

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

Running Agentic AI Locally





In this Workshop:

- **Local vs Closed & Vendor Hosted AI**
- Installing the Tools
- **Reactive vs Agentic**
- Models for you Machine
- MCP Servers
- Advanced Tools

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

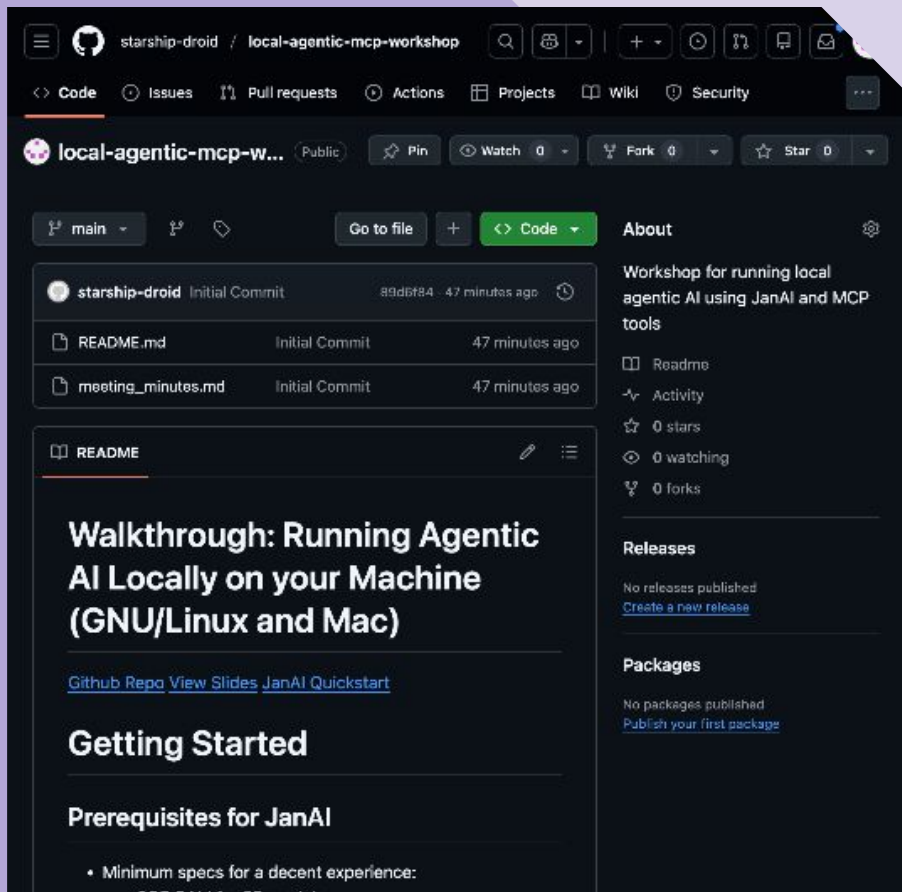
Many options of how
to run, eg locally
through tool, through
API,

Examples:

Llama

Mistral

Deepseek (Local / self-hosted)



Github Repo

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

Go to README.md



Hub

Search

+ New Chat

Projects

Recents

Hello

Hub

Settings

Search for models on Hugging Face...

Newest

Downloaded

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



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

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-part interface for starting a chat. The left panel displays a list of models with a search bar at the top. A red box highlights the 'New Chat' button in the sidebar. Another red box highlights the 'Downloaded' toggle switch at the top of the model list. A third red box highlights the 'Use' button for the selected model, 'Qwen3-VL-8B-Thinking-1M-GGUF'. 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. Two callout boxes provide additional context: one points to the 'Use' button and says 'Shows model downloaded and usable', and the other points to the tool icons and says 'Tools available: image, attachment, web search'.

Search for models on Hugging Face...

Newest ☒ Downloaded

Qwen3-VL-8B-Thinking-1M-GGUF 4.3 GB Use

Qwen3-VL-8B-Thinking is a reasoning-enhanced vision-language model for image-text-to-text tasks with 256K context.

By Unsloth 59334 1

New Chat

Projects

No threads yet

Start a new conversation to see your thread history here.

