A Journey on Building AI Datacenter at Home

Affan Basalamah

IDNOG 10

2025

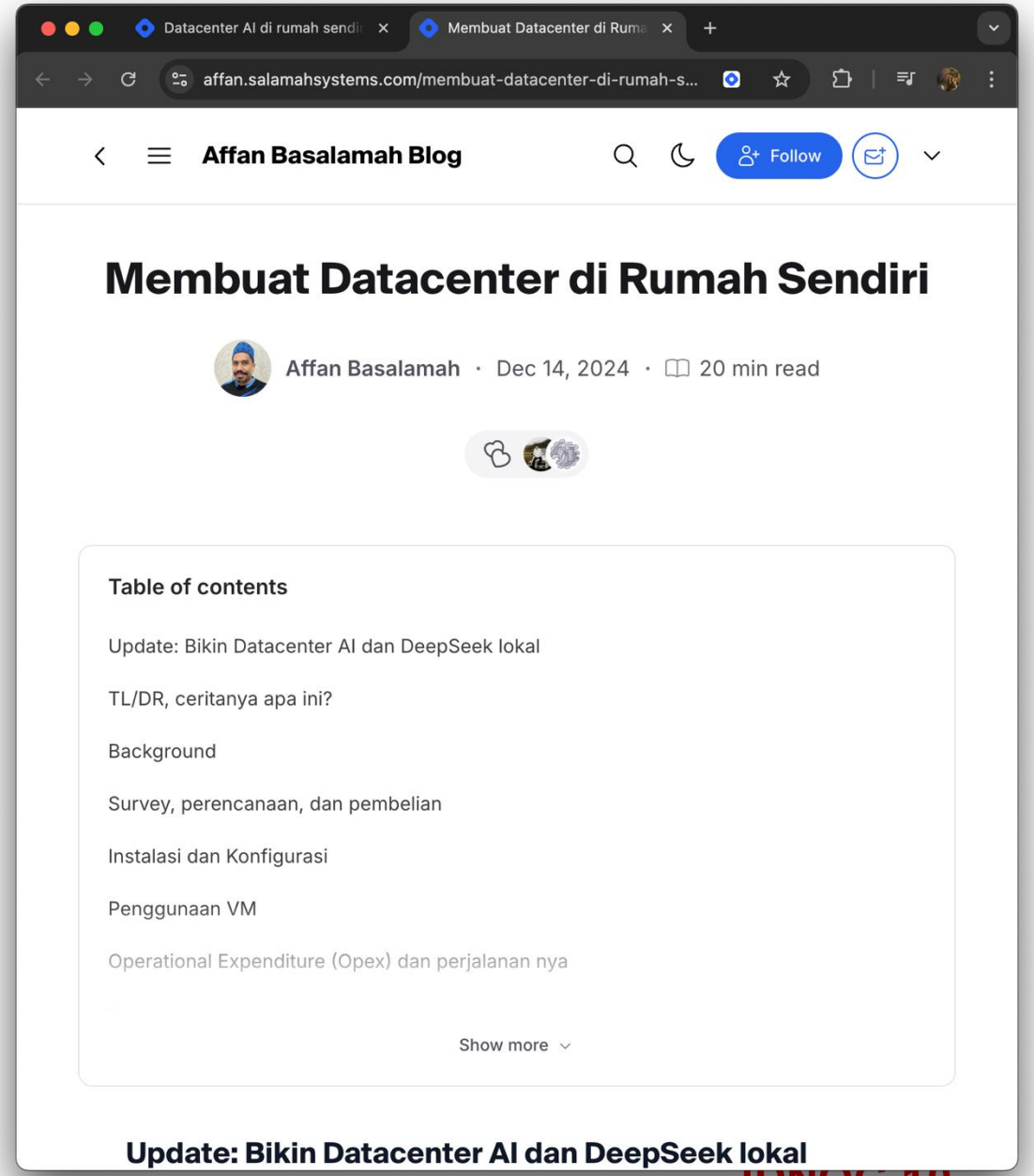
whoami

- Affan Basalamah
- @affanzbasalamah (X)
- affan@salamahsystems.com
- https://s.id/salamahsystems.com

- Consultant
- APNIC Community Trainer

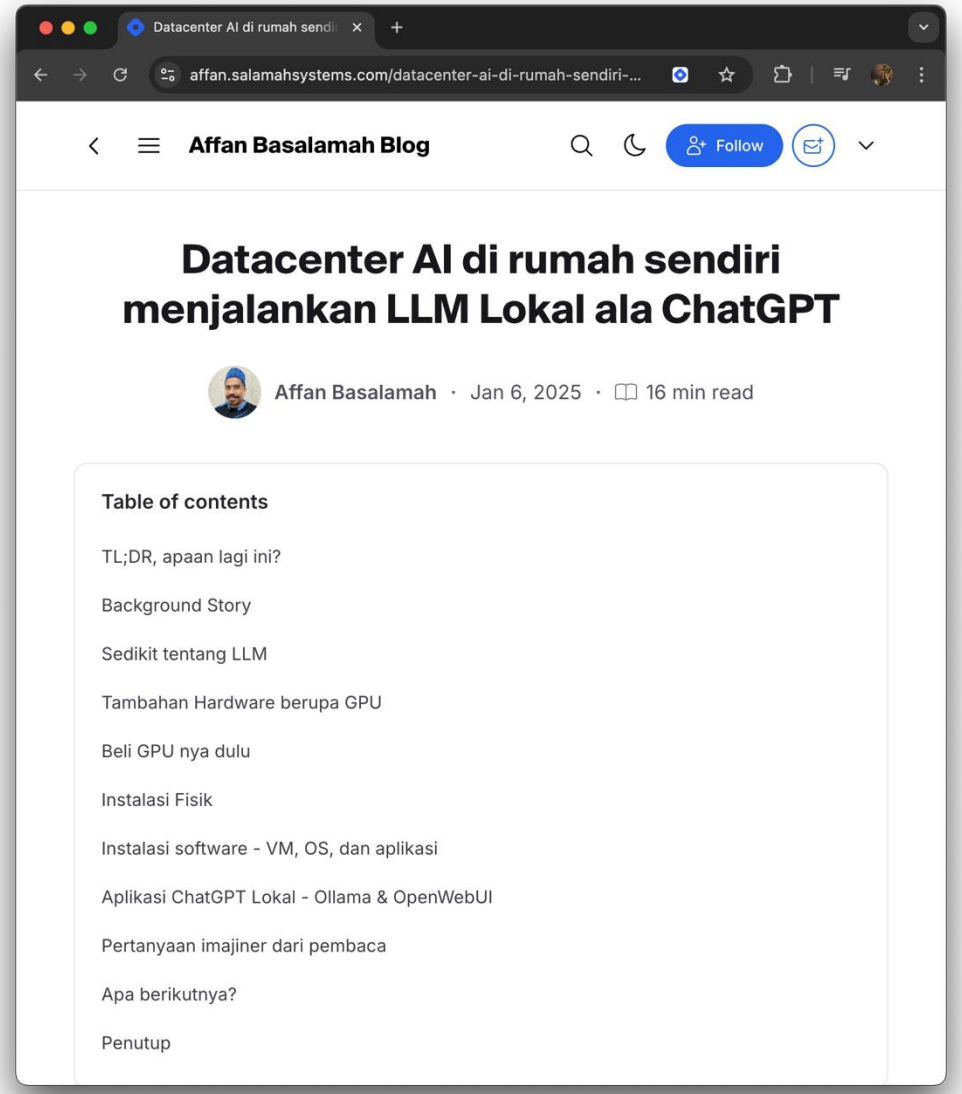
Story begins

- As a consultant, I want to have my own homelab to create PoC easily
  - PNETLAB, Containerlab
  - DNS, NMS, etc.
  - Local k8s, etc.
- And wrote a blog post about it: https://affan.salamahsystems.com/membuat-datacenter-di-rumah-sendiri



Rise of opensource AI and local AI homelab

- In 2024 every homelab youtubers create their own local AI
- Allow me to create one more case study for datacenter at home
- Experiment by installing GPU on server
- Tried the local AI chatbot, works
- Create another blogpost: https://affan.salamahsystems.com/datacenter-ai-di-rumah-sendiri-menjalankan-llm-lokal-ala-chatgpt



This is my journey building AI Datacenter at Home

Today’s Discussion

- What is AI that we’re talking about
- What can you do with AI
- How you build AI on your own datacenter at home

What is an AI?

