346



github **78** hackernews **58** huggingface **28** reddit **138** rss **21** x **23**

DEEP DIVES

FareedKhan-dev/kimi-k3-in-c

github

Fareed Khan has implemented a 2.78-trillion-parameter Kimi K3 model in C, achieving inference on a single CPU in 8.24 GB of RAM. This project demonstrates the potential for efficient LLM deployment on resource-constrained devices. The implementation is notable for its portability, requiring no BLAS, framework, or GPU.

Why it matters. This project showcases the feasibility of running large LLMs on limited hardware, which is crucial for edge AI applications. It also highlights the importance of optimizing LLMs for inference efficiency.

> **TRY THIS**
> Explore the Kimi K3 implementation and experiment with optimizing other LLMs for resource-constrained devices.

https://github.com/FareedKhan-dev/kimi-k3-in-c

llm inference | efficient deployment | edge ai

patchy631/time-to-first-token

github

Patchy631 has created a 10-week roadmap for LLM inference serving and optimization. The project covers topics such as vLLM, SGLang, quantization, speculative decoding, and benchmarking. This resource provides a structured approach to improving LLM performance and efficiency.

Why it matters. Optimizing LLMs is essential for real-world applications, and this roadmap offers a comprehensive guide for achieving better performance and efficiency. By following this roadmap, developers can significantly improve their LLM-based projects.

> **TRY THIS**
> Follow the roadmap and apply the optimization techniques to your own LLM projects.

https://github.com/patchy631/time-to-first-token

llm optimization | inference serving | benchmarking

NeelM0906/Mference

github

NeelM0906 has developed a Swift and Metal-based MoE inference engine for Apple Silicon devices. The project achieves impressive performance, with Gemma 4 26B running in ~2 GB and Qwen 3.6 35B in ~1.45 GB. This engine also supports SSD-streamed experts and includes a native Mac app, CLI, and OpenAI-compatible server.

Why it matters. This project demonstrates the potential for efficient LLM inference on Apple Silicon devices, which is crucial for edge AI applications. The engine's performance and features make it an attractive solution for developers.

> **TRY THIS**
> Explore the Mference engine and experiment with integrating it into your own projects.

https://github.com/NeelM0906/Mference

llm inference | apple silicon | moe

WHAT PEOPLE ACTUALLY BUILT

Show HN: Runthru – open-source Interactive Demos

hackernews

Mark Tolson created Runthru, an open-source tool for creating interactive demos and walkthroughs for web-based software. The tool allows users to point it at a site and either let AI generate the demo or create it manually.

How it works. Runthru uses a combination of AI and user input to generate interactive demos. The tool is designed to be flexible and adaptable to different use cases.

> **STEAL THIS**
> Developers can use Runthru to create engaging and interactive demos for their web-based software, improving user experience and adoption.

https://github.com/marktolson/runthru

Show HN: Texpile, an open-source desktop LaTeX editor with a visual mode

hackernews

The Texpile team created an open-source desktop LaTeX editor with a visual mode, aiming to provide a more user-friendly experience for writing documents in LaTeX.

How it works. Texpile combines the power of LaTeX with a visual editor, allowing users to create and edit documents more efficiently. The tool also supports real-time collaboration and offline use.

> **STEAL THIS**
> Developers can use Texpile to create and edit LaTeX documents more efficiently, and the tool's visual mode makes it more accessible to non-technical users.

https://github.com/texpile/texpile

QUICK HITS

AirLLM 70B inference with single 4GB GPU HACKERNEWS

AirLLM achieves impressive performance with a single 4GB GPU, demonstrating the potential for efficient LLM inference on resource-constrained devices.
https://github.com/lyogavin/airllm

moonshotai/Kimi-K3 HUGGINGFACE

The Kimi-K3 model is trending on Hugging Face, with 9900 likes and 967,622 downloads, indicating its popularity and potential for real-world applications.
https://huggingface.co/moonshotai/Kimi-K3

bybit-exchange/kaas GITHUB

Kaas is an LLM knowledge-base compiler that turns scattered notes, docs, and transcripts into a queryable Markdown wiki, providing a valuable tool for knowledge management and retrieval.
https://github.com/bybit-exchange/kaas

lordx64/pentestkit GITHUB

Pentestkit is an autonomous multi-agent pentest framework that achieves impressive results, with 104/104 (100%) on the XBOW validation benchmarks, powered by Kimi K3.
https://github.com/lordx64/pentestkit

xcosmosbox/Cairn GITHUB

Cairn is an open-source GraphRAG domain knowledge layer that continuously distills scattered skill docs into a queryable, evolvable, human-editable two-layer knowledge graph.
https://github.com/xcosmosbox/Cairn

dinosn/security-research-orchestrator-prompt GITHUB

This project provides a high-assurance, artifact-agnostic prompt for authorized security labs, exploit-chain research, PoC validation, and evidence-led reporting.
https://github.com/dinosn/security-research-orchestrator-prompt

p0nymc1/cee GITHUB

Cee is a deterministic-first execution engine for agent workflows in Go, providing a valuable tool for building reliable and efficient AI systems.
https://github.com/p0nymc1/cee

Funluned/vetresearch-workbench GITHUB

This project provides an evidence-grounded veterinary research workbench with local RAG, bounded LLM agents, auditable tool use, citations, abstention, and human review.
https://github.com/Funluned/vetresearch-workbench

ryan-phq2005h1/github-mcp-server GITHUB

The GitHub MCP Server connects AI agents to the GitHub API, enabling seamless integration and automation of development workflows.
https://github.com/ryan-phq2005h1/github-mcp-server

Wan2.2 14B Fast Preview [NEW] HUGGINGFACE

The Wan2.2 14B model is a trending Space on Hugging Face, with 458 likes, indicating its potential for real-world applications and the interest in large language models.

https://huggingface.co/spaces/cinderholm/wan2-2-i2v-v3

Generated 2026-08-04 · Sources: Hacker News, Reddit, GitHub, Hugging Face, arXiv, engineering RSS, X/Twitter. Filtered for real, implementable work; marketing and cash-grabs removed. Built by the ai-daily-brief automation · LUCA CONARROE — AI Automation.