

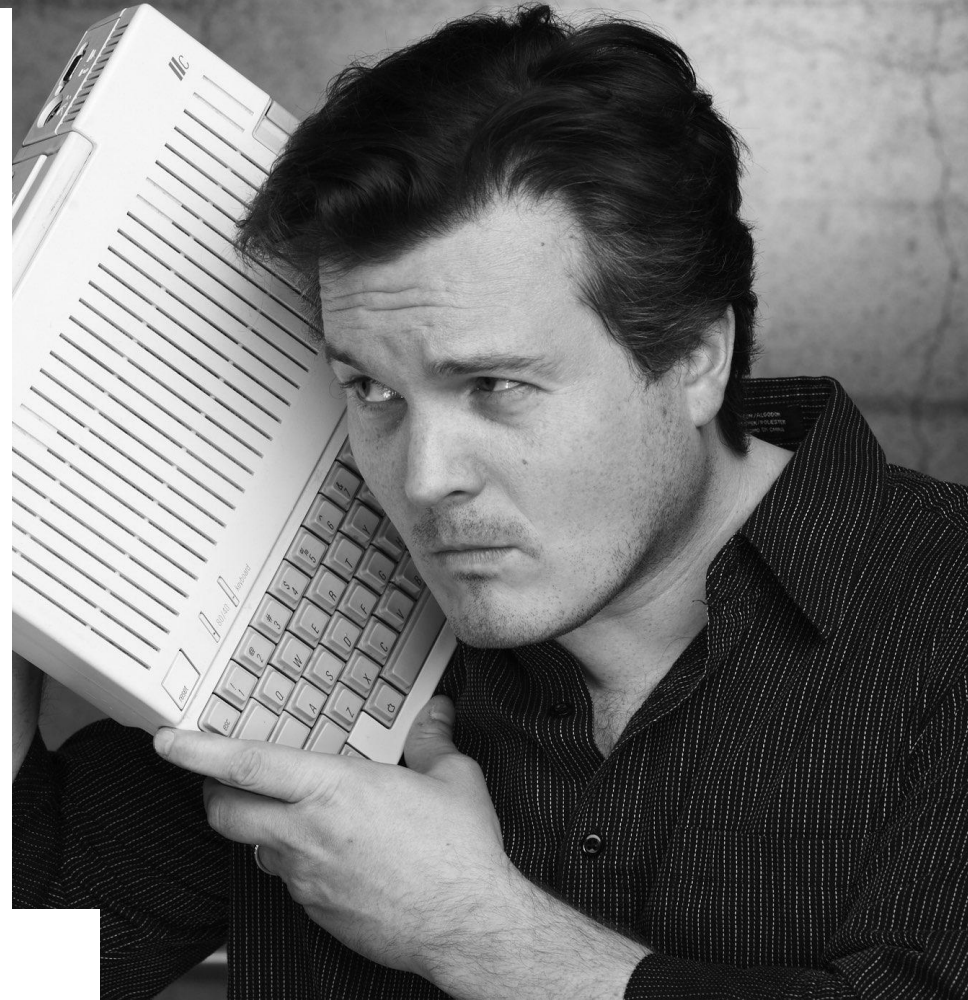
Feb 2, 2025

# Mozilla Builders

# Stephen Hood

Open Source AI Lead

[shood@mozilla.com](mailto:shood@mozilla.com)



# Agenda

- 01 What & why
- 02 Builders projects
- 03 What's next
- 04 Let's build together