Hi, how are you?

How can I help you today?

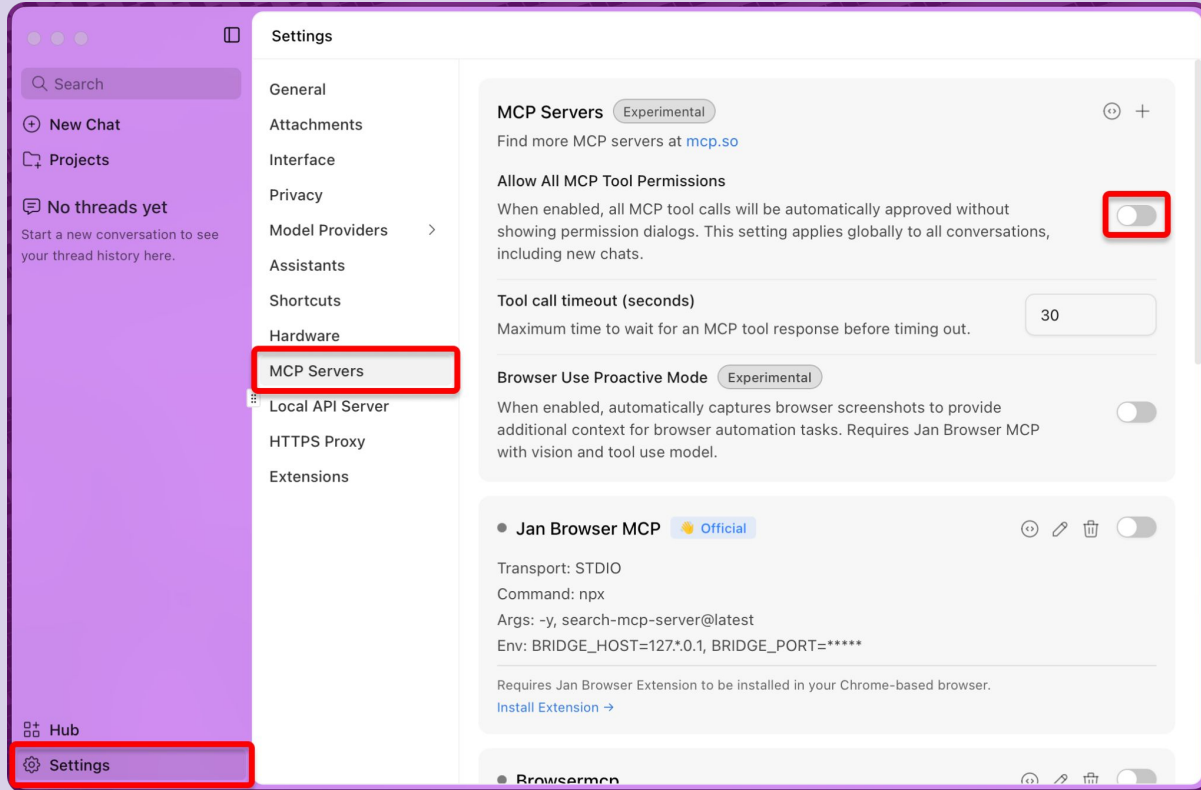
Ask me anything...

Qwen3-VL-8B-Thinking-1M-IQ...

