Serving Local LLM

Jae-ik Shin

2025.06.13

How do you use LLM?

- On-Cloud

OpenAI ChatGPT, Google Gemini, Anthropic Claude, Grok, Deepseek, etc

- Routing

MS Azure, Google, Amazon, NVIDIA, Perplexity, Cursor AI, OpenRouter, etc

- Self-hosting?

Self-hosted Large Language Model

- Open-weight model

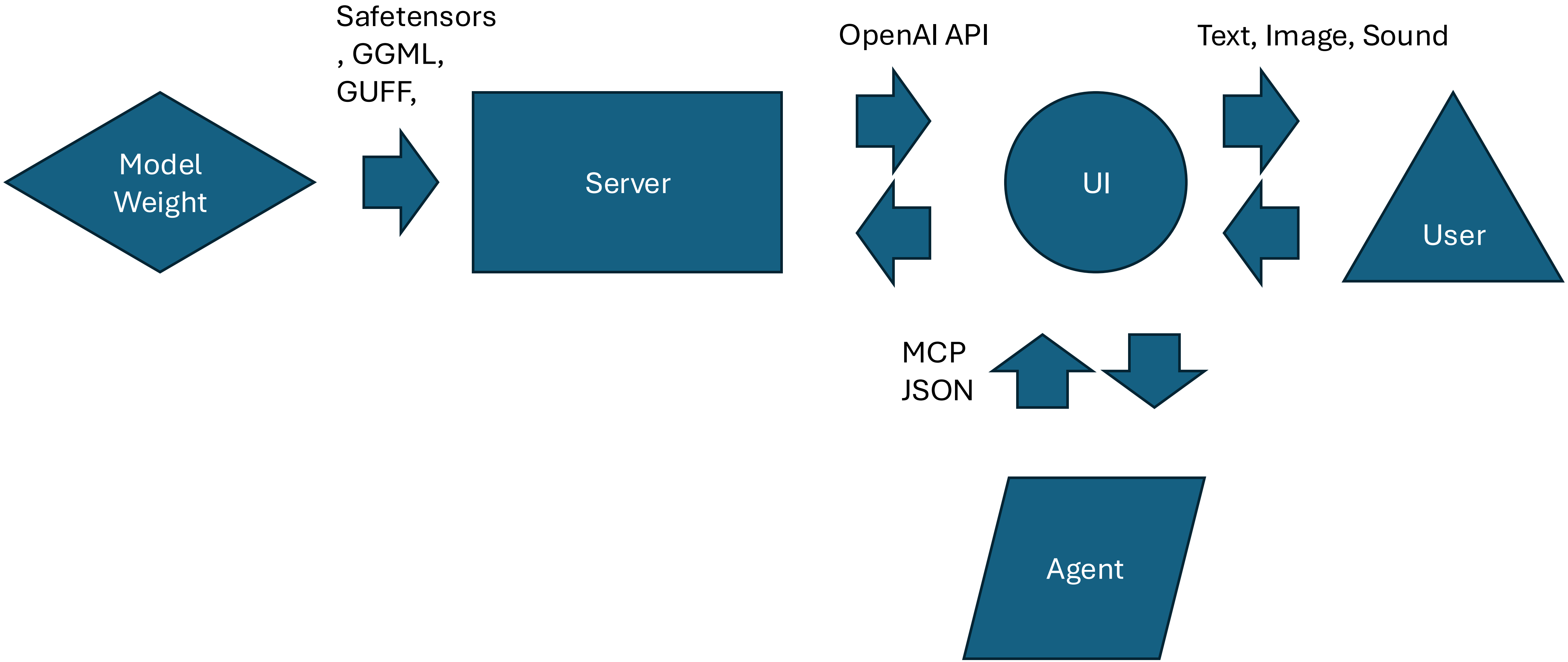
Hugging Face, Model Scope, Ollama

- Open AI API compatible Server
- vLLM, Ollama, LLaMa.cpp

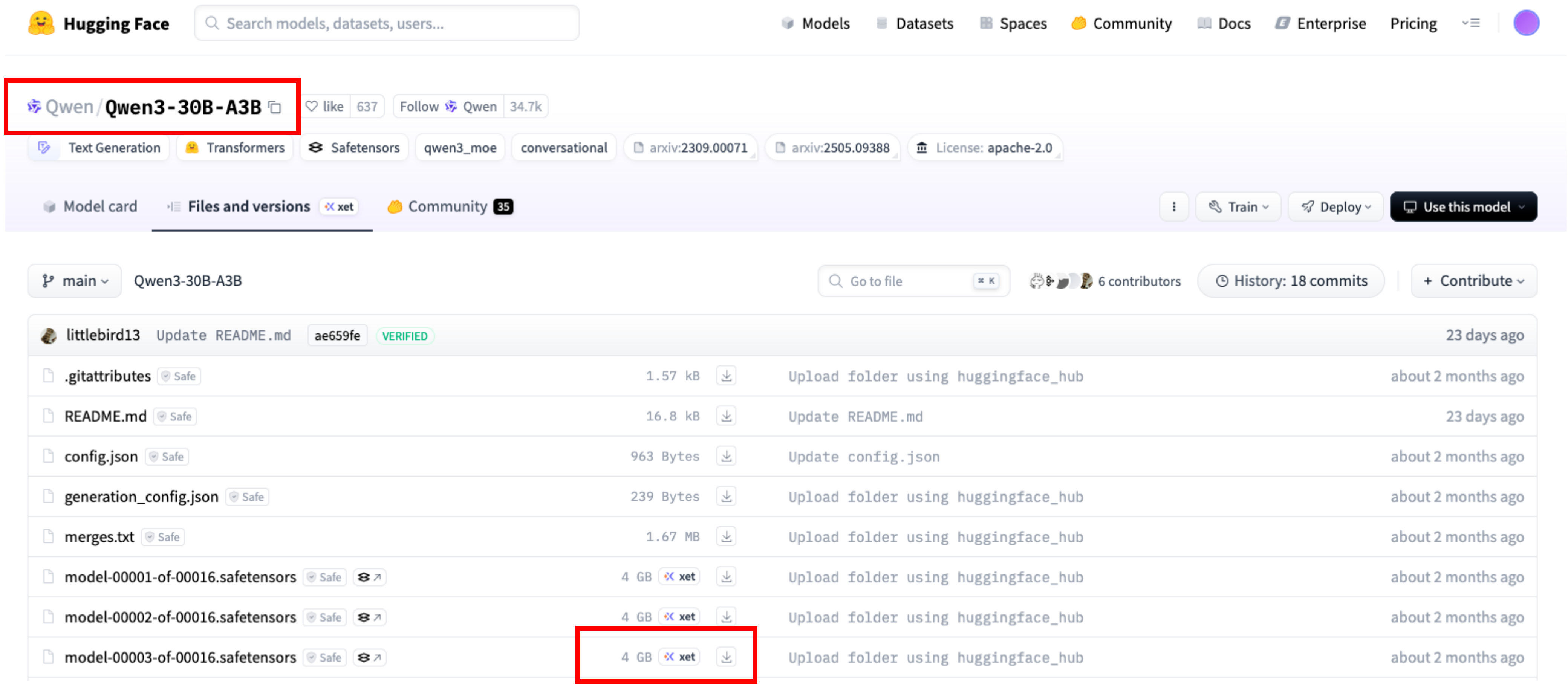
- User Interface

Open WebUI, Msty, LM studio, n8n?