**BRACE YOURSELF**

**QR CODES ARE COMING**

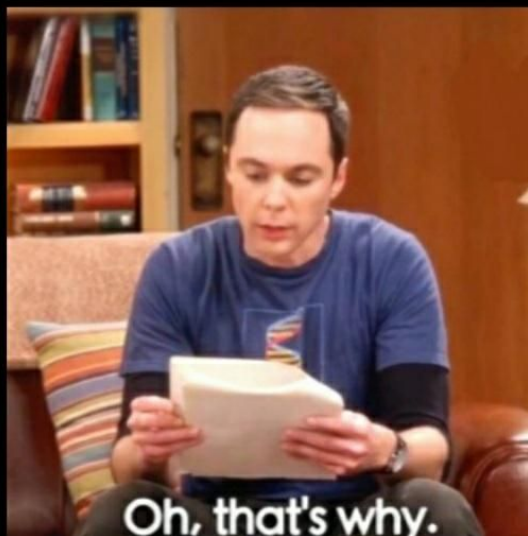
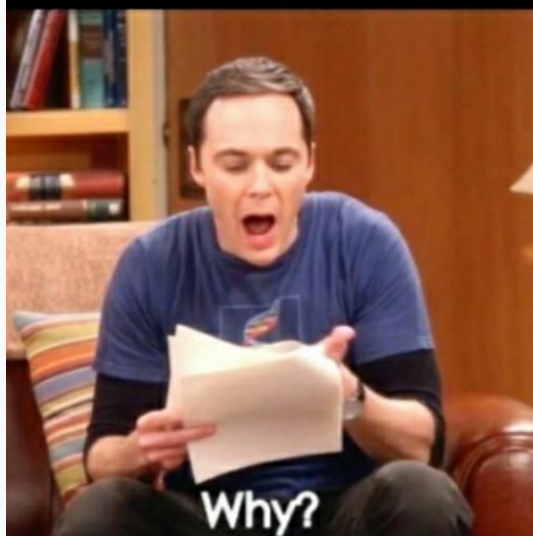
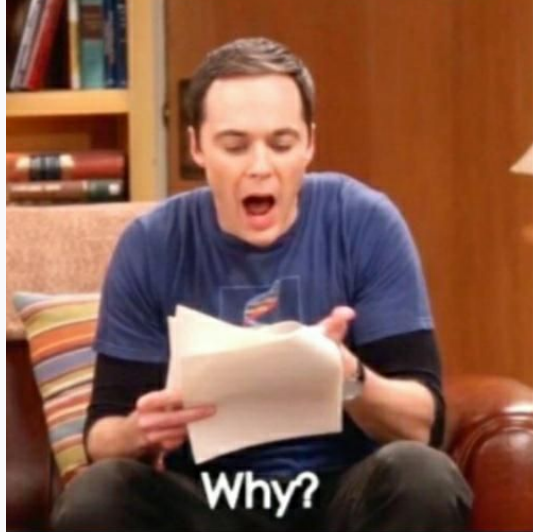
imgflip.com

# Mozilla Builders

sponsors and co-develops  
AI projects that help open  
source compete with  
closed offerings.

[builders.mozilla.org](https://builders.mozilla.org)





**We've been  
here before.**





**It's happening  
again.**





# Things you can already thank open source for:

- **Open models** that you can **run locally**.
- Running models without an expensive GPU.
- Building AI apps that don't depend upon **closed platforms**.
- Using AI without having to share **your data** with 3rd parties.
- Rapid **collaboration** and cross-pollination across public projects.