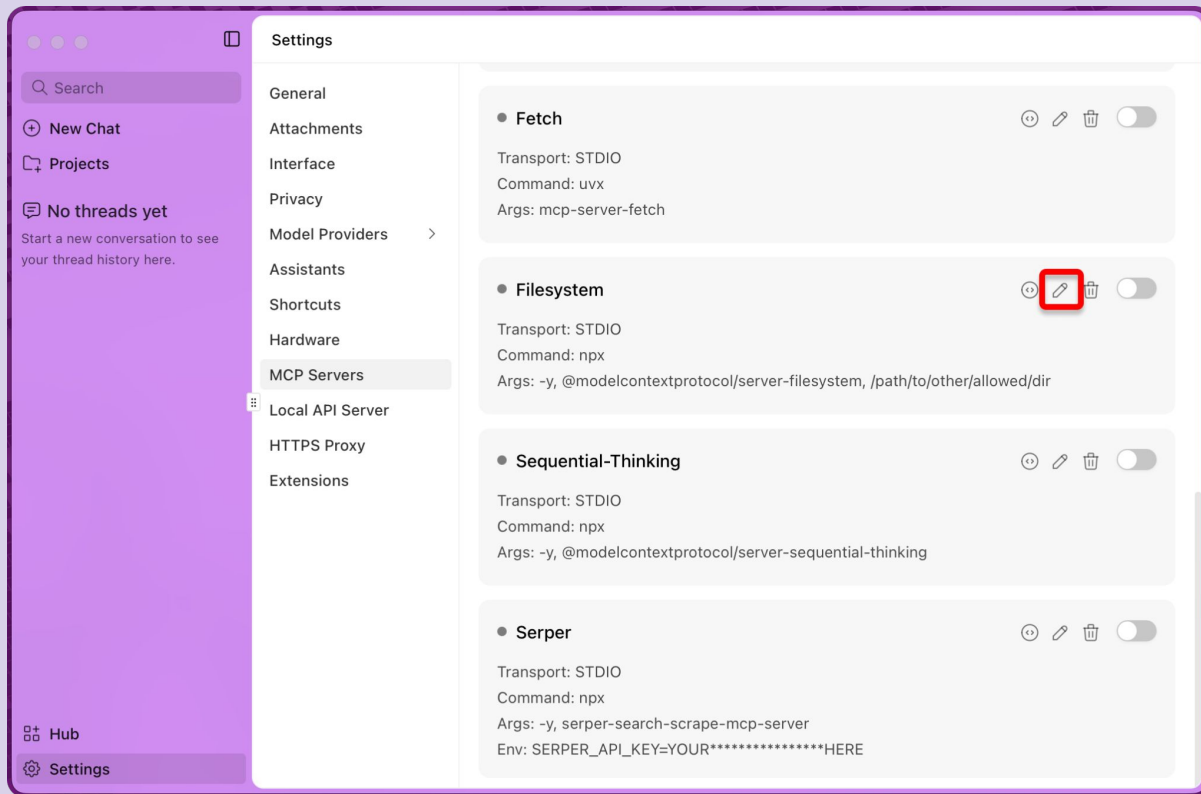
Shows model downloaded and usable

Tools available: image, attachment, web search

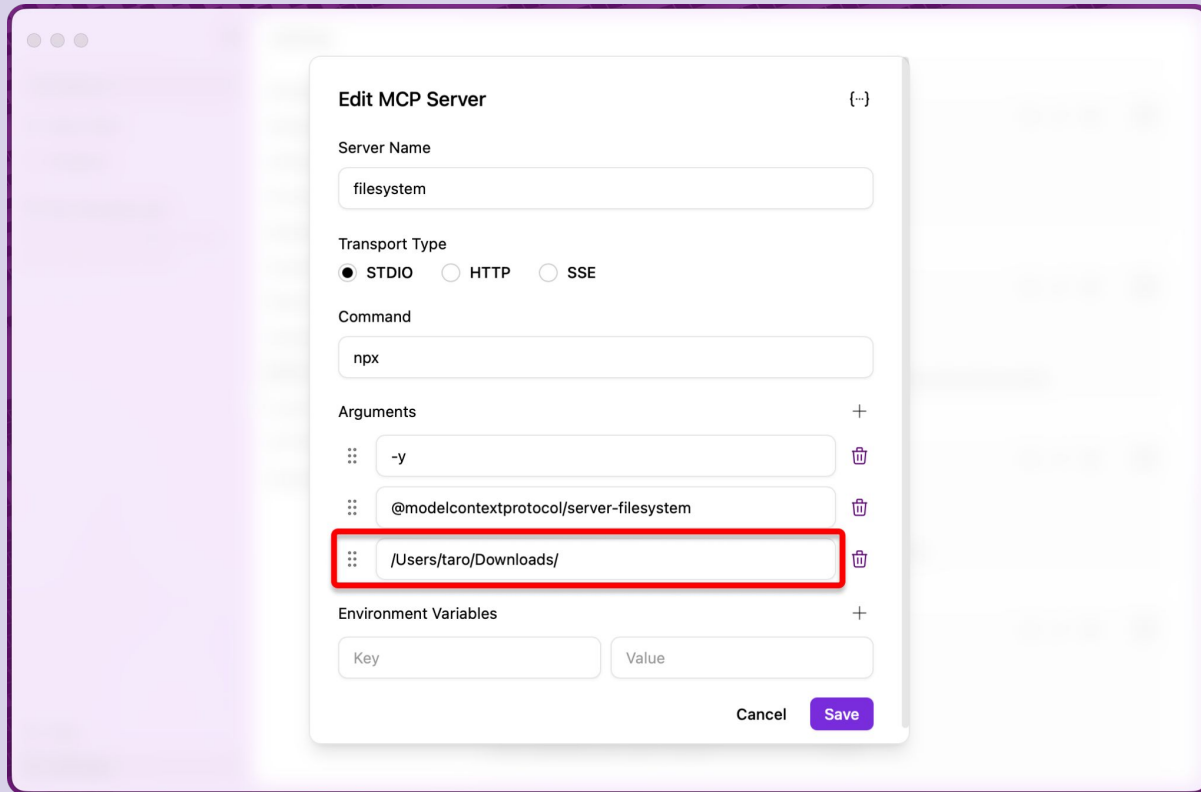
Enable MCP Permission



Edit Filesystem MCP



Enter Allowed Filepath