AI as Large Language Mode (LLM)

- Usually referred to as *Foundational Model*

- Given the instruction, it can generate information (text) or multiple information (text, image, video: multimodal)

Source: O'Reilly Media – AI Engineering

Types of LLM in use

Model license:

- Closed model
  - Can only be accessed by API
- Open weight model
  - Can be downloaded only for weight
- Open source model
  - Can be downloaded all its data

LLM size & properties:

- Measured by billions (B) of parameters
- Base foundational model
  - Very large size (hundreds B of parameter, hundreds of GB size)
- Distilled model
  - Compressed the data from teacher model to student model
  - Between (1.5, 3, 7, 14, 32, 70 B of parameter, size between 1,5 to 80 GB size)
- Quantized model
  - Reduces the precision of model

LLM that you may know and use

Open weight model:

- DeepSeek-R1
- Google Gemma

Open source model:

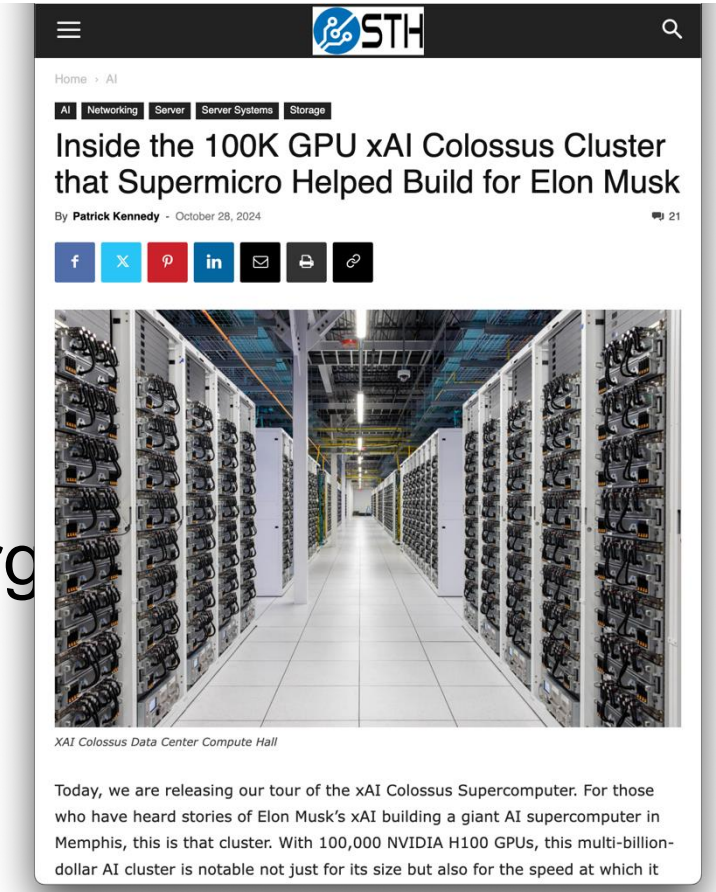
- Meta AI LLama 3.1
- Alibaba Qwen 3
- Mistral AI Mixtral-8x22B

Closed model:

- OpenAI GPT
  - (4o, 4.1, o3, o4, etc.)
- Anthropic Claude
  - (Opus 4, Sonnet 4, Haiku 3.5)
- Google Gemini
  - (2.5 Flash, 2.5 Pro, etc.)
- xAI Grok

How to use the model

- **Training**
  - Build the foundational model from scratch
  - Training foundational model require very large of data
- **Inferencing**
  - Use the trained model for AI application
  - We will use the trained model for AI datacenter at home
- **Finetuning**
  - Some model need to be finetuned for changing the characteristics of results



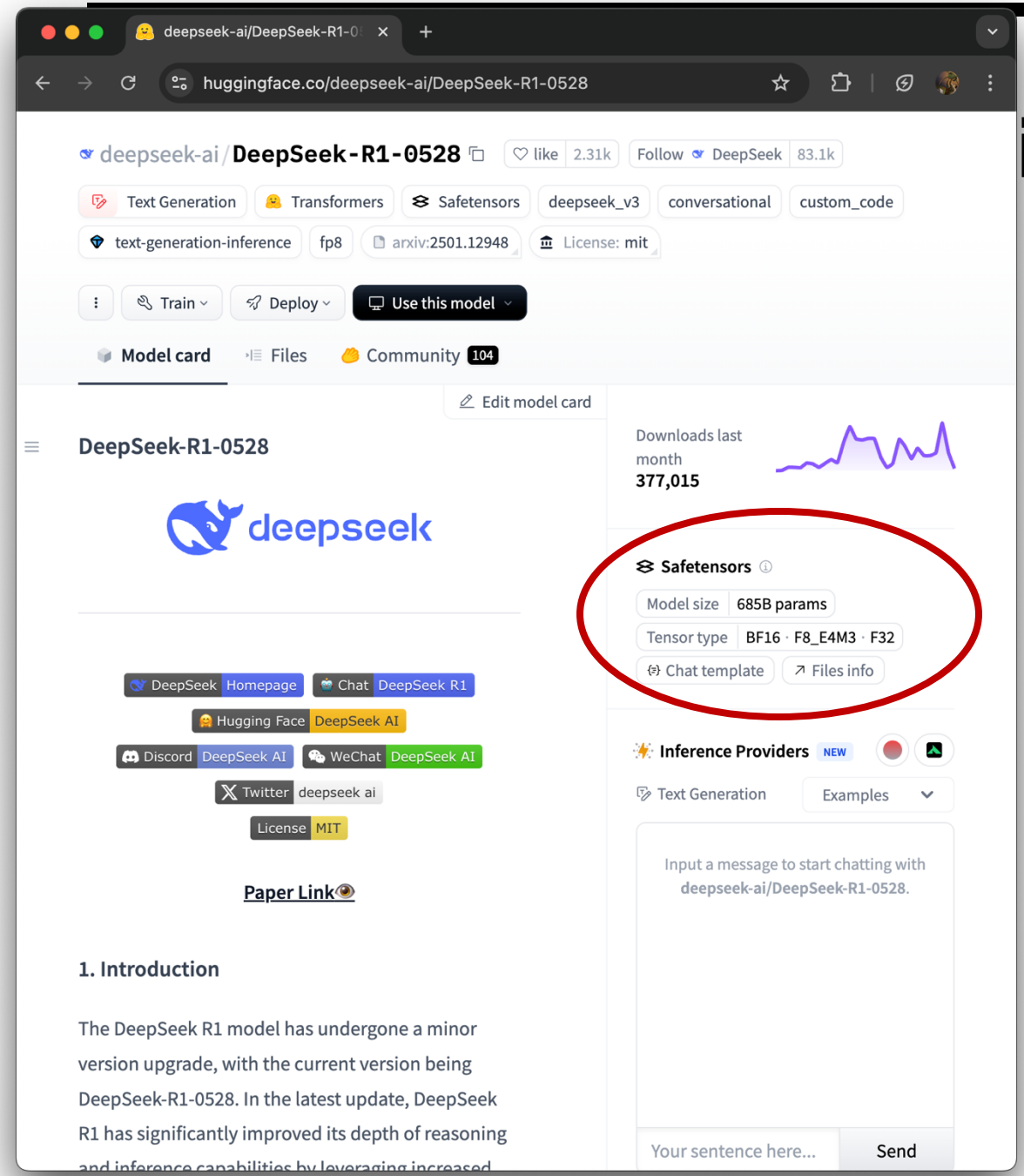
How to host your own AI?