# Open source AI needs to be even simpler.

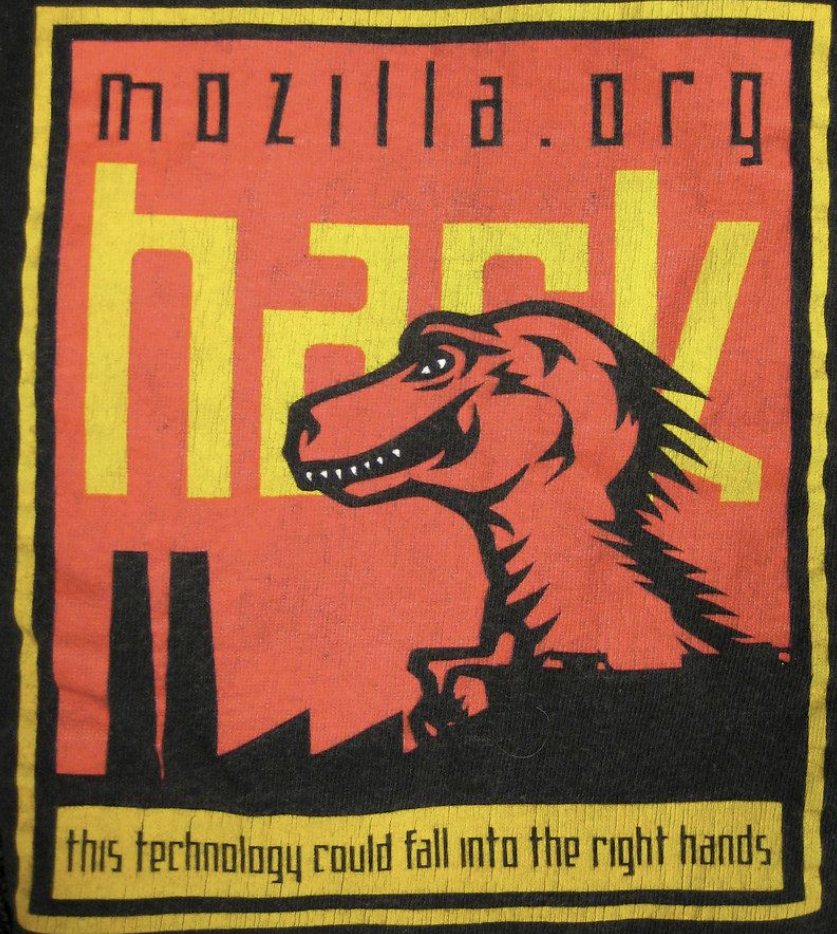


**Open source AI**



**OpenAI**

**We can't do  
this alone.**



# 2024 theme: Local AI

AI applications running “at the edge” on user devices, giving users more control and privacy.

- Developer productivity
- Agents
- Generative UI
- RAG
- Fine-tuning
- Function calling
- Evaluation

# Accelerator

A 12-week program designing, building, and testing 14 high-impact open source AI projects.