Edit MCP Server {--}

Server Name

filesystem

Transport Type

☒ STDIO ☐ HTTP ☐ SSE

Command

npx

Arguments +

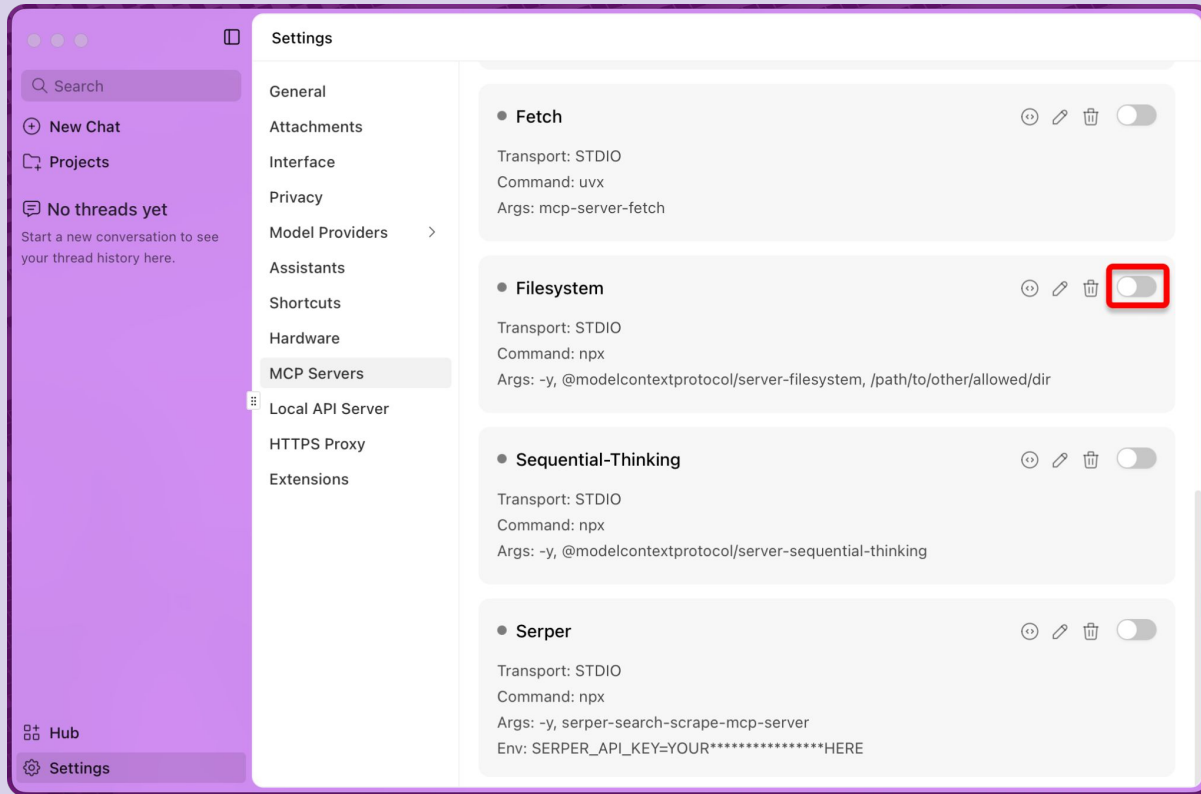
- y
- @modelcontextprotocol/server-filesystem
- /Users/taro/Downloads/**