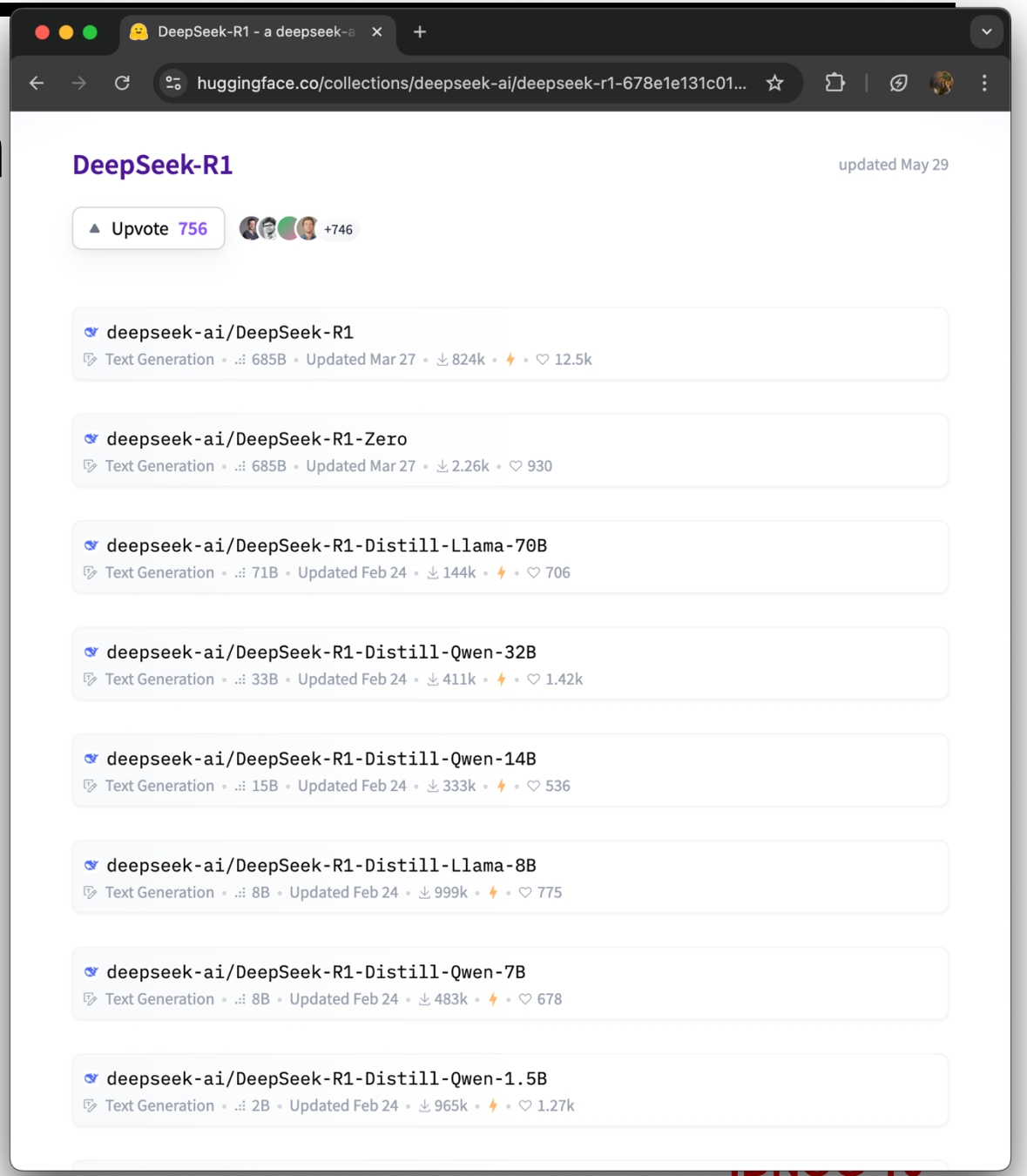
LLM needs to be run in systems with GPU (most of the time)

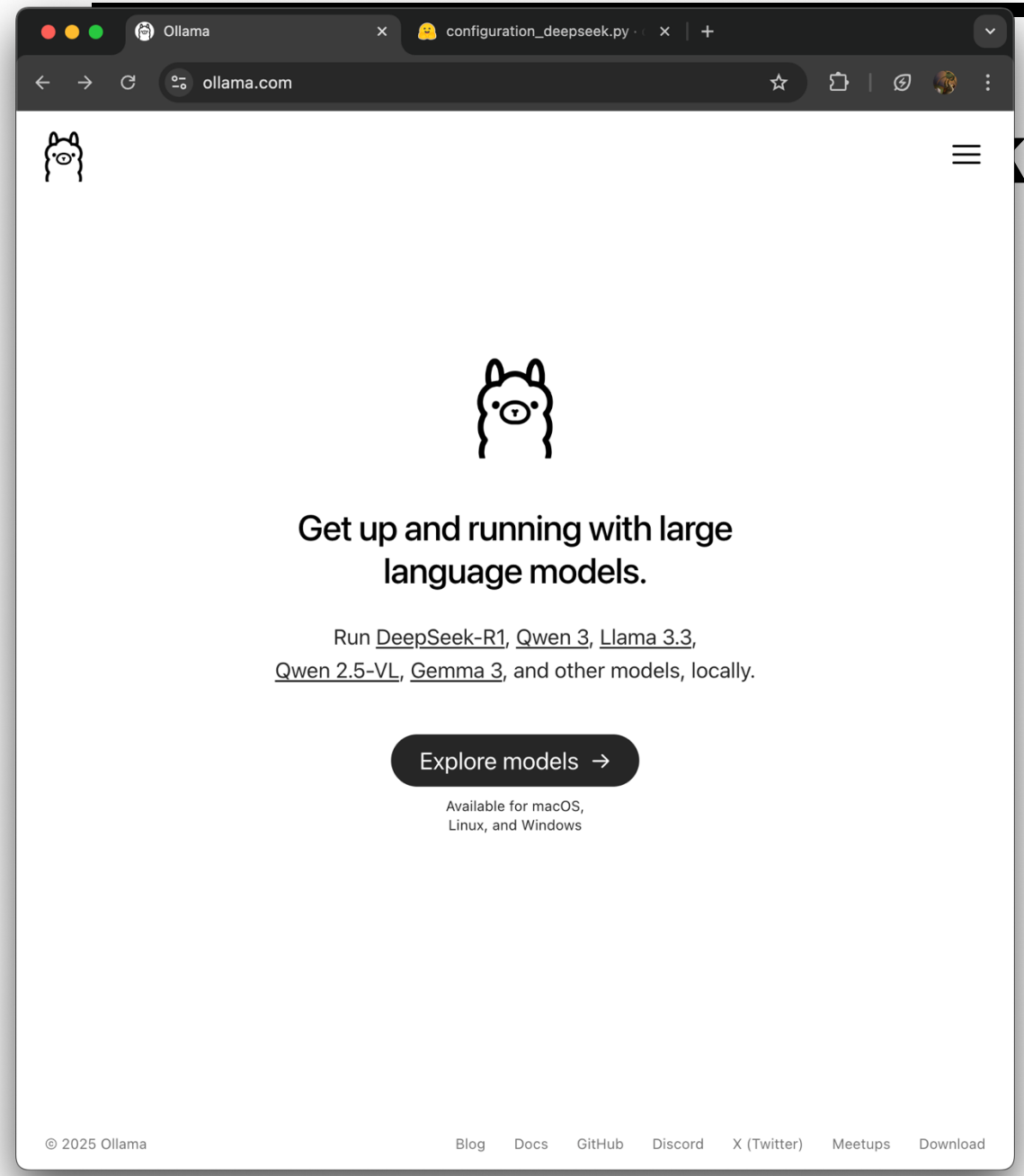
- LLM require heavy computation, basically matrix multiplication and tensor operation executed in parallel
- LLM weight in the system will be loaded to GPU memory to run
- Size of GPU memory (VRAM) will determine the size of the LLM
  - Bigger the model size, bigger the GPU VRAM needed
  - Finetuning require 2x-3x GPU VRAM compared with inferencing
  - Training require more, usually clustering the GPU systems
- There are more than that, but I try to simplify things
  - FLOPS, HBM size & speed, etc.