Inaugural Theme

**Local AI**

## Key Dates

Jun 3, 2024

Applications Open

Jul 8, 2024

Early Application Deadline

Aug 1, 2024

Final Application Deadline

Sep 12, 2024

Accelerator Kick Off

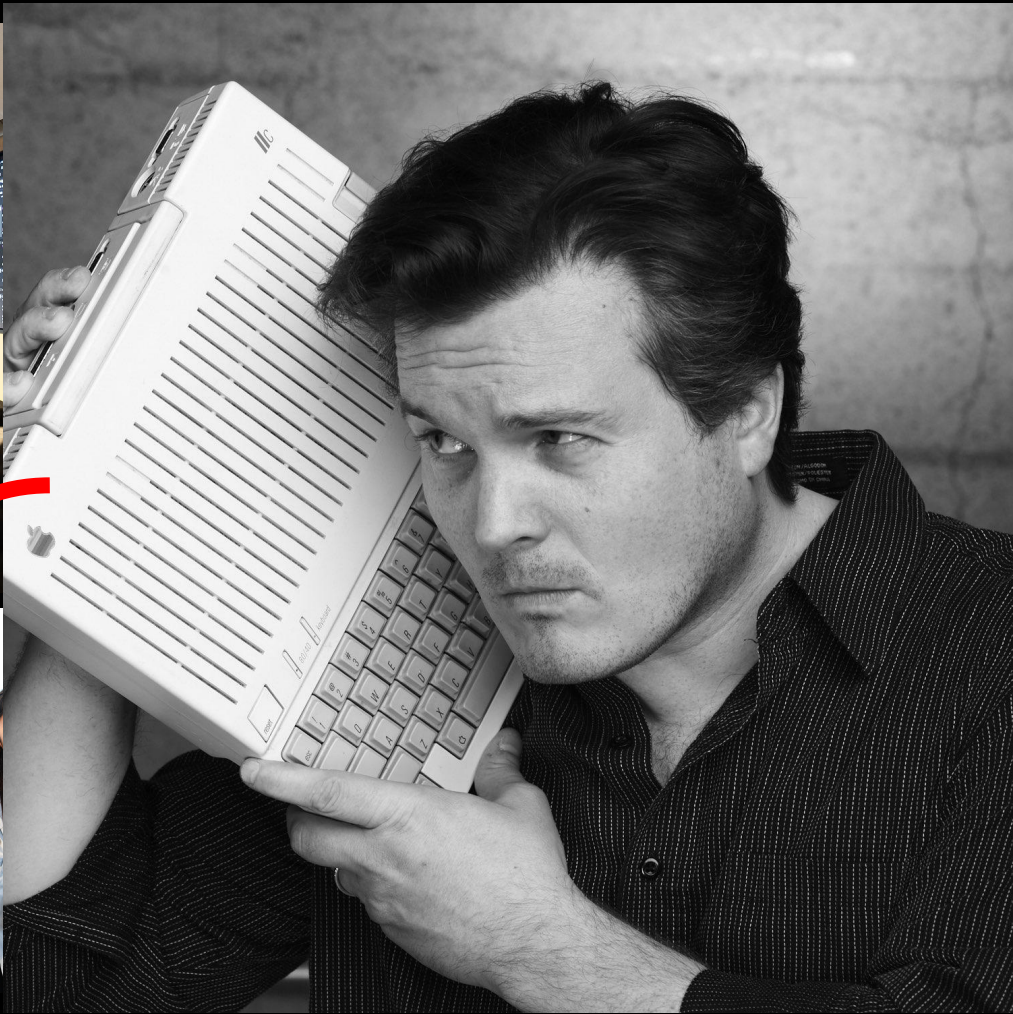
Dec 5, 2024

Demo Day

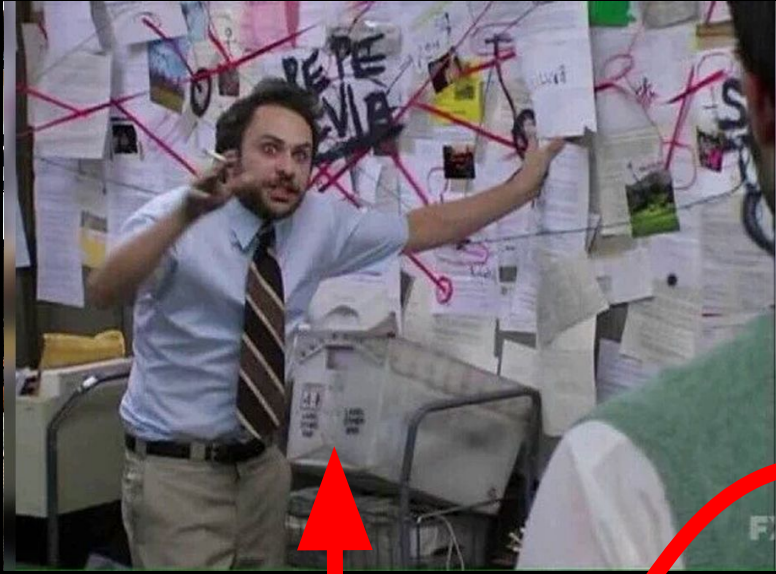




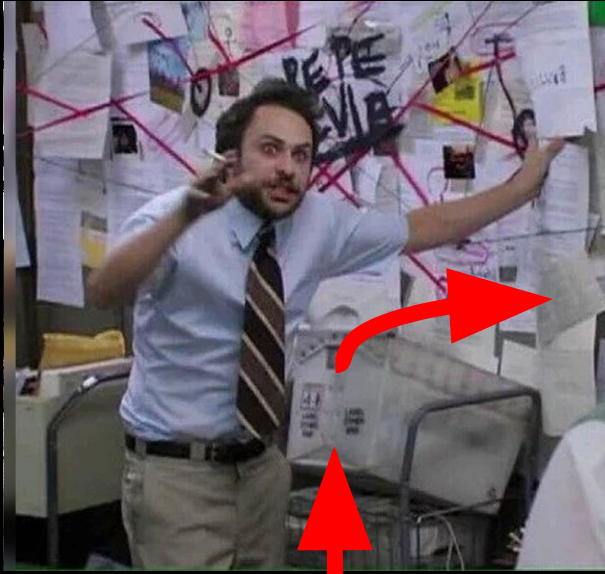














# Transformer Lab

An open-source platform that allows anyone to build, tune, & run LLMs locally, without writing code.

