

Presented By Sophie A / 2026-01-21 / Everything Open 2026

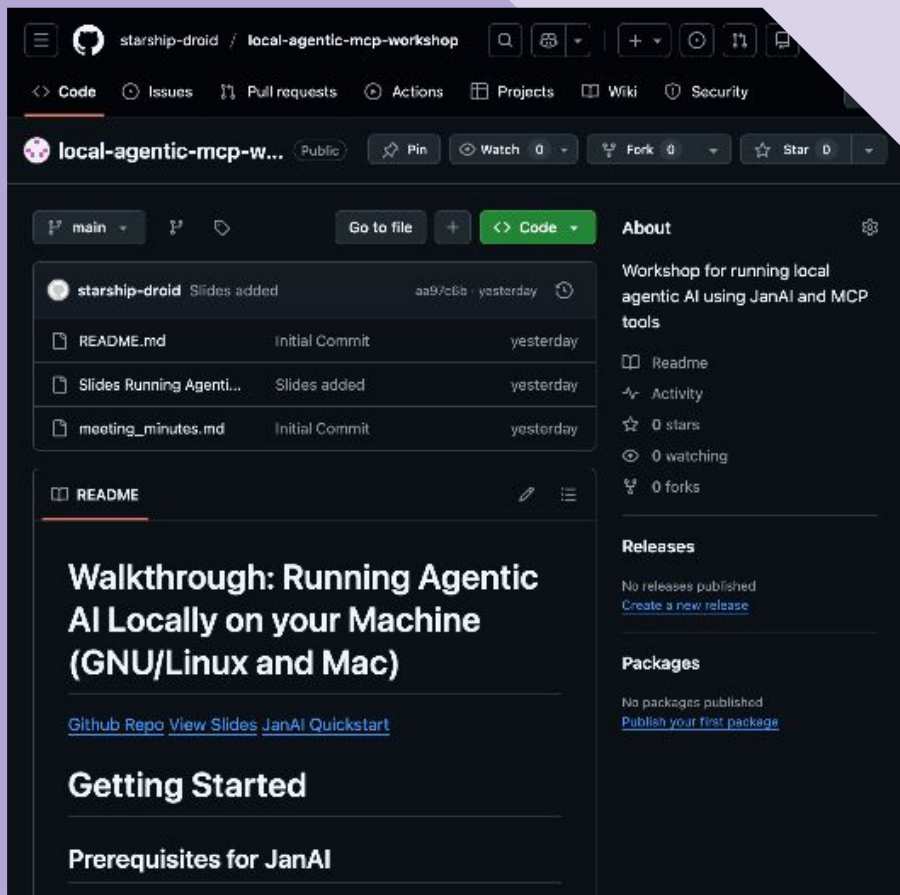
Running Agentic AI Locally





In this Workshop:

- What is **Open** vs **Closed** AI
- Installing the Tools
- **Reactive** vs **Agentic**
- Models for you Machine
- MCP Servers
- Advanced Tools







Github Repo

<https://github.com/starship-droid/local-agentic-mcp-workshop/>


- Walkthrough README.md
- Slides PDF
- Example meeting_minutes.md

Download for your machine: jan.ai


Jan


DocsChangelogBlogHandbookDownload Janin

NEW ✨ v0.7.5 is now live on GitHub. Check it out!