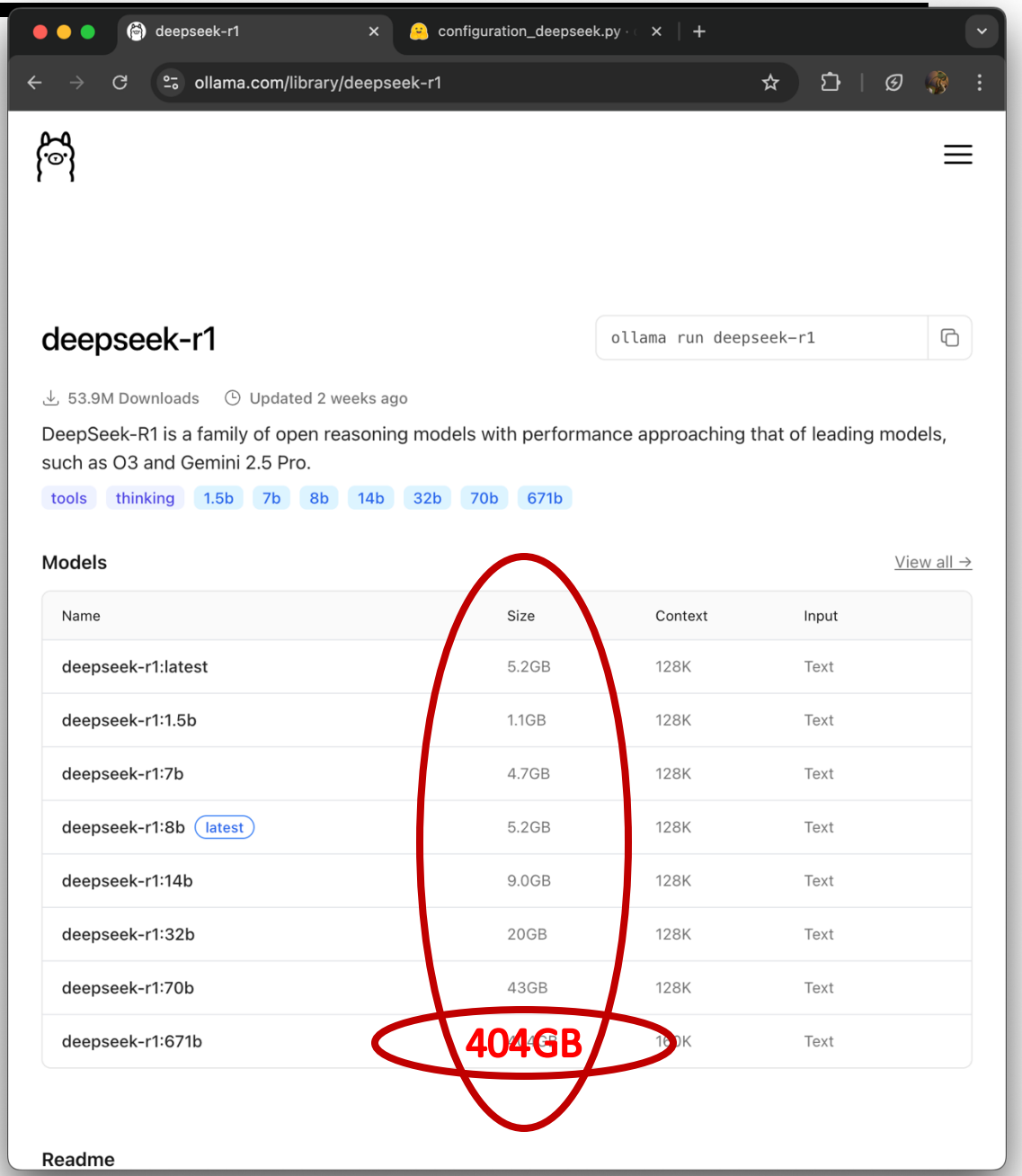
Where can I download LLM?

- You cannot download closed LLMs
- You can only download opensource/openweight LLMs
- https://huggingface.co/ (HF) is the place to go
  - Place for complete opensource/openweight AI model
  - Model file usually with **.gguf** extension (*GPT-generated universal format*)
- For first timer, the fastest route is to use https://ollama.com/
  - Runtime to simplify LLM download in local system
  - Combine with OpenWeb-UI to create local ChatGPT
  - Can be used by AI apps (e.g. n8n, Dify) as local AI model


huggingface.co/deepseek-ai/DeepSeek-R1-0528
deepseek-ai / DeepSeek-R1-0528
like 2.31k
Follow DeepSeek 83.1k
Text Generation
Transformers
Safetensors
deepseek_v3
conversational
custom_code
text-generation-inference
fp8
arxiv:2501.12948
License: mit
Train
Deploy
Use this model
Model card
Files
Community 104
Edit model card
DeepSeek-R1-0528
deepseek
DeepSeek Homepage
Chat DeepSeek R1
Hugging Face DeepSeek AI
Discord DeepSeek AI
WeChat DeepSeek AI
Twitter deepseek ai
License MIT
Paper Link
1. Introduction
The DeepSeek R1 model has undergone a minor version upgrade, with the current version being DeepSeek-R1-0528. In the latest update, DeepSeek R1 has significantly improved its depth of reasoning
Downloads last month
377,015
Safetensors
Model size 685B params
Tensor type BF16 · F8_E4M3 · F32
Chat template
Files info
Inference Providers NEW
Text Generation
Examples
Input a message to start chatting with deepseek-ai/DeepSeek-R1-0528.
Your sentence here...
Send


huggingface.co/collections/deepseek-ai/deepseek-r1-678e1e131c01...
DeepSeek-R1
updated May 29
Upvote 756
+746
deepseek-ai/DeepSeek-R1
Text Generation • 685B • Updated Mar 27 • 824k • 12.5k
deepseek-ai/DeepSeek-R1-Zero
Text Generation • 685B • Updated Mar 27 • 2.26k • 930
deepseek-ai/DeepSeek-R1-Distill-Llama-70B
Text Generation • 71B • Updated Feb 24 • 144k • 706
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
Text Generation • 33B • Updated Feb 24 • 411k • 1.42k
deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
Text Generation • 15B • Updated Feb 24 • 333k • 536
deepseek-ai/DeepSeek-R1-Distill-Llama-8B
Text Generation • 8B • Updated Feb 24 • 999k • 775
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
Text Generation • 8B • Updated Feb 24 • 483k • 678
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
Text Generation • 2B • Updated Feb 24 • 965k • 1.27k


Ollama
configuration_deepseek.py
ollama.com
Get up and running with large language models.
Run DeepSeek-R1, Qwen 3, Llama 3.3, Qwen 2.5-VL, Gemma 3, and other models, locally.
Explore models →
Available for macOS, Linux, and Windows
© 2025 Ollama
Blog
Docs
GitHub
Discord
X (Twitter)
Meetups
Download