## *Impact*

- 43% increase in GitHub project stars
- Doubled Discord server membership



# PLEIAS

## Pleias

Fully truly open models and data sets for European languages.

### *Impact*

- Released Common Corpus, a 2+ Trillion token data set
- Released Pleias-pico, an SLM for multilingual RAG



Mozilla Builders



Helping users access high-quality scientific data in multiple languages. ScholarAI uses open-source AI models to process open licensed scientific documents. This approach ensures accessibility of top-tier research while remaining private and reducing the need for remote servers. [@NASTASIASTASENKO](#) [PIERREGARLANGAIS](#)



# THEIA AI IDE

# Theia

An open source AI-powered IDE with full developer control.

## *Impact*

- Launched first version of AI copilot capabilities.
- Integrated with Llamafire.



Mozilla Builders



Offering full control over AI interactions, including customizable prompts and the freedom to choose any language models, the AI-powered Theia IDE provides a suite of AI agents for developers on the desktop and the browser. Built on the Theia AI Framework, it allows users to tailor AI-driven features to their workflows with full flexibility and privacy, whether using on-device or cloud AI solutions. <https://www.mozilla.org/en-US/ai/>

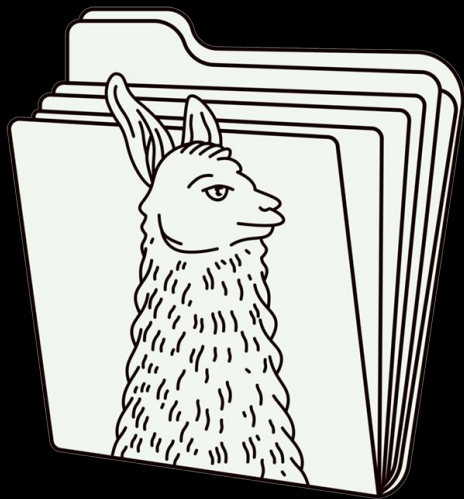




# Collaborations

Working one-on-one with builders on impactful OSS efforts.

- Independent open source projects
- Academic research teams
- Either sponsorship or co-development
- Variable funding and flexible terms



# Llamafile

Turns open models into single-file executables that run anywhere without installation.

*Goal:* democratize AI so that anyone can use open source AI as easily as using commercial offerings.



Justine Tunney





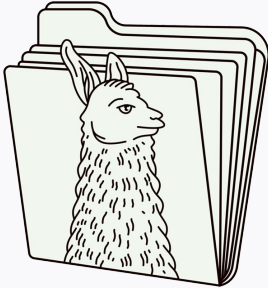
# The Reese's Peanut Butter Cup of Open Source AI



+



=



# Batteries included

**User:** Why is CPU inference speed important in edge computing?

**Llamafile:** Hello there! I'd be happy to help explain the importance of CPU inference speed in edge computing.

In edge computing, data processing occurs closer to where it is generated, rather than being sent to a cloud server for analysis. This approach can offer several benefits such as reduced latency and improved privacy. However, since edge devices typically have limited computational resources compared to cloud servers, optimizing CPU inference speed becomes crucial.

CPU inference speed refers to how quickly the device's central processing unit (CPU) can process machine learning models or artificial intelligence algorithms. The faster the CPU inference speed, the more real-time and responsive the edge computing application will be. This is particularly important for applications that require immediate feedback or decision making based on data analysis, such as autonomous vehicles, industrial IoT systems, or smart homes.

Therefore, optimizing CPU inference speed is essential for edge computing applications that require real-time processing of large amounts of data or complex machine learning models. This can involve techniques such as model quantization, pruning, or using specialized hardware like GPUs or TPUs to accelerate the inference process.