Communication



Model : Hugging Face (HF)



littlebird13 Update README.md ae659fe VERIFIED			23 days ago
.gitattributes Safe	1.57 kB	Upload folder using huggingface_hub	about 2 months ago
README.md Safe	16.8 kB	Update README.md	23 days ago
config.json Safe	963 Bytes	Update config.json	about 2 months ago
generation_config.json Safe	239 Bytes	Upload folder using huggingface_hub	about 2 months ago
merges.txt Safe	1.67 MB	Upload folder using huggingface_hub	about 2 months ago
model-00001-of-00016.safetensors Safe	4 GB xet	Upload folder using huggingface_hub	about 2 months ago
model-00002-of-00016.safetensors Safe	4 GB xet	Upload folder using huggingface_hub	about 2 months ago
model-00003-of-00016.safetensors Safe	4 GB xet	Upload folder using huggingface_hub	about 2 months ago

https://huggingface.co/Qwen/Qwen3-30B-A3B/tree/main

Download from HF

```
from huggingface_hub import snapshot_download

snapshot_download(repo_id=“Qwen/Qwen3-30B-A3B")
```

https://huggingface.co/docs/huggingface_hub/guides/download#download-an-entire-repository
https://huggingface.co/docs/huggingface_hub/v0.33.0/en/package_reference/file_download

vLLM

- https://github.com/vllm-project/vllm

- https://pypi.org/project/vllm/

- Download model from HF
- Run server with model by python script
- Connect to server in UI

Ollama

- Download binary or docker
- https://github.com/ollama/ollama

- Turn on server

- Download(Pull) model
- https://ollama.com/search

- Run model

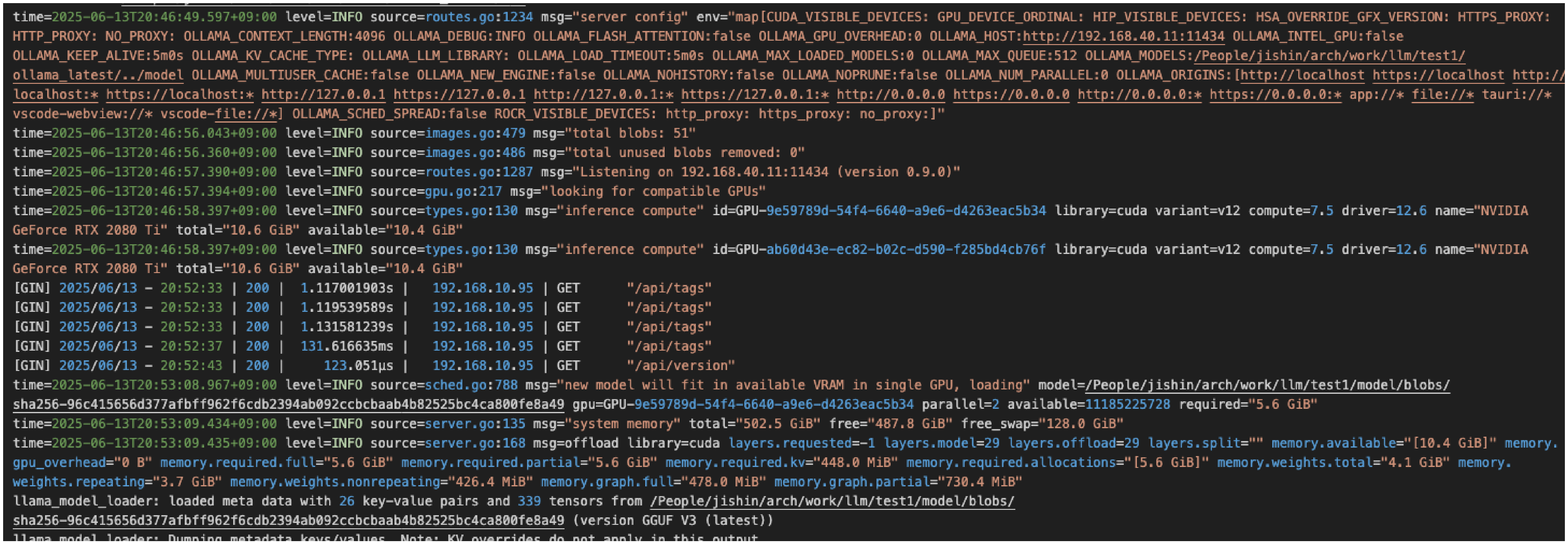
Environment variable & CLI

- export OLLAMA_HOME="/People/jishin/arch/work/llm/test1/ollama_latest"
- export HOST_IP="192.168.40.11"
- export OLLAMA_HOST="${HOST_IP}:11434"
- export OLLAMA_MODELS="${OLLAMA_HOME}/../model"

- ./ollama serve
- ./ollama pull deepseek-r1:8b
- ./ollama list
- ./ollama run deepseek-r1:8b
- ./ollama ps