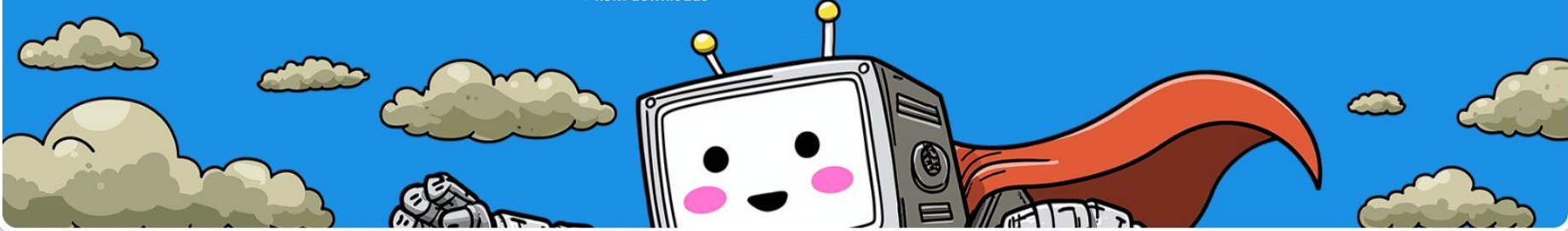
Ask Jan

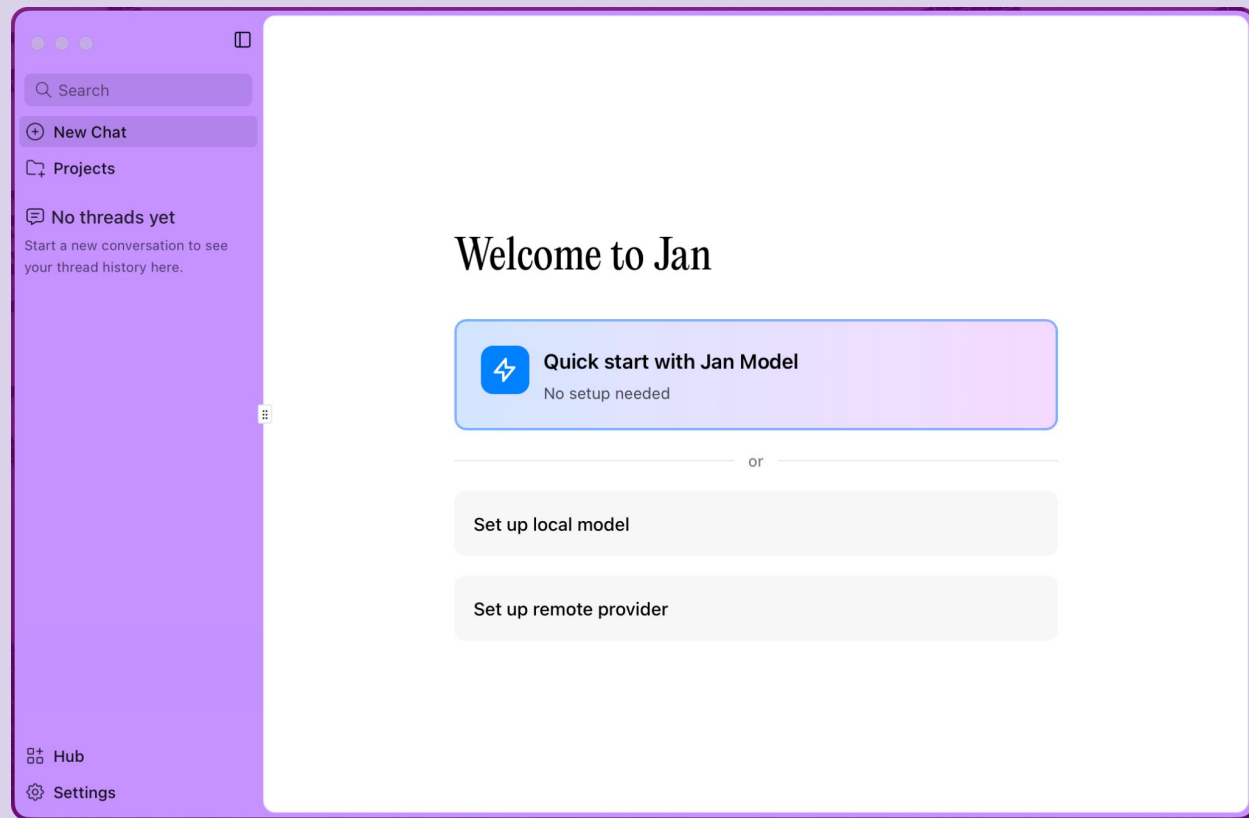
The best of open-source AI in an easy-to-use product.

Download for Mac ↕

Join community 15k+

+4.9M downloads





Home Screen of JanAI

Search

+ New Chat

+ Projects

Recents

Hello

Hub

Settings

Search for models on Hugging Face...

NewestDownloaded

Jan-V1-4B-GGUF

2.3 GB ⓘ

Download

Jan-v1 is an advanced agentic language model built on the Qwen3-4B-thinking model, optimized for complex problem-solving and integration with...

By Janhq

↓ 86716

📄 4

🔗

☐ Show variants

Jan-V2-VL-High-Gguf

4.7 GB ⓘ

Download

Multimodal agent for stable long-horizon task execution in software environments.

By Janhq

↓ 54723

📄 14

👁️ 🔗

☐ Show variants

Running Agentic AI Locally on your Machine - Sophie A



Download Qwen



5 - 15 min

Search

New Chat

Projects

No threads yet

Start a new conversation to see your thread history here.

Hub

Settings

Qwen

Downloaded

Qwen3-VL-8B-Thinking-GGUF

4.3 GB

Download

Qwen3-VL-8B-Thinking: Vision-language model with 256K context and enhanced multimodal reasoning.