Say something...

Send Upload Image Stop Reset

336 tokens predicted, 53 ms per token, 18.88 tokens per second  
prompt evaluation speed is 0.00 prompt tokens evaluated per second  
powered by [llama.cpp](#), [ggml.ai](#), and [llamafile](#)

## Web & Console UI

```
...ape-1.10
ggml_metal_init: simdgroup reduction support = true
ggml_metal_init: simdgroup matrix mul. support = true
ggml_metal_init: hasUnifiedMemory = true
ggml_metal_init: recommendedMaxWorkingSetSize = 17179.89 MB
ggml_backend_metal_buffer_type_alloc_buffer: allocated buffer, size = 64.00 MiB, ( 3913.44 / 16384.02)
llama_kv_cache_init: Metal KV buffer size = 64.00 MiB
llama_new_context_with_model: KV self size = 64.00 MiB, K (f16): 32.00 MiB, V (f16): 32.00 MiB
llama_new_context_with_model: CPU output buffer size = 0.14 MiB
ggml_backend_metal_buffer_type_alloc_buffer: allocated buffer, size = 81.02 MiB, ( 3994.45 / 16384.02)
llama_new_context_with_model: Metal compute buffer size = 81.00 MiB
llama_new_context_with_model: CPU compute buffer size = 9.01 MiB
llama_new_context_with_model: graph nodes = 1030
llama_new_context_with_model: graph splits = 2
{"function":"initialize","level":"INFO","line":485,"msg":"initializing slots","n_slots":1,"tid":"1099515688016","timestamp":1719092605}
{"function":"initialize","level":"INFO","line":494,"msg":"new slot","n_ctx_slot":512,"slot_id":0,"tid":"1099515688016","timestamp":1719092605}
{"function":"server_cli","level":"INFO","line":3080,"msg":"model loaded","tid":"1099515688016","timestamp":1719092605}
llama server listening at http://127.0.0.1:8080
```

## OpenAI-compatible API

```
./llava-v1.5-7b-g4.llamafile --cli \
--image lemurs.jpg \
--temp 0 \
--grammar 'root ::= [a-z]+ -n 16
\
-e -p '### User: The image \
has...\n### Assistant:'
\
--silent-prompt 2>/dev/null |
sed -n 1s/ / /g; s/ /$ /g; s/ /$ /g;
```

## Scriptable CLI

# Summarizing Dijkstra w/ Mistral 7B on Threadripper

```
main jart@luna:~/llamafile$ /mnt/videos/llamafile-0.6.2/bin/llamafile -m main jart@luna:~/llamafile$ llamafile -m /weights/Mistral-7B-Instruct-v0.3.F16.gguf -f ~/prompt1625.txt -n 40 -c 03.BF16.gguf -f ~/prompt1625.txt -c 0 -b 256  
-b 256
```

# What's next for Llamafile

## 1. Llamafiler



A new “clean room” API server with improved performance, stability, and functionality.

## 2. Developer enablement



Making Llamafile the go-to solution for building local AI-powered apps.

## 3. New governance

Transitioning from a single primary maintainer to multiple maintainers.



# SQLite-vec

Bringing high-efficiency vector search to SQLite.

*Goal:* Enable developers to build RAG apps powered by local AI models.



Alex Garcia





# Web Applets

An open spec for building software that both humans and AI can understand and use together.

*Goal:* Ensure that the Web doesn't get shut out of a new wave of apps.