Turn on Ollama Server & Load model

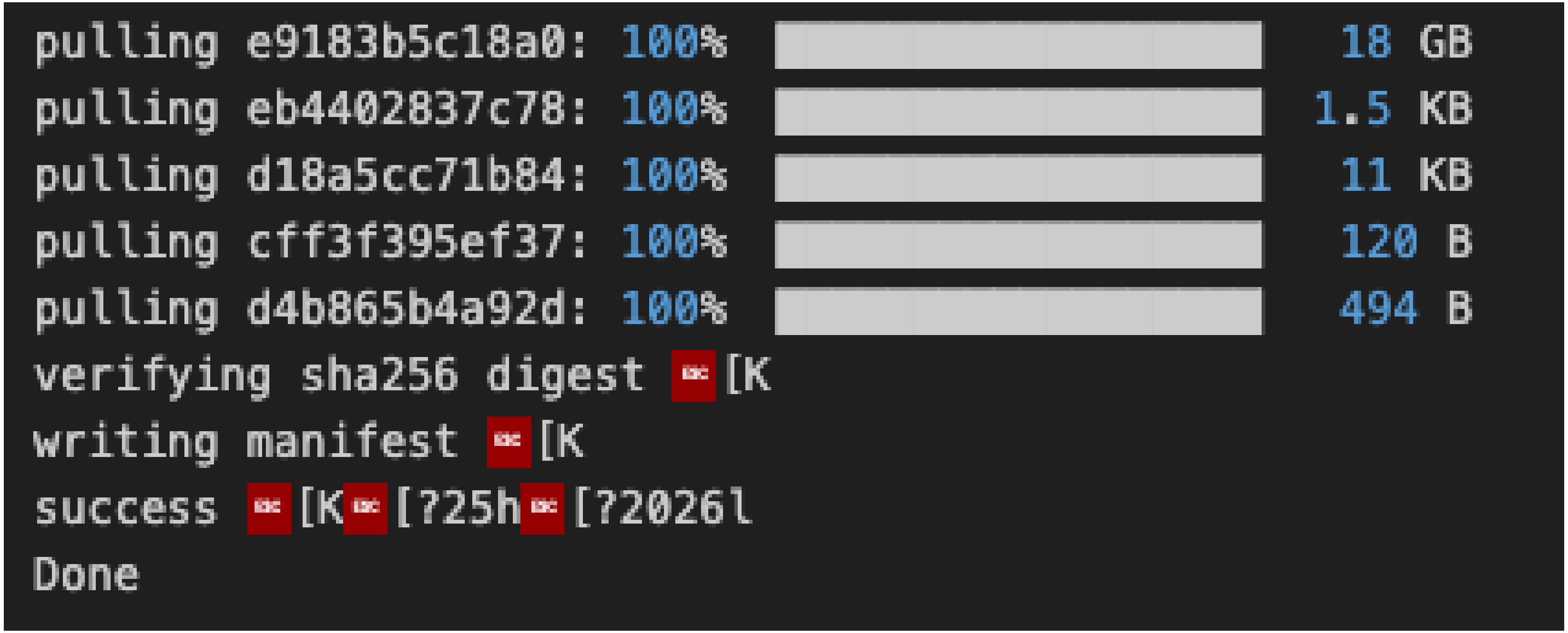
```
./ollama  serve
./ollama run deepseek-r1:8b
```



```
time=2025-06-13T20:46:49.597+09:00 level=INFO source=routes.go:1234 msg="server config" env="map[CUDA_VISIBLE_DEVICES: GPU_DEVICE_ORDINAL: HIP_VISIBLE_DEVICES: HSA_OVERRIDE_GFX_VERSION: HTTPS_PROXY:
HTTP_PROXY: NO_PROXY: OLLAMA_CONTEXT_LENGTH:4096 OLLAMA_DEBUG:INFO OLLAMA_FLASH_ATTENTION:false OLLAMA_GPU_OVERHEAD:0 OLLAMA_HOST:http://192.168.40.11:11434 OLLAMA_INTEL_GPU:false
OLLAMA_KEEP_ALIVE:5m0s OLLAMA_KV_CACHE_TYPE: OLLAMA_LLM_LIBRARY: OLLAMA_LOAD_TIMEOUT:5m0s OLLAMA_MAX_LOADED_MODELS:0 OLLAMA_MAX_QUEUE:512 OLLAMA_MODELS:/People/jishin/arch/work/llm/test1/
ollama_latest/../model OLLAMA_MULTIUSER_CACHE:false OLLAMA_NEW_ENGINE:false OLLAMA_NOHISTORY:false OLLAMA_NOPRUNE:false OLLAMA_NUM_PARALLEL:0 OLLAMA_ORIGINS:[http://localhost https://localhost http://
localhost:* https://localhost:* http://127.0.0.1 https://127.0.0.1 http://127.0.0.1:* https://127.0.0.1:* http://0.0.0.0 https://0.0.0.0 http://0.0.0.0:* https://0.0.0.0:* app://* file://* tauri://*
vscode-webview://* vscode-file://*] OLLAMA_SCHED_SPREAD:false ROCR_VISIBLE_DEVICES: http_proxy: https_proxy: no_proxy:]"
time=2025-06-13T20:46:56.043+09:00 level=INFO source=images.go:479 msg="total blobs: 51"
time=2025-06-13T20:46:56.360+09:00 level=INFO source=images.go:486 msg="total unused blobs removed: 0"
time=2025-06-13T20:46:57.390+09:00 level=INFO source=routes.go:1287 msg="Listening on 192.168.40.11:11434 (version 0.9.0)"
time=2025-06-13T20:46:57.394+09:00 level=INFO source=gpu.go:217 msg="looking for compatible GPUs"
time=2025-06-13T20:46:58.397+09:00 level=INFO source=types.go:130 msg="inference compute" id=GPU-9e59789d-54f4-6640-a9e6-d4263eac5b34 library=cuda variant=v12 compute=7.5 driver=12.6 name="NVIDIA
GeForce RTX 2080 Ti" total="10.6 GiB" available="10.4 GiB"
time=2025-06-13T20:46:58.397+09:00 level=INFO source=types.go:130 msg="inference compute" id=GPU-ab60d43e-ec82-b02c-d590-f285bd4cb76f library=cuda variant=v12 compute=7.5 driver=12.6 name="NVIDIA
GeForce RTX 2080 Ti" total="10.6 GiB" available="10.4 GiB"
[GIN] 2025/06/13 - 20:52:33 | 200 |  1.117001903s |   192.168.10.95 | GET      "/api/tags"
[GIN] 2025/06/13 - 20:52:33 | 200 |  1.119539589s |   192.168.10.95 | GET      "/api/tags"
[GIN] 2025/06/13 - 20:52:33 | 200 |  1.131581239s |   192.168.10.95 | GET      "/api/tags"
[GIN] 2025/06/13 - 20:52:37 | 200 |  131.616635ms |   192.168.10.95 | GET      "/api/tags"
[GIN] 2025/06/13 - 20:52:43 | 200 |     123.051µs |   192.168.10.95 | GET      "/api/version"
time=2025-06-13T20:53:08.967+09:00 level=INFO source=sched.go:788 msg="new model will fit in available VRAM in single GPU, loading" model=/People/jishin/arch/work/llm/test1/model/blobs/
sha256-96c415656d377afbff962f6cdb2394ab092ccbcbaab4b82525bc4ca800fe8a49 gpu=GPU-9e59789d-54f4-6640-a9e6-d4263eac5b34 parallel=2 available=11185225728 required="5.6 GiB"
time=2025-06-13T20:53:09.434+09:00 level=INFO source=server.go:135 msg="system memory" total="502.5 GiB" free="487.8 GiB" free_swap="128.0 GiB"
time=2025-06-13T20:53:09.435+09:00 level=INFO source=server.go:168 msg=offload library=cuda layers.requested=-1 layers.model=29 layers.offload=29 layers.split="" memory.available="[10.4 GiB]" memory.
gpu_overhead="0 B" memory.required.full="5.6 GiB" memory.required.partial="5.6 GiB" memory.required.kv="448.0 MiB" memory.required.allocations="[5.6 GiB]" memory.weights.total="4.1 GiB" memory.
weights.repeating="3.7 GiB" memory.weights.nonrepeating="426.4 MiB" memory.graph.full="478.0 MiB" memory.graph.partial="730.4 MiB"
llama_model_loader: loaded meta data with 26 key-value pairs and 339 tensors from /People/jishin/arch/work/llm/test1/model/blobs/
sha256-96c415656d377afbff962f6cdb2394ab092ccbcbaab4b82525bc4ca800fe8a49 (version GGUF V3 (latest))
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not apply in this output.
```

Pull Model

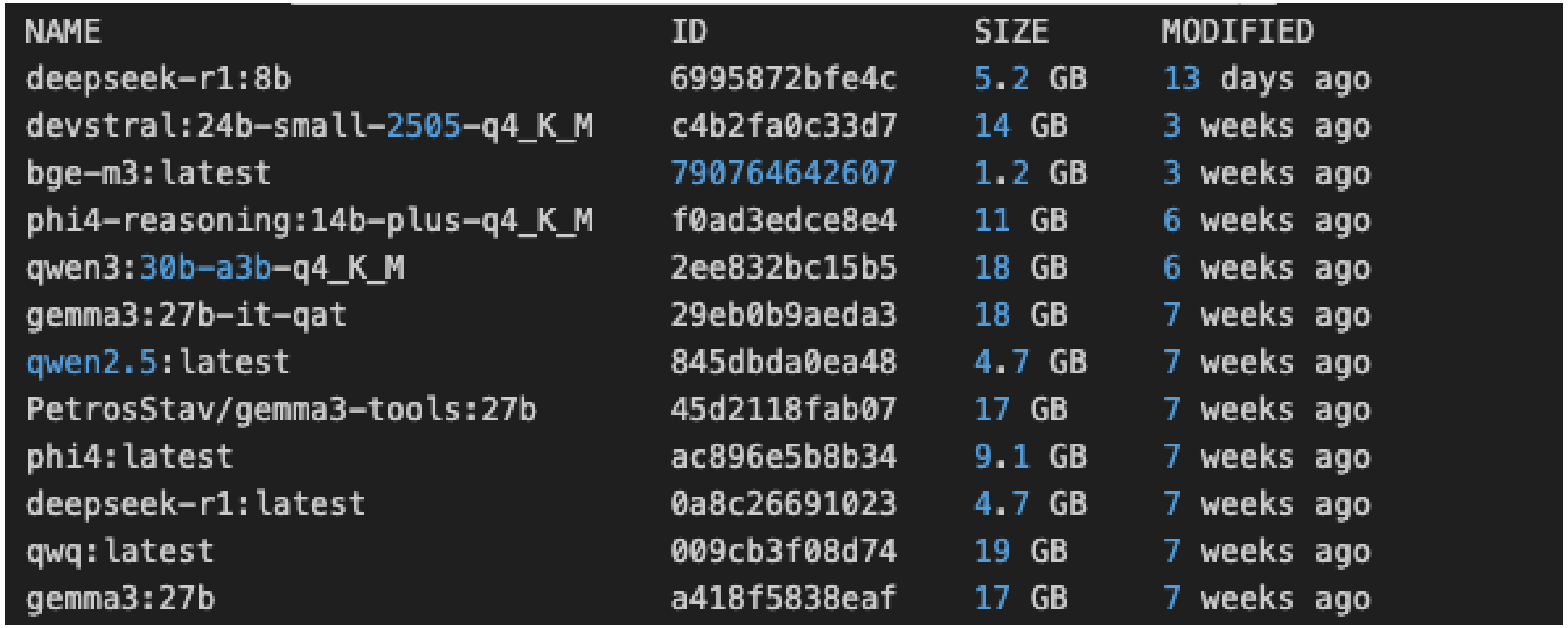
./ollama pull deepseek-r1:8b



```
pulling e9183b5c18a0: 100% ██████████████████   18 GB
pulling eb4402837c78: 100% ██████████████████  1.5 KB
pulling d18a5cc71b84: 100% ██████████████████   11 KB
pulling cff3f395ef37: 100% ██████████████████  120 B
pulling d4b865b4a92d: 100% ██████████████████  494 B
verifying sha256 digest ␛[K
writing manifest ␛[K
success ␛[K␛[?25h␛[?2026l
Done
```