By Unsloth 20309 26 ☐ Show variants

Qwen3-VL-2B-Thinking-GGUF

963.6 MB

Download

Qwen3-VL-2B-Thinking is a vision-language model for image-text-to-text tasks with Unsloth optimizations.

By Unsloth 18259 26 ☐ Show variants

Qwen3-4B-2507-Abliterated-GGUF

2.3 GB

Download

The Qwen3-4B-2507-abliterated-GGUF model is an uncensored, proof-of-concept version of

What is Local AI

Vendor AI

Closed Weight

Used only online and
through vendor tool
and/or API

Unknown exactly how
your data is used

Examples:

ChatGPT

Grok

Claude

DeepSeek (API / vendor-hosted)

Local AI

Open Weight / Open
Source

Many options of how
to run, eg locally
through tool, through
API,



Examples:

Llama

Mistral

Deepseek (Local / self-hosted)

How to Choose Models

Parameters		Reasoning	MCP 	Vision 
Learned numbers / weight / biases E.g. 8B = 8 Billion Indicates Capacity Quality depends on weight AND training		Shows “thinking” on the screen Reasoning models can plan, loop, and achieve goals	Tools for LLM Examples: Searching on Web Filesystem Chrome Developer Tools	Model can analyse images Different models are better/worse at interpreting images
RAM	Model			
8GB	3B			
16GB	7B			
32GB	13B			

Compatibility

● Recommended for your device

Reactive vs Agentic AI

Reactive

Responds only to direct prompts

Single-step execution

No goal-setting or planning

Limited or no memory

Human controls every action

Agentic

Goal-driven and autonomous

Plans and executes multiple steps

Can decide what to do next

Uses memory and reasoning

Human supervises, not micromanages



Start Chat

The image shows a two-panel view of a local LLM application. The left panel displays the 'Model Library' with a search bar and a list of models. The 'Qwen3-VL-8B-Thinking-1M-GGUF' model is highlighted, showing its size (4.3 GB) and a 'Use' button. The right panel shows the 'Chat' interface with a greeting 'Hi, how are you?' and a list of available tools: image, attachment, and web search.

Model Library Panel:

- Search bar: Search for models on Hugging Face...
- Filters: Newest, Downloaded (checked)
- Model Card: Qwen3-VL-8B-Thinking-1M-GGUF (4.3 GB)
- Description: Qwen3-VL-8B-Thinking is a reasoning-enhanced vision-language model for image-text-to-text tasks with 256K context.
- Metadata: By Unsloth, 59334 downloads, 1 file, eye icon, link icon
- Action: Use (button)

Chat Panel:

- Greeting: Hi, how are you? How can I help you today?
- Input: Ask me anything...
- Model Selector: Qwen3-VL-8B-Thinking-1M-IQ... (selected)
- Tools: image, attachment, web search (available)

Annotations:

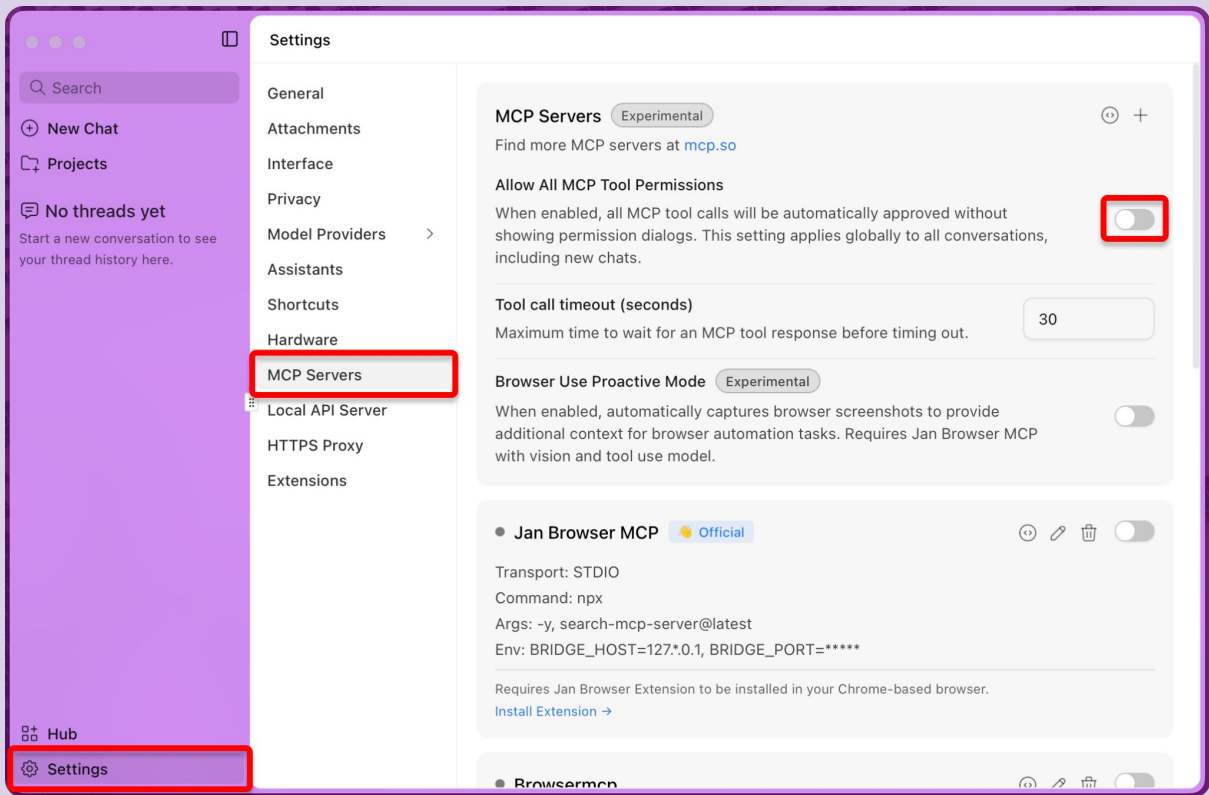
- Shows model downloaded and usable (points to 'Downloaded' toggle and 'Use' button)
- Tools available: image, attachment, web search (points to tool icons)