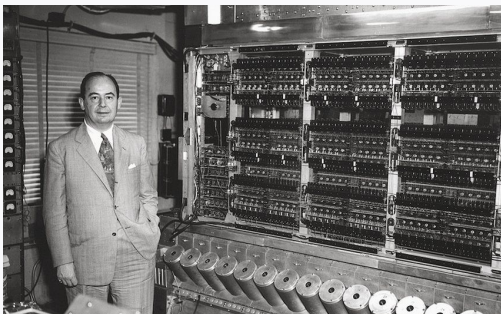
Rupert Manfredi



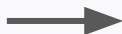
# **2025 theme: The next user agent**



# Computing is changing



**1949:** EDVAC and von Neumann architecture



**2024:** OpenAI multimodal agent demo

**75 years:** *computers can't understand us*

**Now:** *they can*

**Applications**

**Programming  
languages**

**Operating  
systems**



**Browsers**

**Applications**

**Programming  
languages**

**Operating  
systems**



# 2025 theme: The next user agent

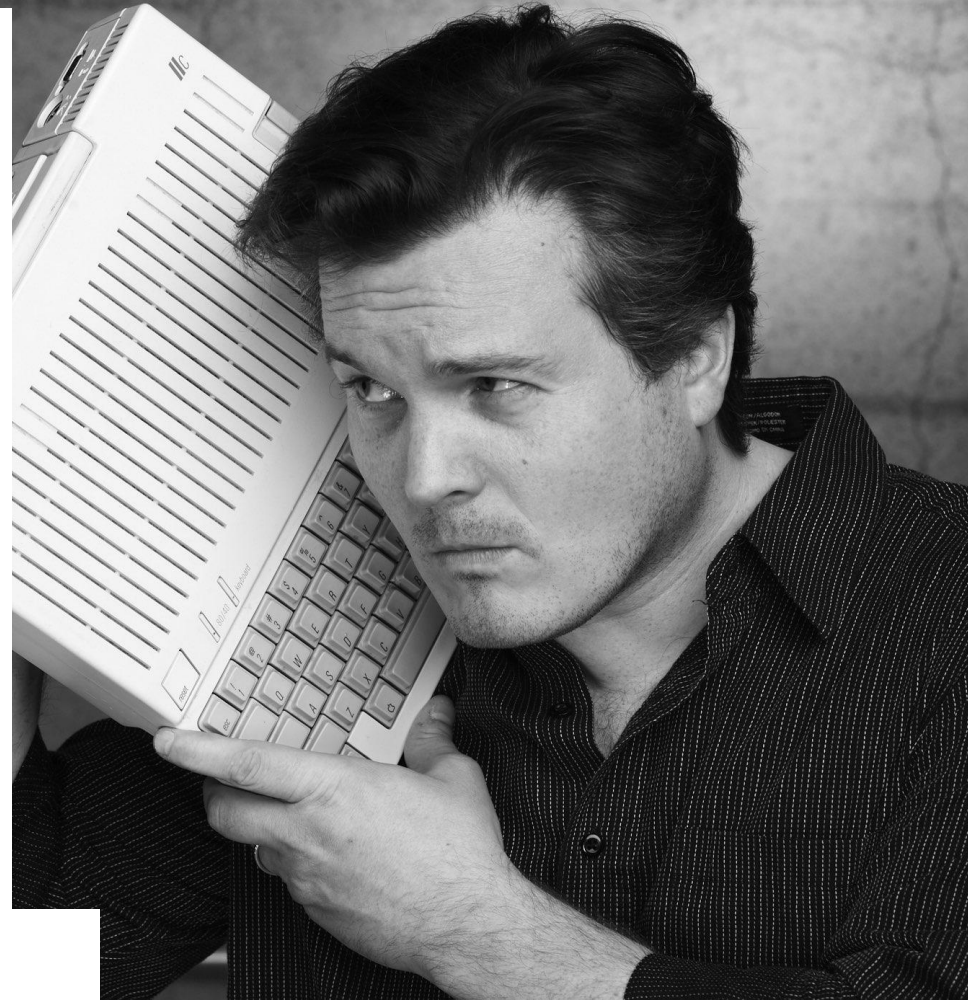
- Local/hybrid intelligence
- Adaptive, generative UI
- Multimodality
- Human-in-the-loop
- Personalization
- Browser interaction
- Collaboration

Ensuring a healthy Web

# Stephen Hood

Open Source AI Lead

[shood@mozilla.com](mailto:shood@mozilla.com)



# **Mozilla**

# **Builders**

**[builders.mozilla.org](https://builders.mozilla.org)**