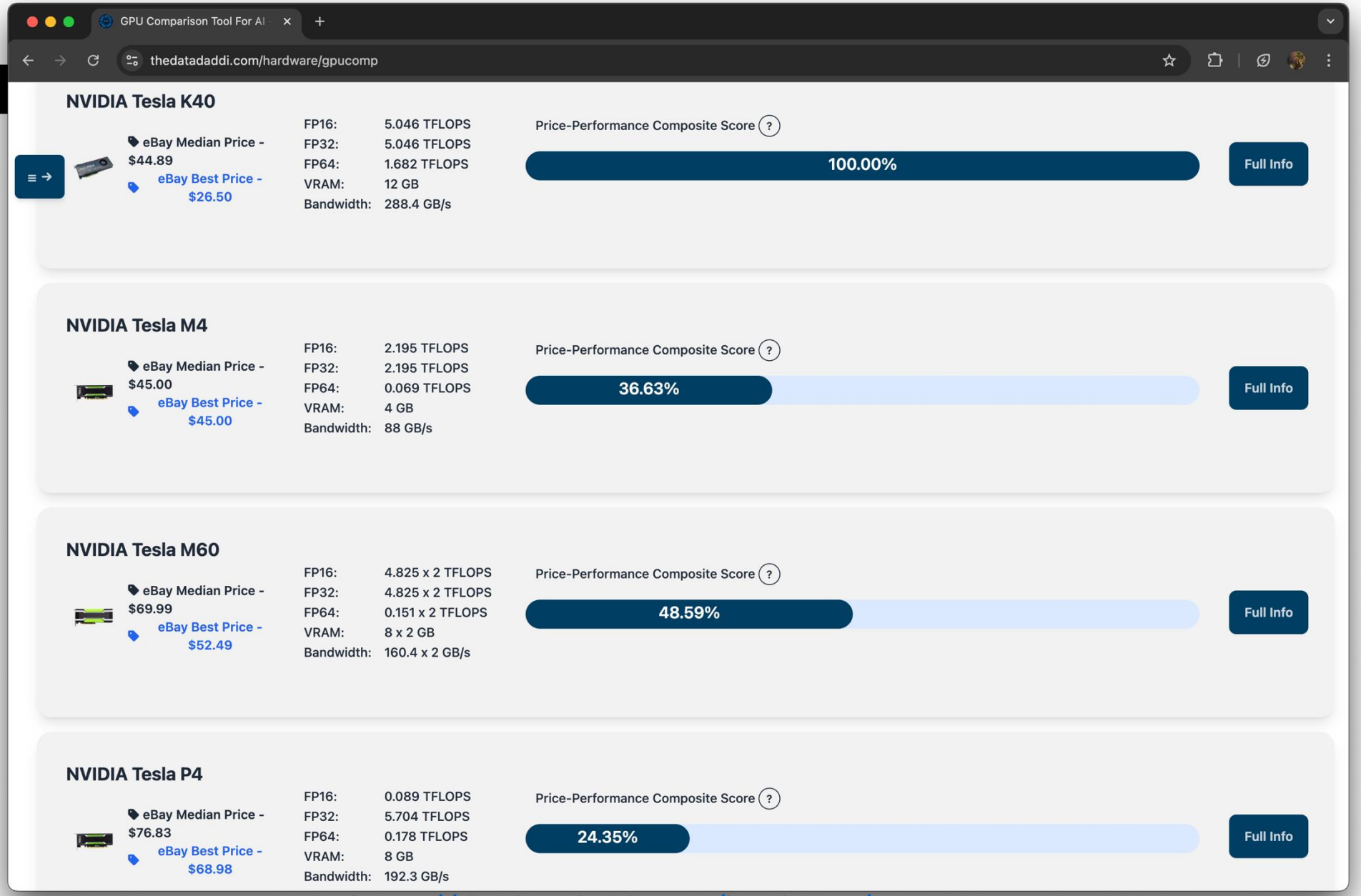

deepseek-r1
configuration_deepseek.py
ollama.com/library/deepseek-r1
deepseek-r1
ollama run deepseek-r1
53.9M Downloads
Updated 2 weeks ago
DeepSeek-R1 is a family of open reasoning models with performance approaching that of leading models, such as O3 and Gemini 2.5 Pro.
tools
thinking
1.5b
7b
8b
14b
32b
70b
671b
Models
View all →
Name
Size
Context
Input
deepseek-r1:latest
5.2GB
128K
Text
deepseek-r1:1.5b
1.1GB
128K
Text
deepseek-r1:7b
4.7GB
128K
Text
deepseek-r1:8b
latest
5.2GB
128K
Text
deepseek-r1:14b
9.0GB
128K
Text
deepseek-r1:32b
20GB
128K
Text
deepseek-r1:70b
43GB
128K
Text
deepseek-r1:671b
404GB
Text
Readme

IDNOG 10

Which LLMs require which GPU?

Model Size (full or distilled)	GPU type	Application Types
1.3B – 2B (e.g. **DistilGPT2, TinyLlama, Phi-2**)	4 – 6GB GPU or CPU only	Embedded AI, lightweight chatbot
7B (e.g. **LLaMA 2 7B, Mistral 7B, DeepSeek 7B, Gemma 7B**)	8GB GPU	General-purpose assistants, RAG, code help
13– 14B (e.g. **LLaMA 2 13B, StarCoder 14B**)	24 – 32GB GPU	Smarter assistants with better reasoning, multi-turn chat, dev aid
30 – 34B (e.g **LLaMA 2 30B, Mixtral, Yi-34B**)	48– 80GB GPU	Complex assistants with solid reasoning, agents, domain-specific
65 – 70B (e.g. **LLaMA 2 65B, Falcon 180B, Claude 2/3 class**)	96 – 128GB GPU	GPT-4-class apps, enterprise copilots

NVid



https://thedatadaddi.com/hardware/gpucomp

NVIDIA Tesla P4

Designed for AI inference and edge deployments, the Tesla P4 is compact and power-efficient, featuring 8GB of GDDR5 memory, and is ideal for recommendation systems, video processing, and low-latency inference tasks.

General Information

Number Of GPUs	1
GPU Variant	GP104-895-A1
Architecture	Pascal
Process Size	16 nm
Release Date	2016-09-13
Generation	Tesla Pascal
Bus Interface	PCIe 3.0 x16
NVLink	Yes

Performance Rates

Pixel Rate	71.3 GPixel/s
Texture Rate	178.2 GTexel/s
FP16	0.089 TFLOPS
FP32	5.704 TFLOPS
FP64	0.178 TFLOPS

Memory Specifications

VRAM Size	8 GB
VRAM Type	GDDR5
VRAM Bus	256 bit

Clock Speeds

Base	

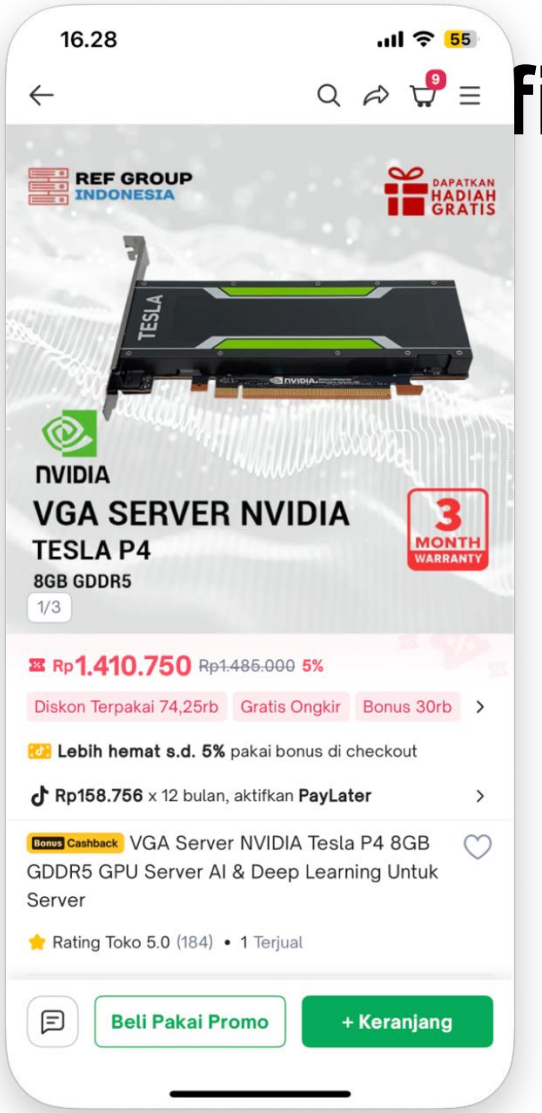
GPU Prices Over Time

Start Date: 13/11/2024 End Date: 20/07/2025

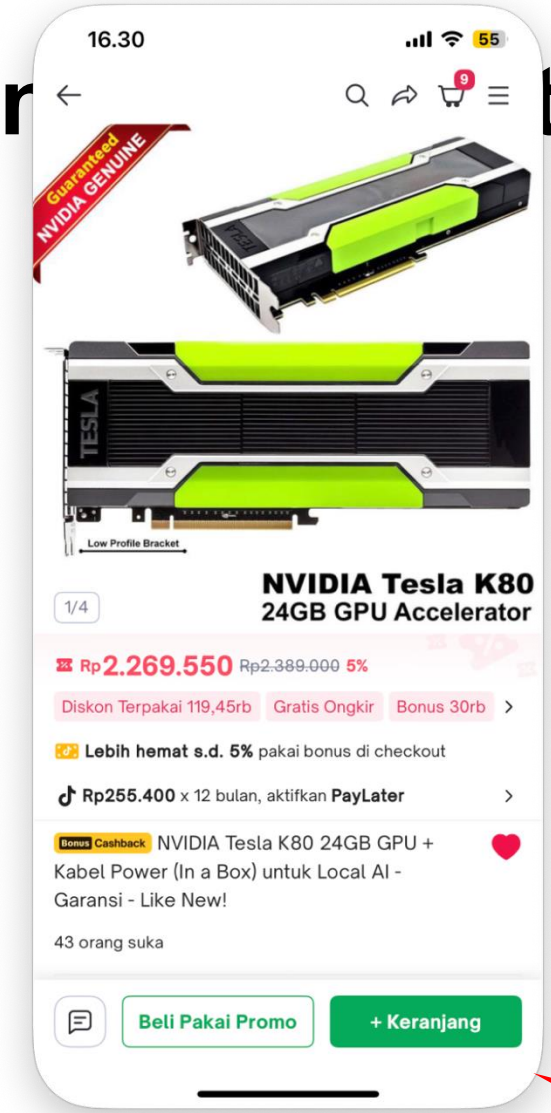
Apply Filter

Last Week Price History

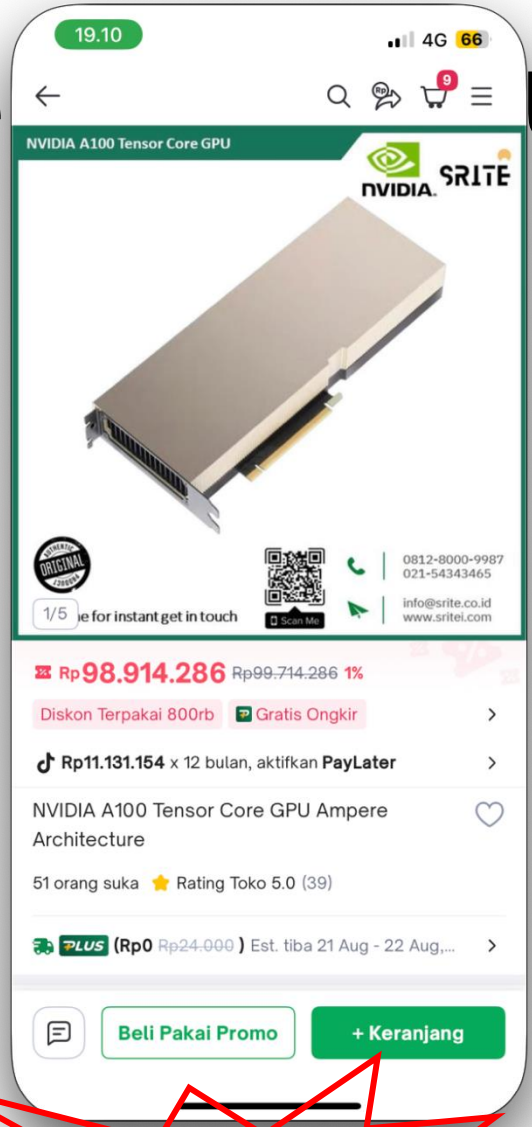
Date Checked	Median Price	Best Price	Link
7/20/2025	$76.83	$68.98	ebay
7/19/2025	$76.88	$68.98	ebay
7/18/2025	$76.88	$68.98	ebay