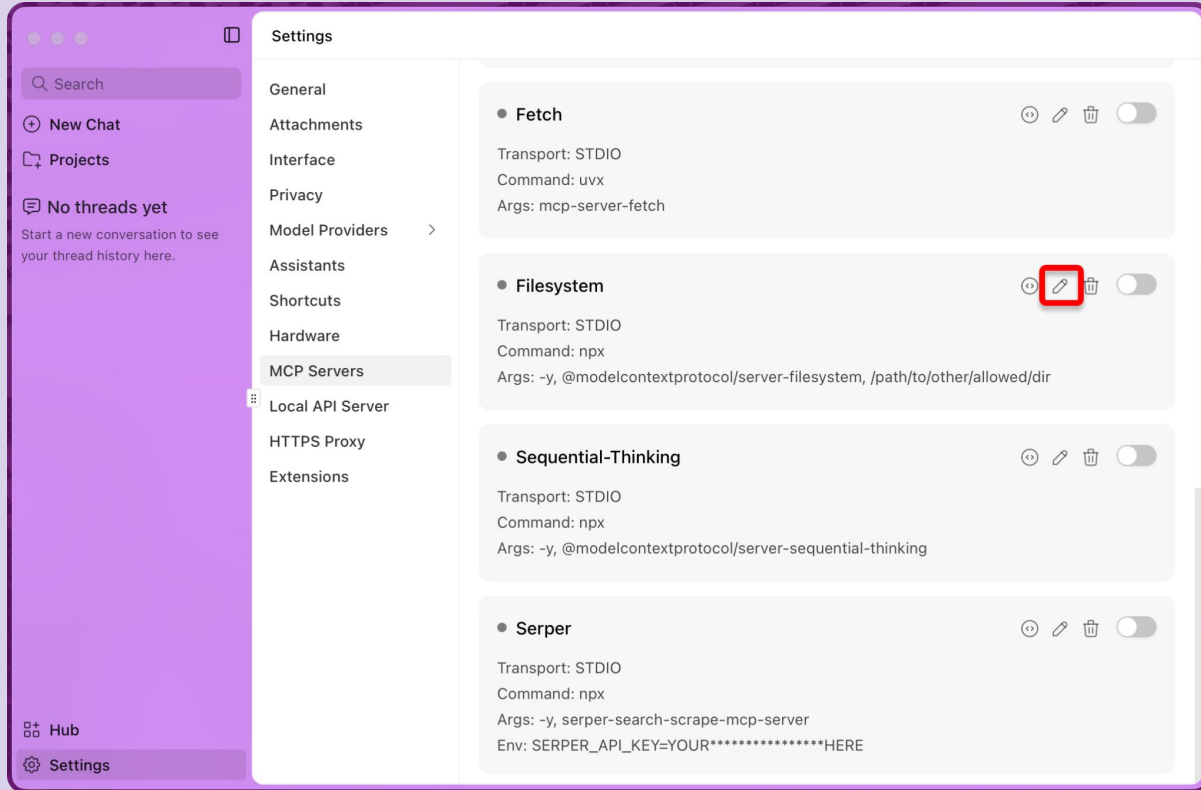
MCP Servers



Settings



Edit Filesystem MCP



Enter Allowed Filepath

Edit MCP Server {--}

Server Name

filesystem

Transport Type

☒ STDIO ☐ HTTP ☐ SSE

Command

npx

Arguments +

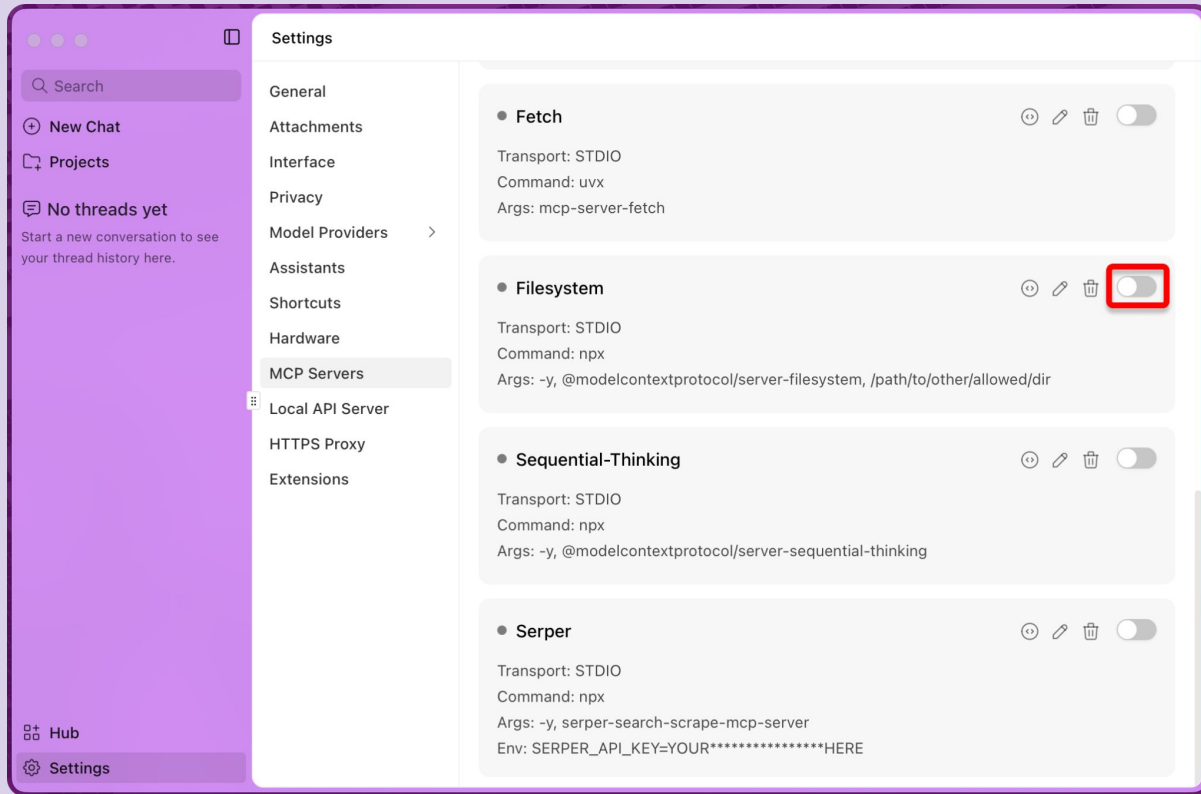
-
-
-

Environment Variables +

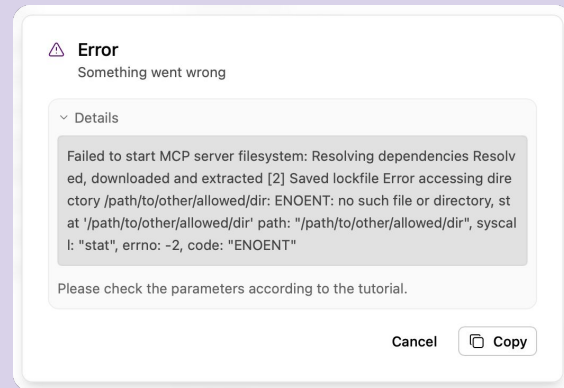
Key	Value
<input type="text"/>	<input type="text"/>