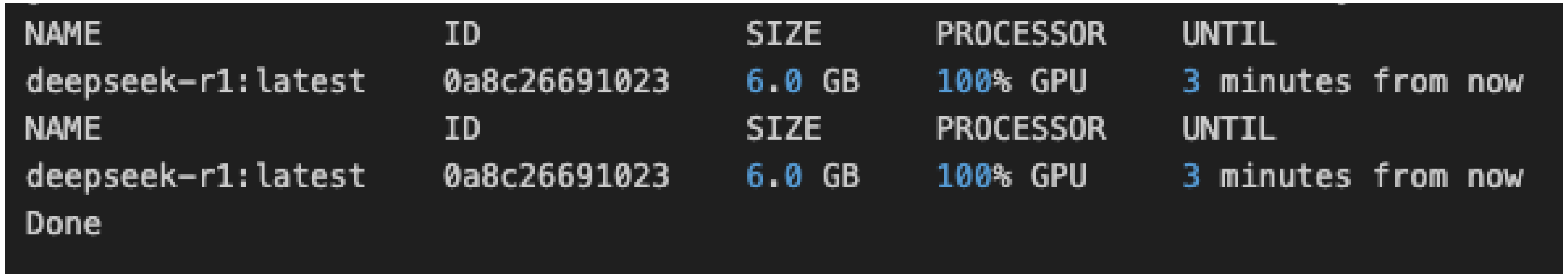
Listing Model & PS

./ollama list



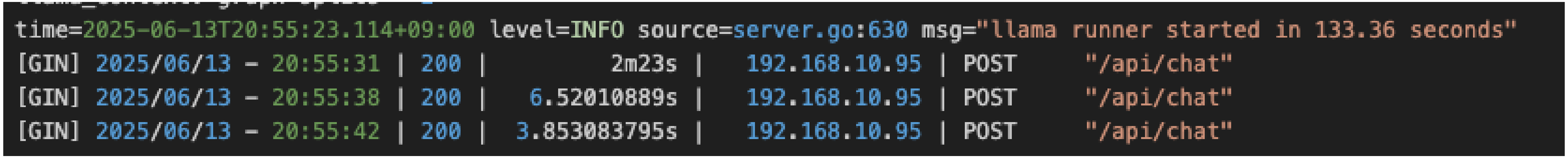
```
NAME                            ID              SIZE      MODIFIED
deepseek-r1:8b                  6995872bfe4c    5.2 GB    13 days ago
devstral:24b-small-2505-q4_K_M  c4b2fa0c33d7    14 GB     3 weeks ago
bge-m3:latest                   790764642607    1.2 GB    3 weeks ago
phi4-reasoning:14b-plus-q4_K_M  f0ad3edce8e4    11 GB     6 weeks ago
qwen3:30b-a3b-q4_K_M            2ee832bc15b5    18 GB     6 weeks ago
gemma3:27b-it-qat               29eb0b9aeda3    18 GB     7 weeks ago
qwen2.5:latest                  845dbda0ea48    4.7 GB    7 weeks ago
PetrosStav/gemma3-tools:27b     45d2118fab07    17 GB     7 weeks ago
phi4:latest                     ac896e5b8b34    9.1 GB    7 weeks ago
deepseek-r1:latest              0a8c26691023    4.7 GB    7 weeks ago
qwq:latest                      009cb3f08d74    19 GB     7 weeks ago
gemma3:27b                      a418f5838eaf    17 GB     7 weeks ago
```

./ollama ps



```
NAME                  ID              SIZE      PROCESSOR    UNTIL
deepseek-r1:latest    0a8c26691023    6.0 GB    100% GPU     3 minutes from now
NAME                  ID              SIZE      PROCESSOR    UNTIL
deepseek-r1:latest    0a8c26691023    6.0 GB    100% GPU     3 minutes from now
Done
```

Chat with model

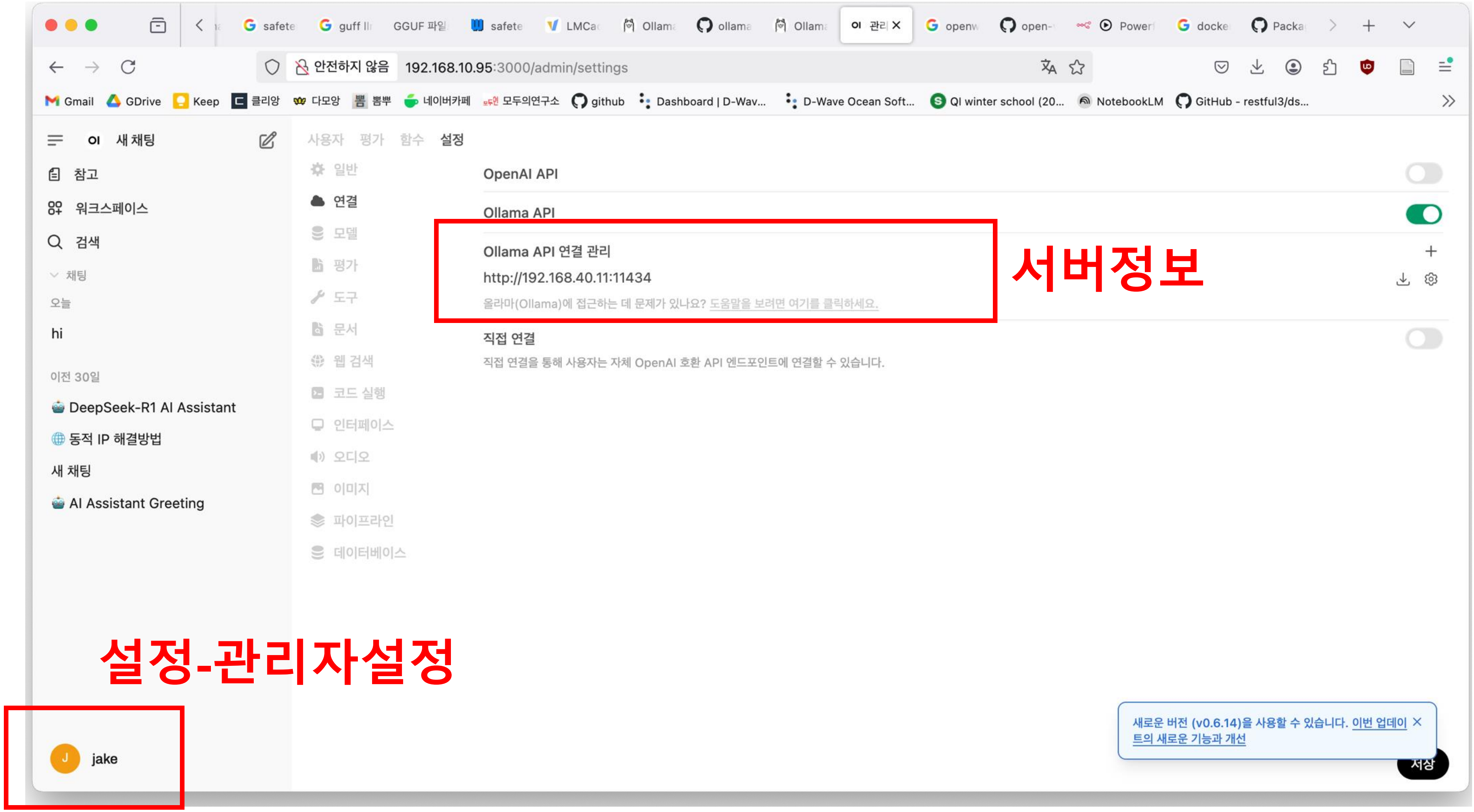


```
time=2025-06-13T20:55:23.114+09:00 level=INFO source=server.go:630 msg="llama runner started in 133.36 seconds"
[GIN] 2025/06/13 - 20:55:31 | 200 |          2m23s |    192.168.10.95 | POST     "/api/chat"
[GIN] 2025/06/13 - 20:55:38 | 200 |    6.52010889s |    192.168.10.95 | POST     "/api/chat"
[GIN] 2025/06/13 - 20:55:42 | 200 |   3.853083795s |    192.168.10.95 | POST     "/api/chat"
```