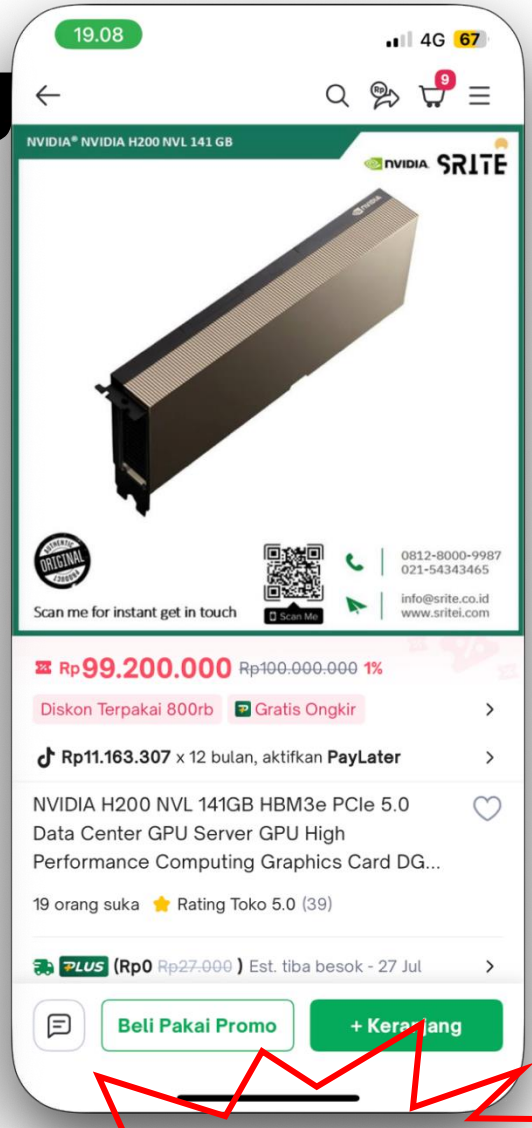
Tesla P4 – 8GB
Rp. 1.410.750



Tesla K80 – 2 x 12GB
Rp. 2.269.550



NVidia A100 – 80GB
Rp. 98.914.286

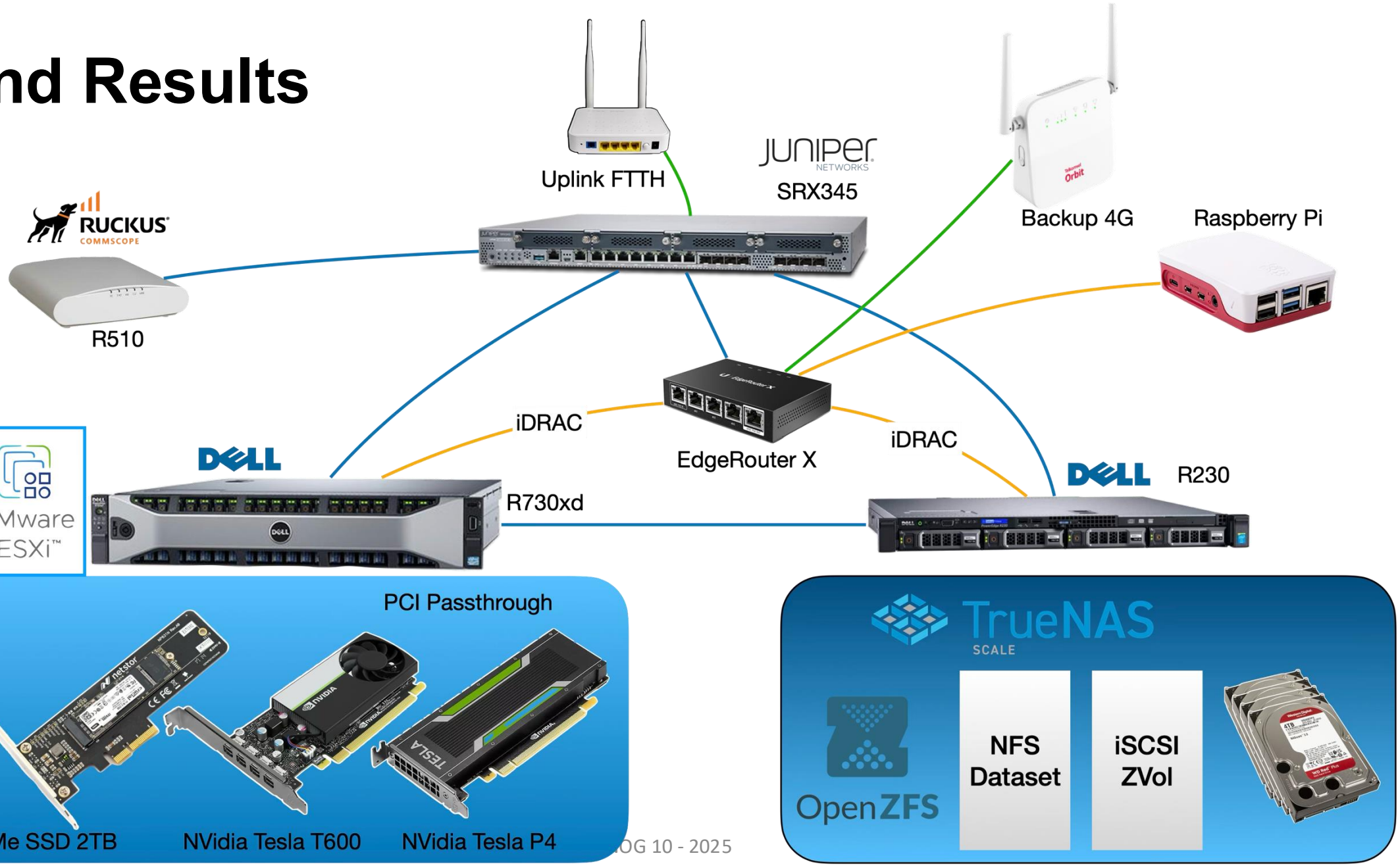


NVidia H200 – 141GB
Rp. 99.200.000

IDMOG 10

How I do it in my home

End Results


Uplink FTTH
JUNIPER NETWORKS
SRX345
Backup 4G
Raspberry Pi
RUCKUS COMMSCOPE
R510
iDRAC
iDRAC
EdgeRouter X
DELL
R730xd
DELL
R230
VMware ESXi™
PCI Passthrough
NVMe SSD 2TB
NVidia Tesla T600
NVidia Tesla P4
TrueNAS SCALE
OpenZFS
NFS Dataset
iSCSI ZVol


FIPS BLNK
PowerEdge R230

Choose hardware for the systems

- For starter, better check aftermarket GPU
- Any PC or servers with a lot of RAM and PCIe slot
- Power depends on the GPU
  - Small form factor (SFF) GPU usually have lower power (75W), can be powered by PCIe slot itself
  - Dual width GPU usually have bigger power (300-500W), needs additional power cables