Environment Variables +

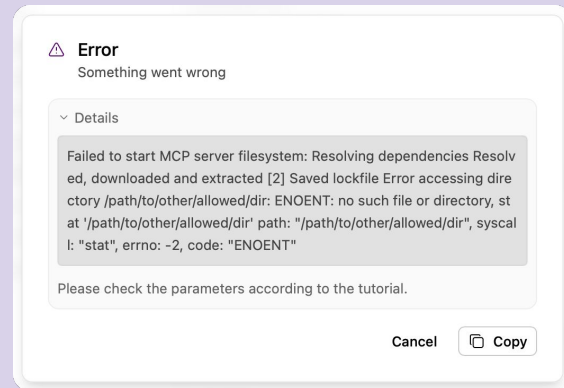
Key	Value
-----	-------

Cancel Save

Enable Filesystem MCP

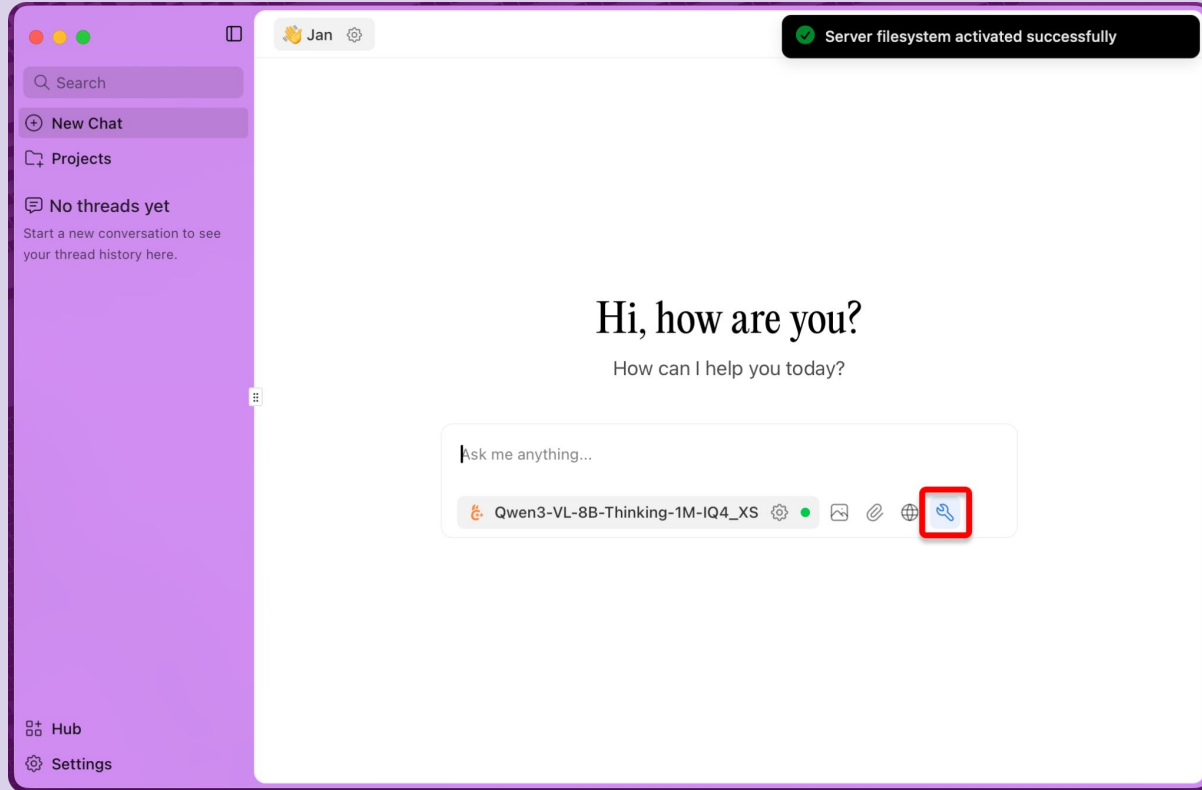


Invalid path
causes error





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. 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 from the data engineering team. 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