Open WebUI

- https://github.com/open-webui/open-webui
- https://pypi.org/project/open-webui/
- docker
- https://github.com/open-webui/open-webui/pkgs/container/open-webui

UI connect to Ollama server



192.168.10.95:3000/admin/settings

OpenAI API

Ollama API

Ollama API 연결 관리

http://192.168.40.11:11434

올라마(Ollama)에 접근하는 데 문제가 있나요? 도움말을 보려면 여기를 클릭하세요.

직접 연결

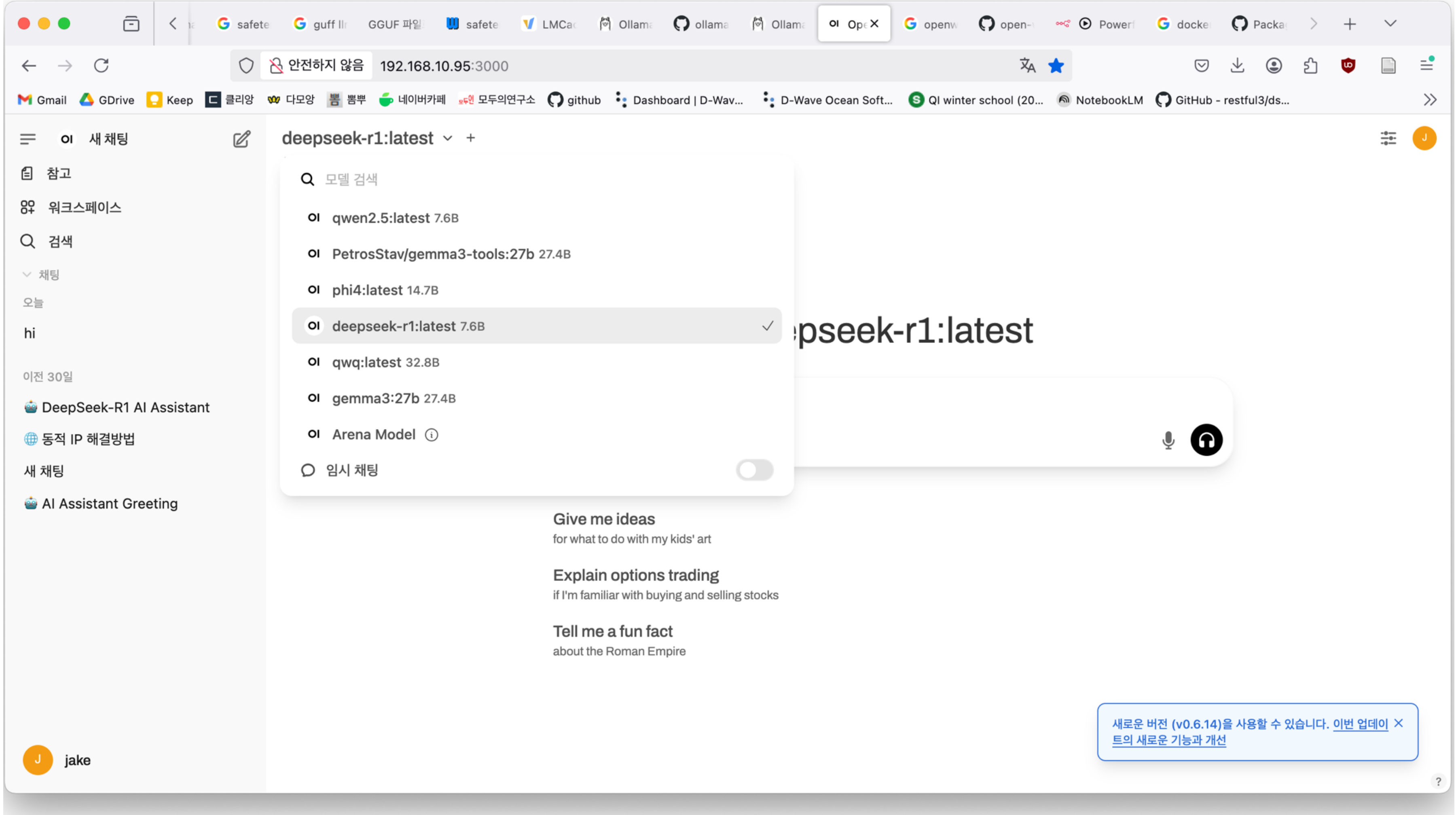
직접 연결을 통해 사용자는 자체 OpenAI 호환 API 엔드포인트에 연결할 수 있습니다.

새로운 버전 (v0.6.14)을 사용할 수 있습니다. 이번 업데이트의 새로운 기능과 개선

서버정보

설정-관리자설정

Select Model (Run model in Ollama server)



안전하지 않음 192.168.10.95:3000

deepseek-r1:latest

pseek-r1:latest

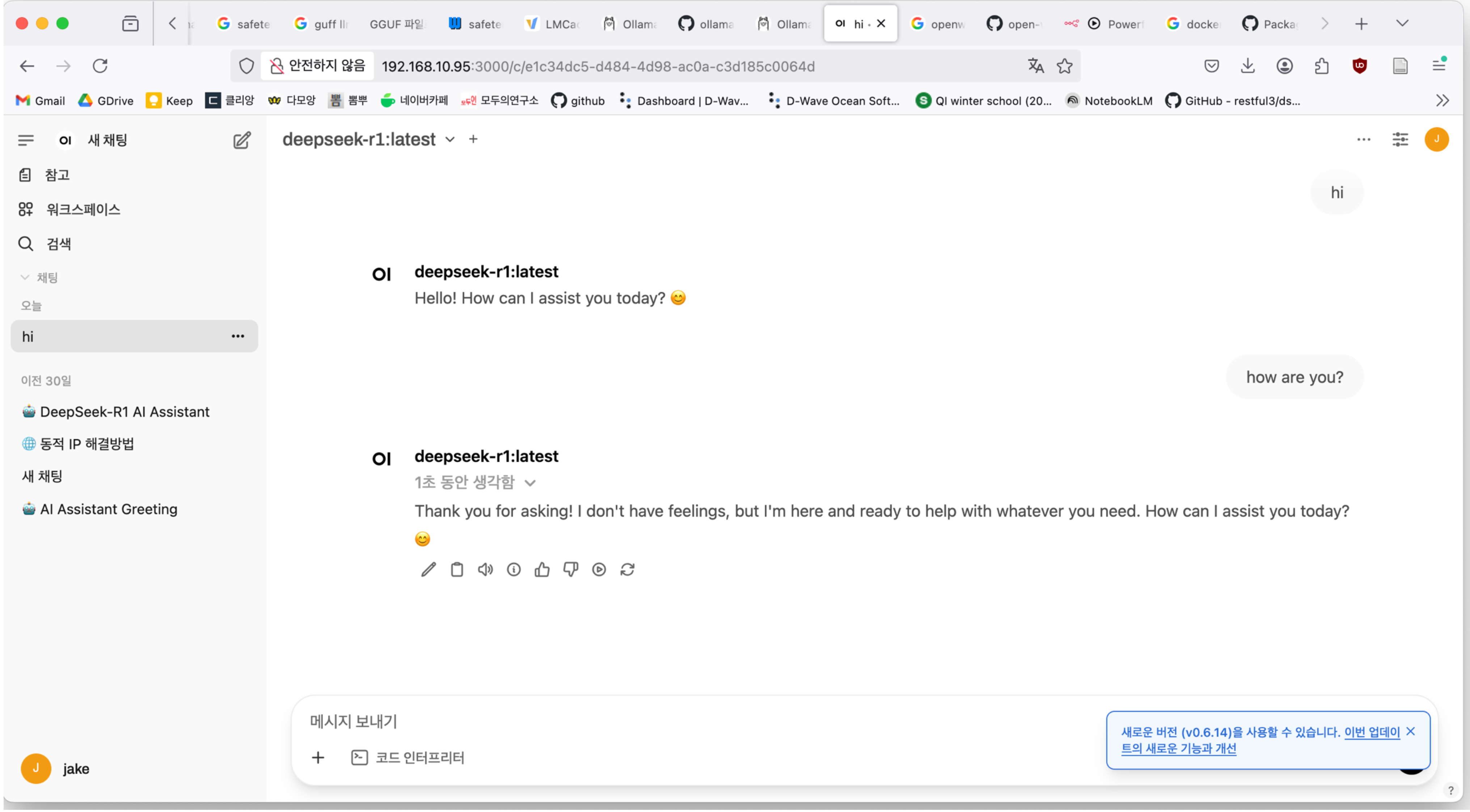
Give me ideas
for what to do with my kids' art