Tesla K80 – 24GB
Double width GPU

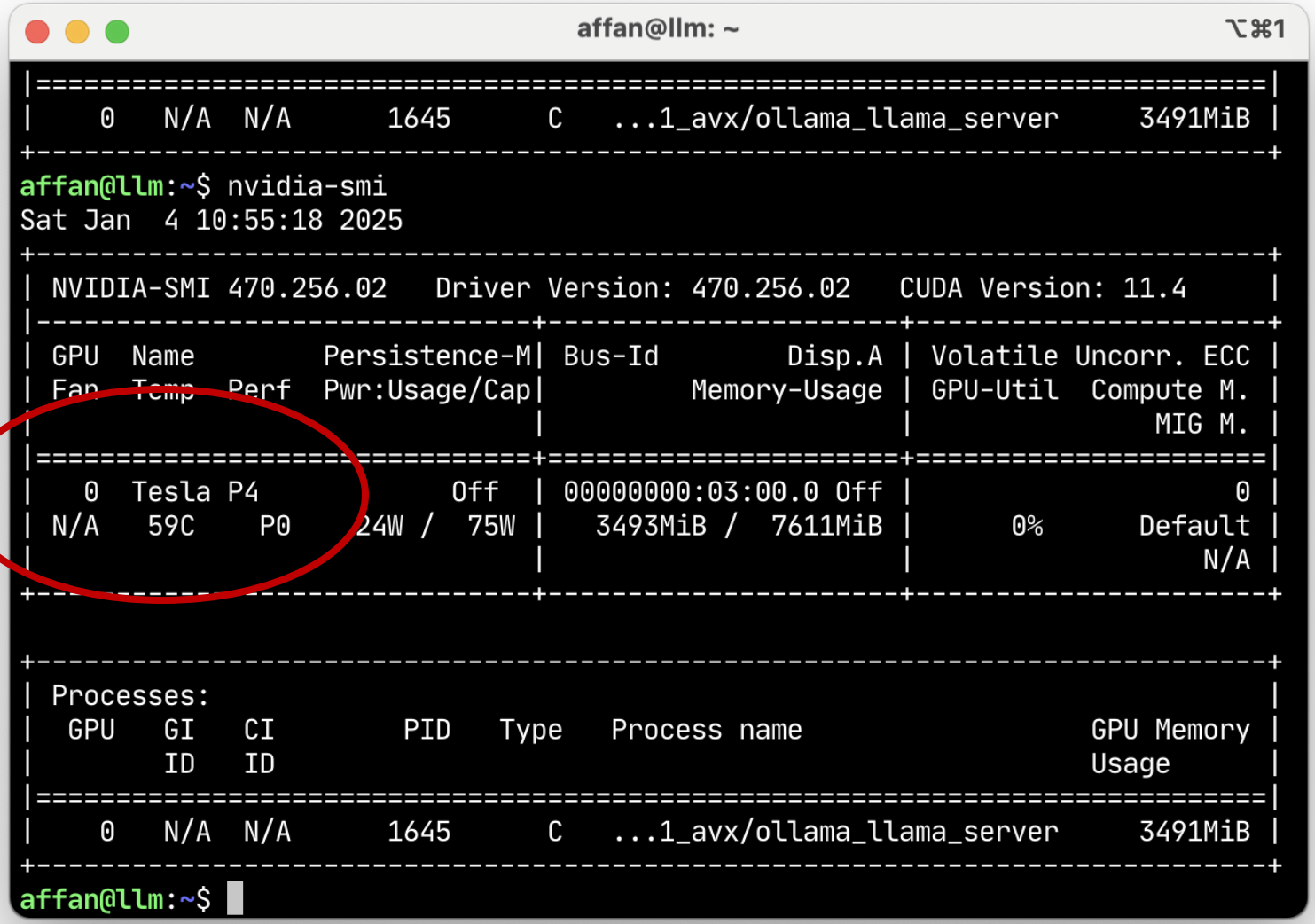
Tesla P4 – 8GB
SFF Single width GPU

Software Stack to run AI Datacenter

- For desktop use use LMStudio https://lmstudio.ai/
  - Windows, macOS, Linux
- For server use Linux distro (myself use Ubuntu)
- Usually use Docker for easier installation
- GPU driver installed on OS
  - Learn how to install NVidia driver on Ubuntu
- Faster results: use integrated Ollama + OpenWeb-UI to create chatbot https://openwebui.com/
- Or better: use Ollama separately

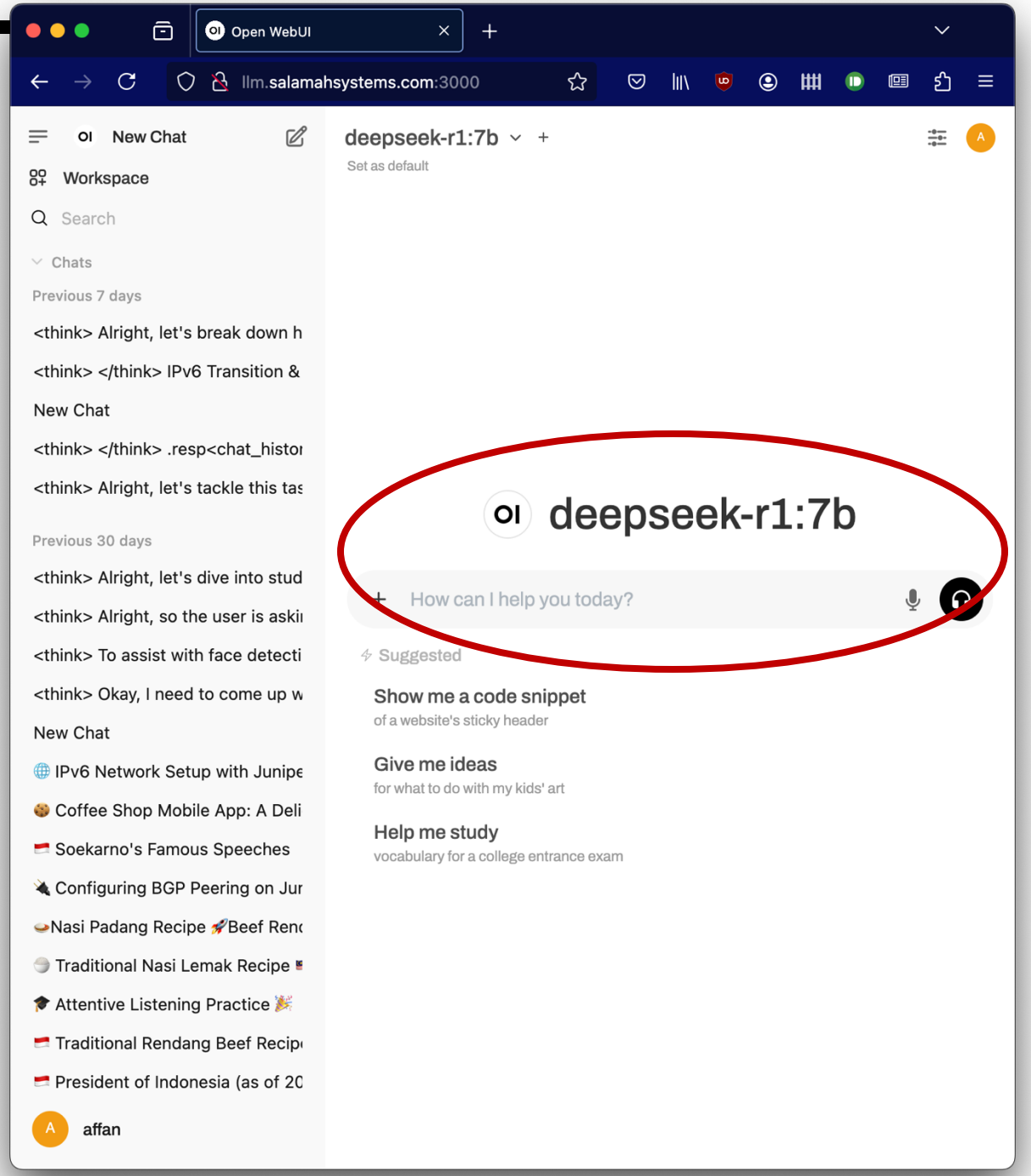
Install NVidia Driver

- Sometimes installation becomes a bit intricate
- "`sudo nvidia-smi`" and see your GPU model to begin install Ollama



Open WebUI

- Frontend for local LLM
- Suddenly you have local chatbot + RAG interface
  - Lots of function, check the docs
- Docker installation integrated with Ollama or separated