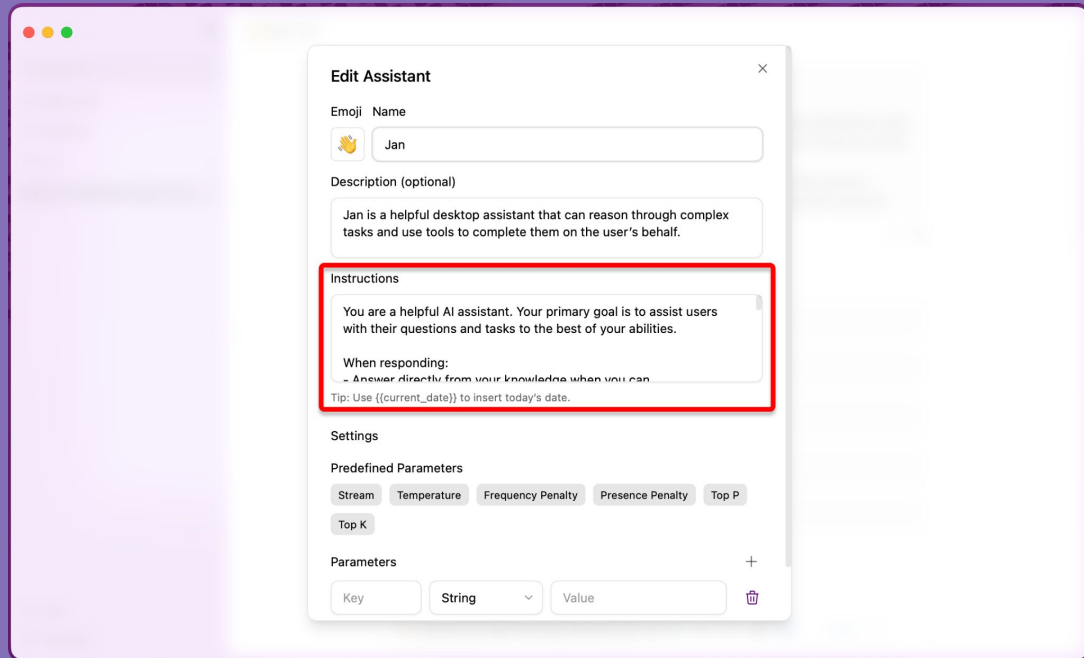
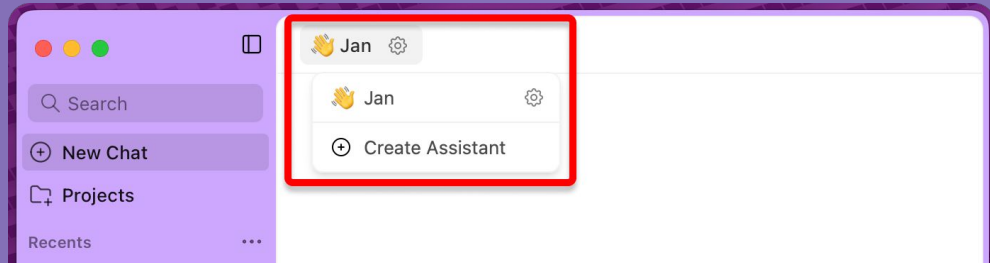
Plans were discussed to add additional event logging to support future A/B testing. The team can consume them.

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

Summarise Meeting

meeting_mintues.md

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)

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

QR CODE

Thank you!