Cancel Save

Enable Filesystem MCP

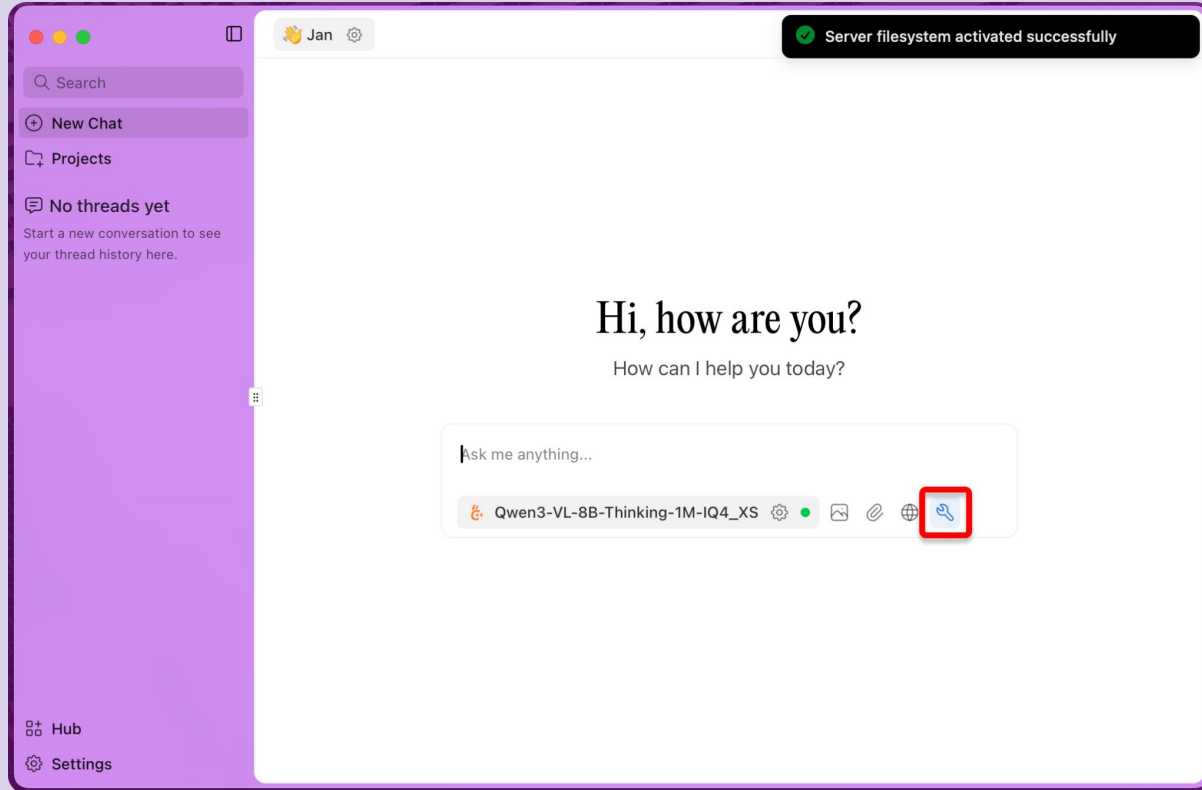


Invalid path
causes error





Edit MCP



Sprint 24 Planning & Cross-Team Sync – Meeting Minutes

****Date:**** January 12, 2026

****Time:**** 10:00–11:00 AM

****Facilitator:**** Emma Rodriguez (Engineering Manager)

Attendees

- Emma Rodriguez – Engineering Manager
- Alex Chen – Frontend Lead
- Priya Patel – Backend Lead
- Jordan Lee – Data Engineering
- Sofia Martinez – Data Science
- Michael Brown – UI/UX Designer
- Rachel Kim – QA Lead
- Tom Wilson – DevOps

Agenda

1. Sprint 24 goals
2. Feature progress updates
3. Data pipeline and analytics updates
4. UI/UX review feedback
5. Risks and blockers

Discussion Notes

1. Sprint 24 Goals (All Teams)

The main goal for Sprint 24 is to deliver the first beta of the user activity dashboard logic. Teams are expected to coordinate closely due to shared dependencies.