Explain options trading
if I'm familiar with buying and selling stocks

Tell me a fun fact
about the Roman Empire

새로운 버전 (v0.6.14)을 사용할 수 있습니다. 이번 업데이트의 새로운 기능과 개선

Chat!



192.168.10.95:3000/c/e1c34dc5-d484-4d98-ac0a-c3d185c0064d

deepseek-r1:latest

hi

deepseek-r1:latest

Hello! How can I assist you today? 😊

how are you?

deepseek-r1:latest

1초 동안 생각함

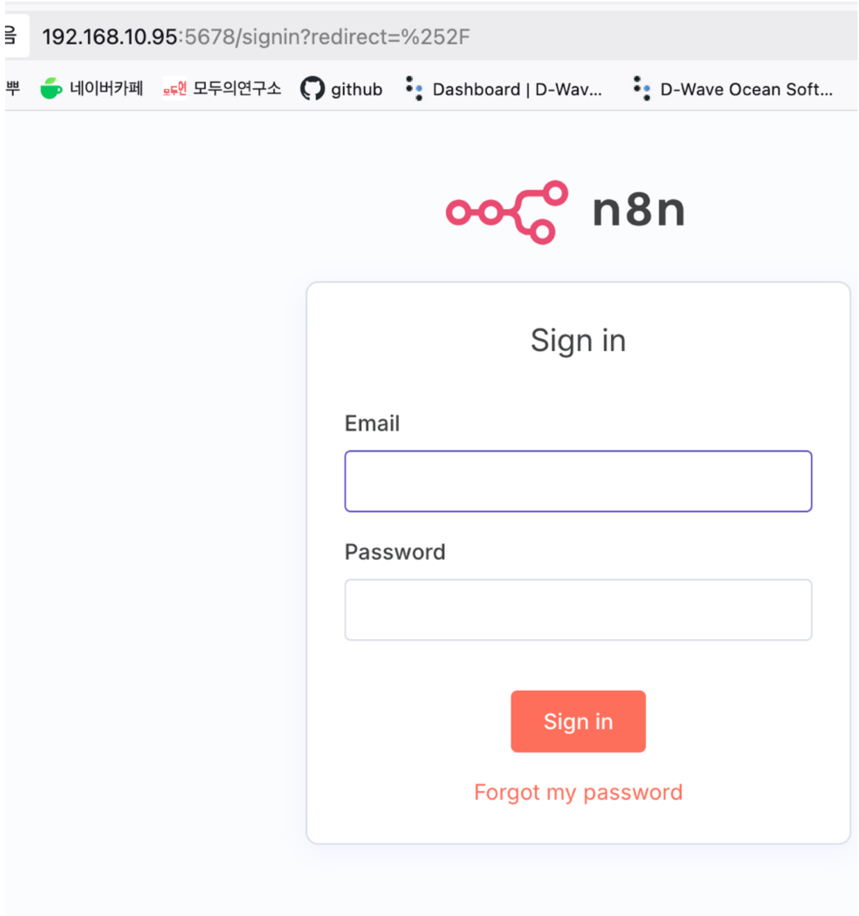
Thank you for asking! I don't have feelings, but I'm here and ready to help with whatever you need. How can I assist you today? 😊

새로운 버전 (v0.6.14)을 사용할 수 있습니다. 이번 업데이트의 새로운 기능과 개선

n8n

- n8n Self-hosting kit
- https://github.com/n8n-io/n8n-hosting/blob/main/docker-compose/withPostgres/docker-compose.yml
- docker-compose up -d
- docker-compose stop

Self hosting n8n



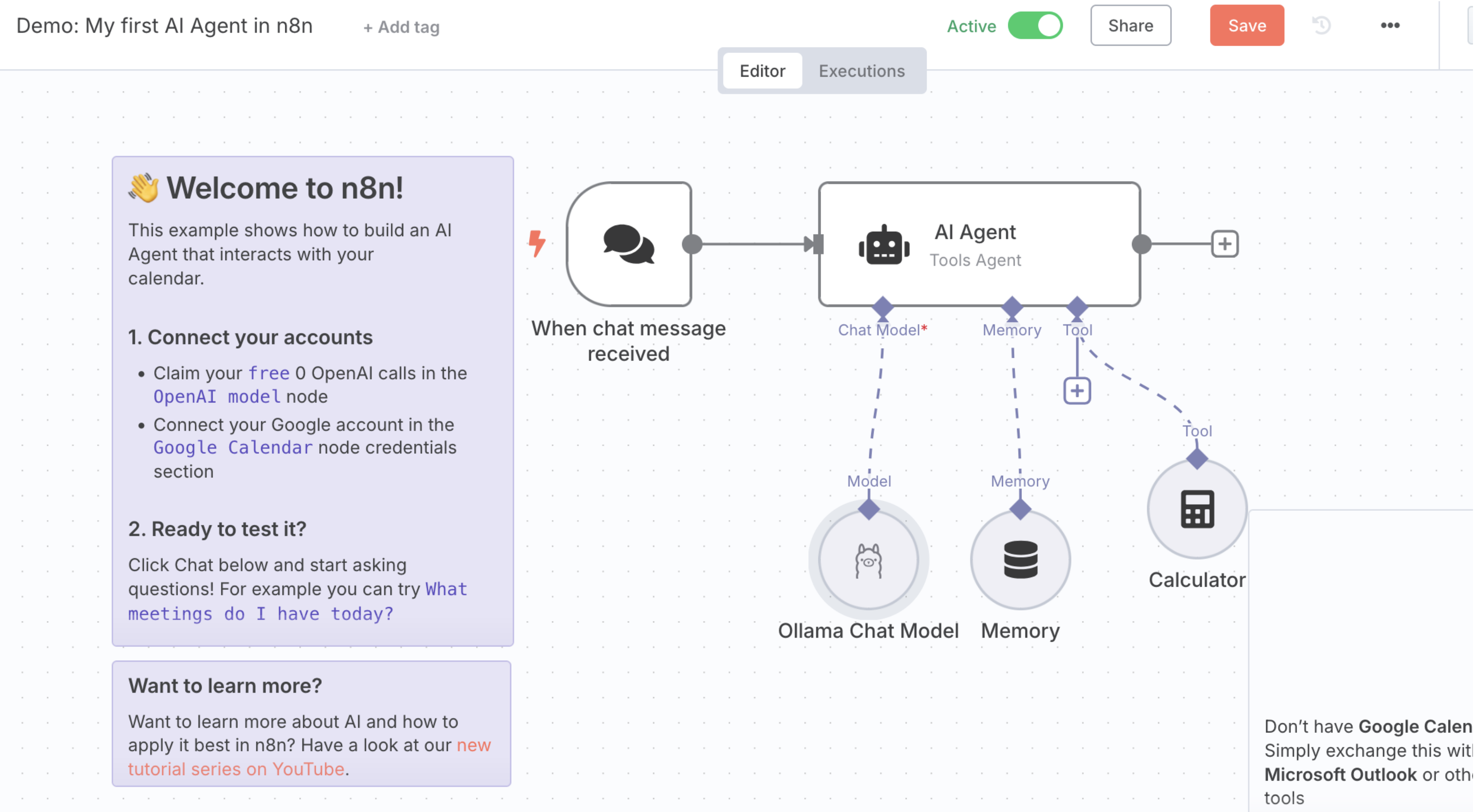
192.168.10.95:5678/signin?redirect=%252F

Sign in

Email

Password

n8n workflow example



👋 Welcome to n8n!

This example shows how to build an AI Agent that interacts with your calendar.

1. Connect your accounts

- Claim your `free` 0 OpenAI calls in the `OpenAI model` node
- Connect your Google account in the `Google Calendar` node credentials section

2. Ready to test it?

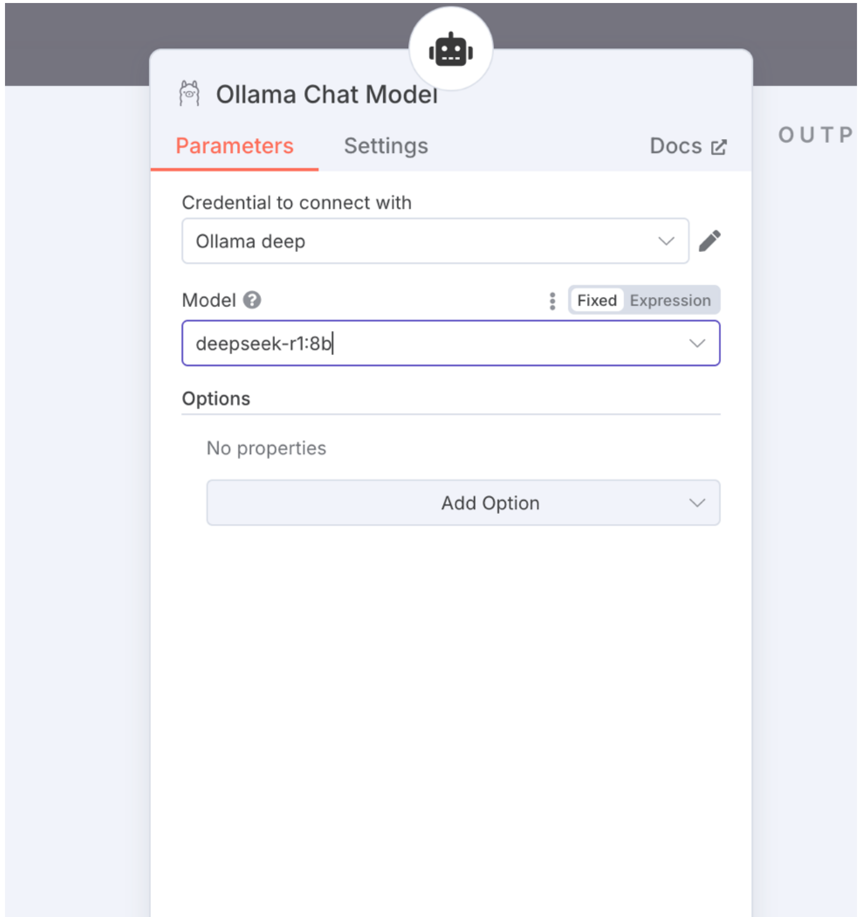
Click Chat below and start asking questions! For example you can try `What meetings do I have today?`

Want to learn more?

Want to learn more about AI and how to apply it best in n8n? Have a look at our new tutorial series on YouTube.

Don't have **Google Calen**
Simply exchange this wit
Microsoft Outlook or oth
tools