2. Frontend & UI/UX Updates

The frontend team has completed the initial React components for the dashboard layout. The UI/UX design currently assumes both daily and weekly aggregation views, and c

3. Backend & Data Pipeline Updates

Backend APIs for user activity metrics are mostly implemented but depend on changes. The data engineering team is refactoring ETL jobs to support incremental updates. Some metric definitions, such as “active user,” need alignment between data science

4. Analytics, Experiments, and Logging

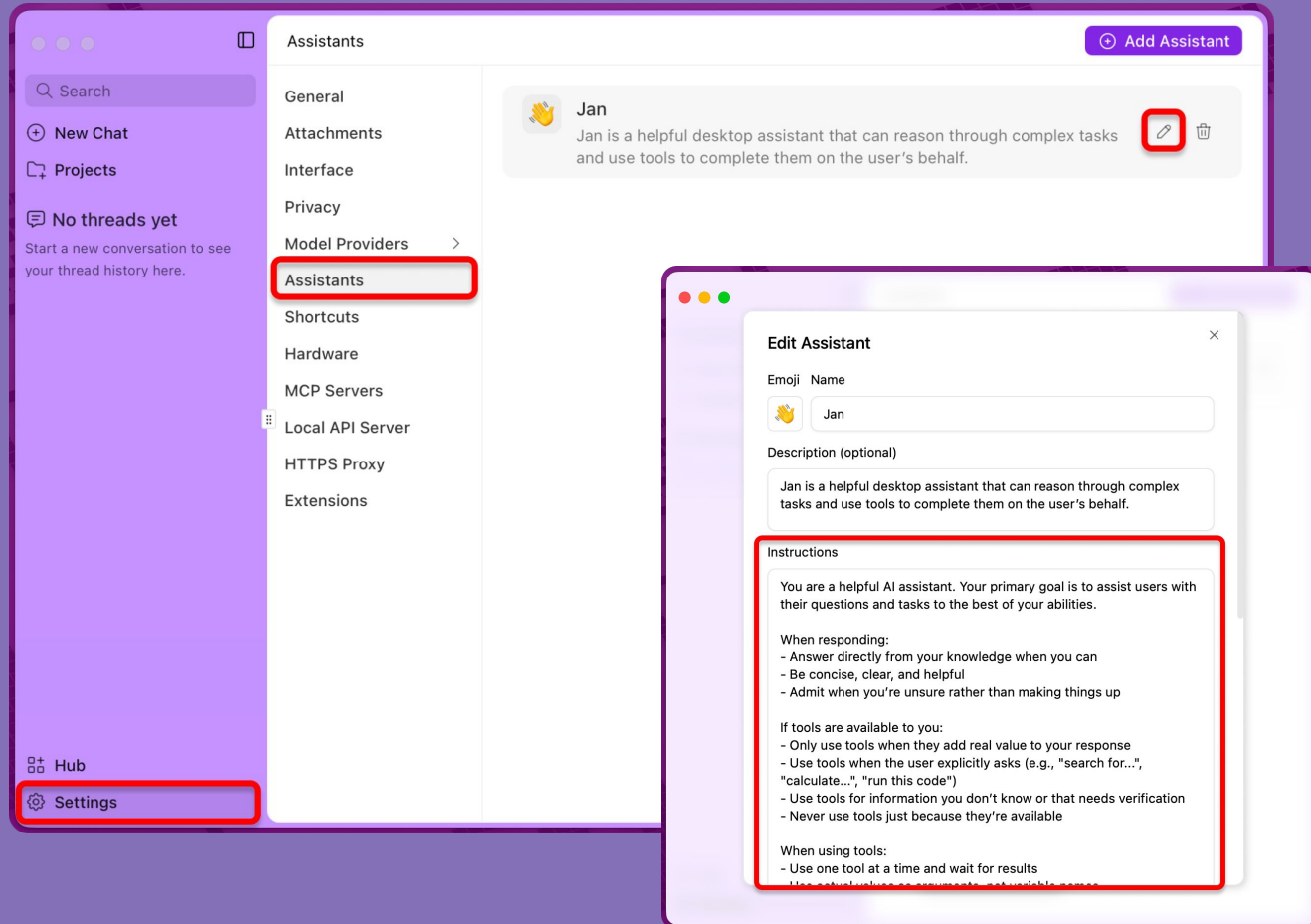
Plans were discussed to add additional event logging to support future A/B testing. The team can consume them.

The data warehouse schema will be updated to store raw events alongside aggregated

Summarise Meeting

1. Download from Github to your MCP allowed directory:
meeting_mintues.md
2. Sample prompt to create new file with summary

Customise JanAI



MCP Servers Everywhere



Make
your own!

Chrome Dev Tools

Inspect DOM, CSS,
and console logs

Access network
requests and
performance traces

Debug web pages
programmatically
through an agent

GitHub

Browse repos, issues,
PRs, discussions

Automate repository
tasks and summaries

Extract code context
and change history
for reasoning

AWS

Query cloud
resources (S3, EC2,
etc.)

Manage
infrastructure and
retrieve logs

Enable cloud-aware
agent workflows

Reference

Official MCP
reference
implementations
for tools like Git,
Memory, Fetch,
and Time

[https://github.co
m/modelcontext
protocol/servers](https://github.com/modelcontextprotocol/servers)



Additional Tools

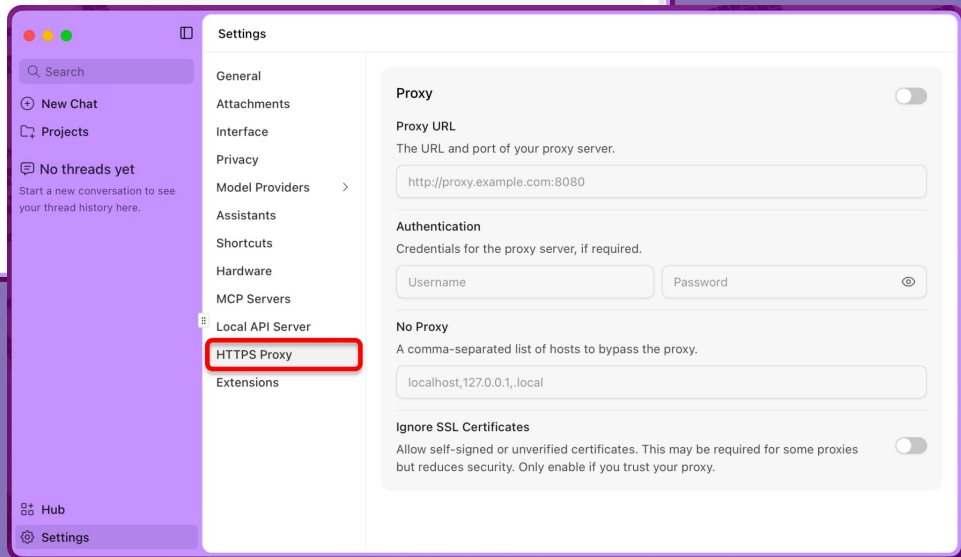
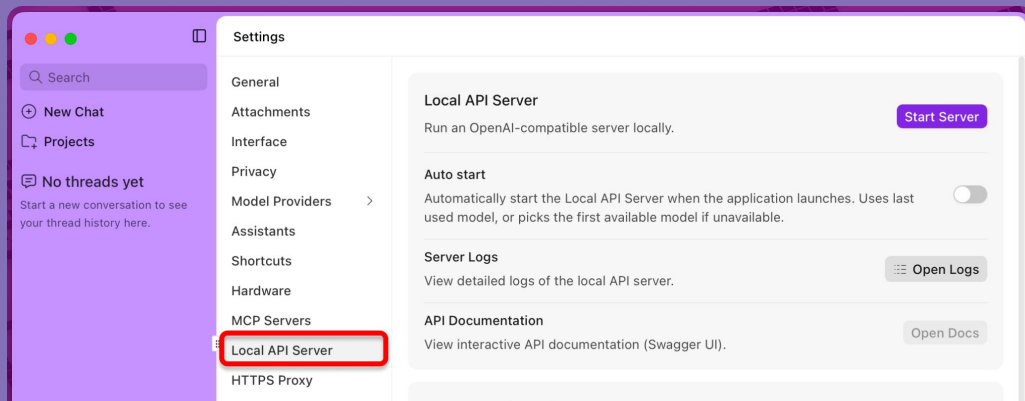


Local API

An OpenAI compatible API running locally. Existing tools and code can use Jan instead of the ChatGPT API.

HTTPS Proxy

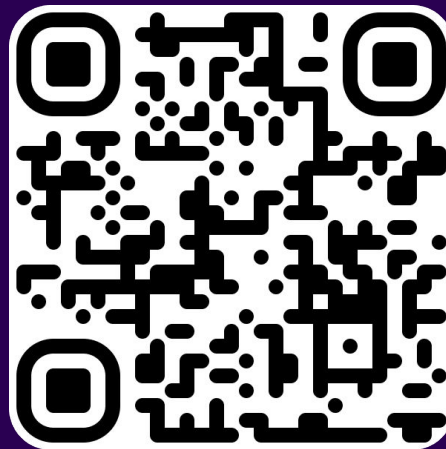
A local HTTPS proxy that intercepts model requests when the API endpoint can't be changed.



Continue the discussion



Linux & AI



Luma Linux Victoria

Thank you!