Set server and select model



Ollama Chat Model

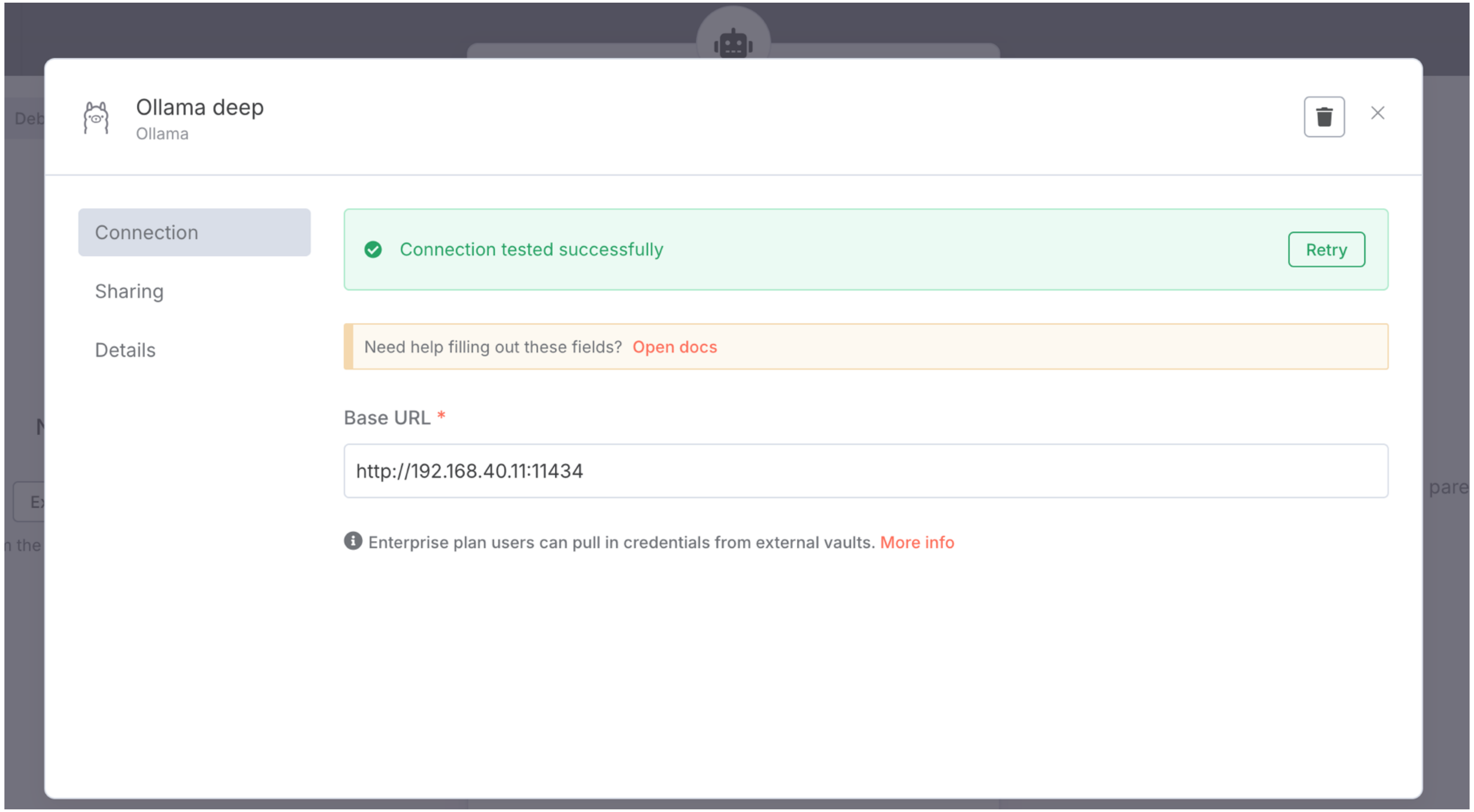
Credential to connect with: Ollama deep

Model: deepseek-r1:8b

Options

No properties

Set server



Ollama deep

Ollama

Connection tested successfully

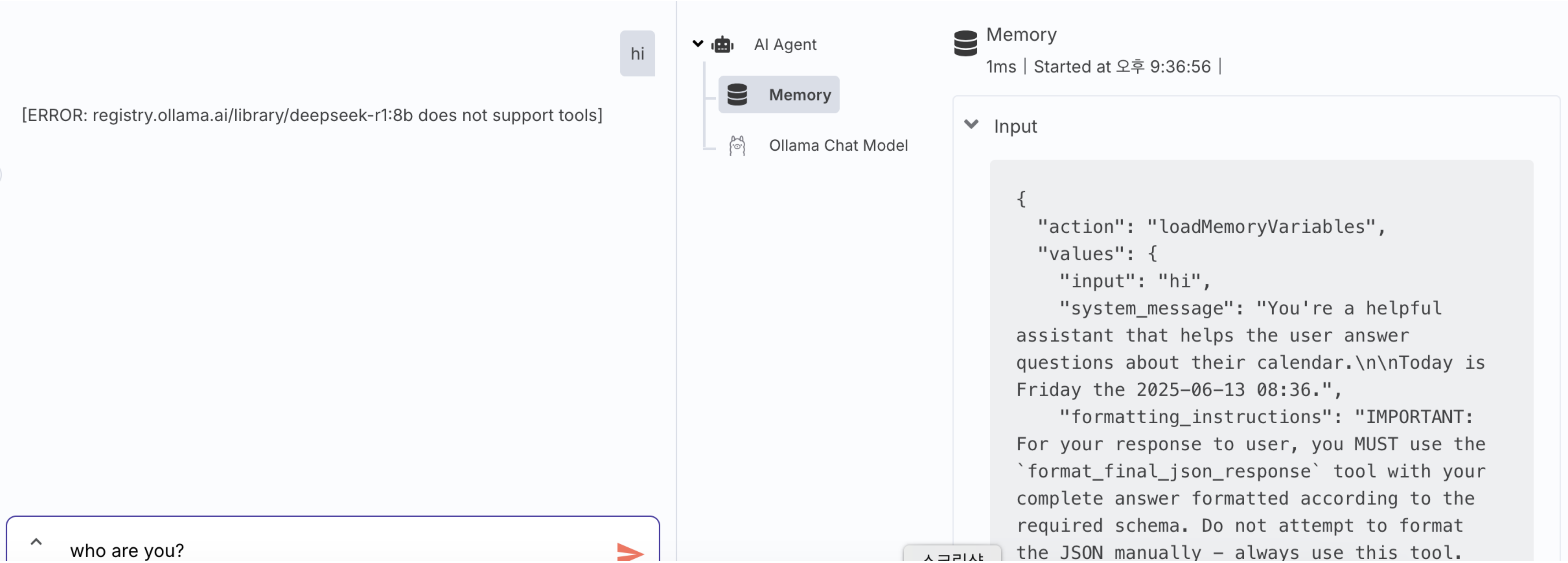
Need help filling out these fields? Open docs

Base URL *

http://192.168.40.11:11434

Enterprise plan users can pull in credentials from external vaults. More info

Chat and response



hi

[ERROR: registry.ollama.ai/library/deepseek-r1:8b does not support tools]

who are you?

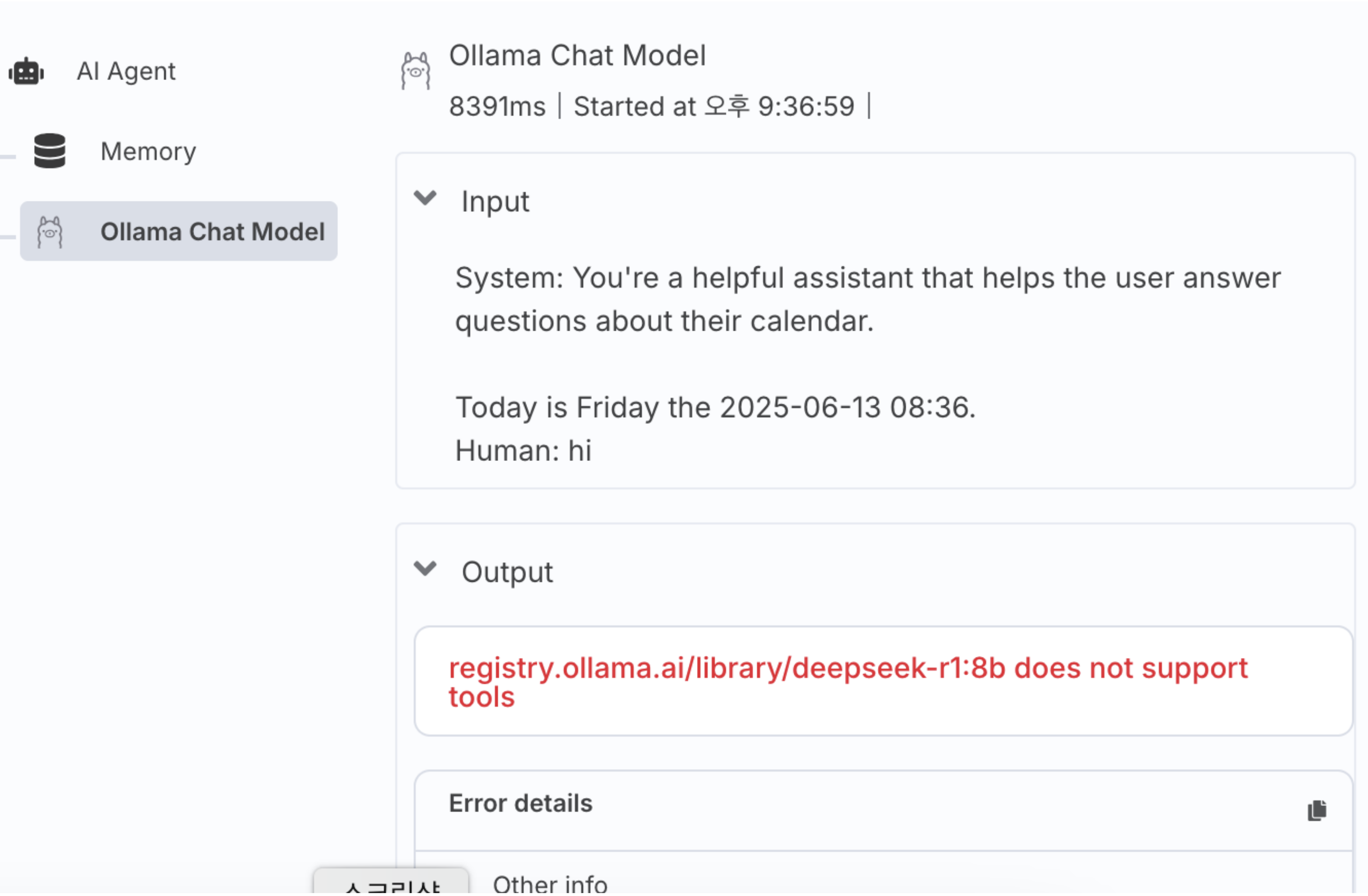
Memory

1ms | Started at 오후 9:36:56 |

Input

```
{
  "action": "loadMemoryVariables",
  "values": {
    "input": "hi",
    "system_message": "You're a helpful
assistant that helps the user answer
questions about their calendar.\n\nToday is
Friday the 2025-06-13 08:36.",
    "formatting_instructions": "IMPORTANT:
For your response to user, you MUST use the
`format_final_json_response` tool with your
complete answer formatted according to the
required schema. Do not attempt to format
the JSON manually - always use this tool.
```

Log of chatting



Ollama Chat Model

8391ms | Started at 오후 9:36:59 |

Input

System: You're a helpful assistant that helps the user answer questions about their calendar.

Today is Friday the 2025-06-13 08:36.
Human: hi

Output

registry.ollama.ai/library/deepseek-r1:8b does not support tools

Error details

Other info