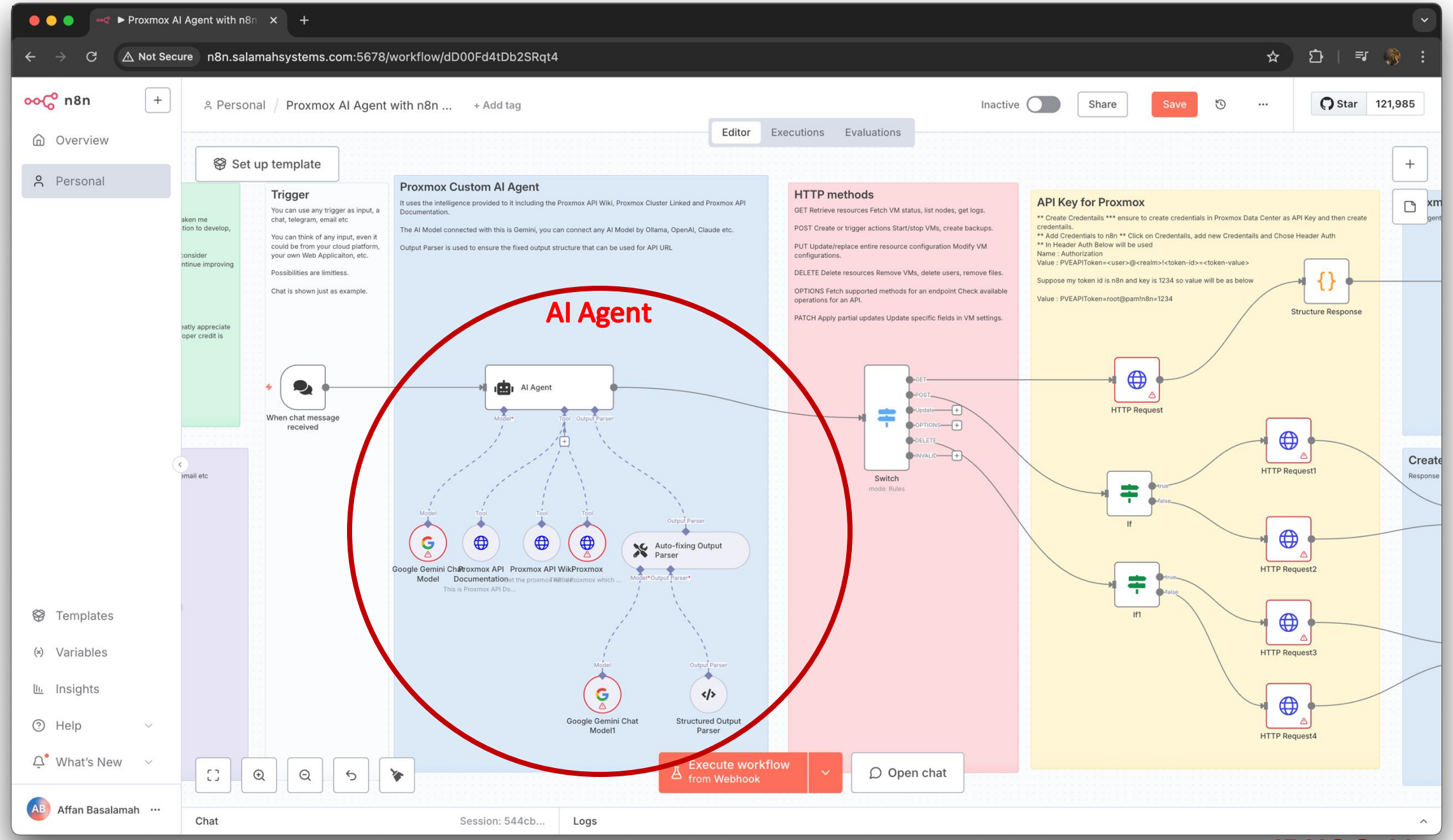


AI apps running on local LLM installation

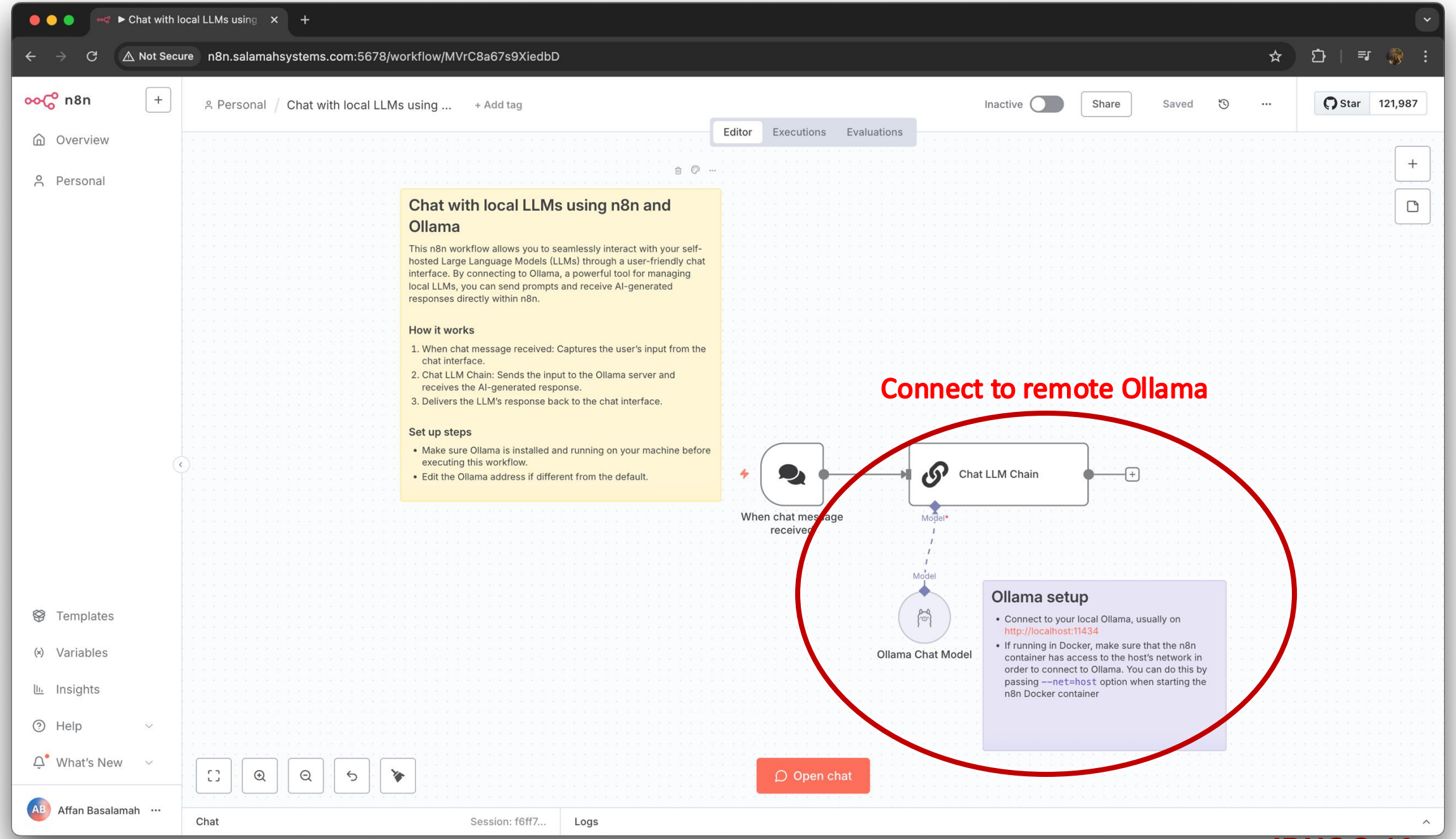
- Local AI means private or secret document stay home and don’t go to the remote LLM provider

AI Apps:

- Chatbot (ChatGPT local clone) with OpenWeb-UI
- Retrieve knowledge from documents using RAG
- Workflow automation from n8n accessing local Ollama instances


Proxmox AI Agent with n8n
Not Secure
n8n.salamahsystems.com:5678/workflow/dD00Fd4tDb2SRqt4
n8n
Overview
Personal
Templates
Variables
Insights
Help
What's New
Affan Basalamah
Personal / Proxmox AI Agent with n8n ...
+ Add tag
Inactive
Share
Save
Star 121,985
Editor
Executions
Evaluations
Set up template
Trigger
You can use any trigger as input, a chat, telegram, email etc
You can think of any input, even it could be from your cloud platform, your own Web Applicaiton, etc.
Possibilities are limitless.
Chat is shown just as example.
When chat message received
Proxmox Custom AI Agent
It uses the intelligence provided to it including the Proxmox API Wiki, Proxmox Cluster Linked and Proxmox API Documentation.
The AI Model connected with this is Gemini, you can connect any AI Model by Ollama, OpenAI, Claude etc.
Output Parser is used to ensure the fixed output structure that can be used for API URL
AI Agent
Google Gemini Chat Model
Proxmox API Documentation
Proxmox API Wiki
Auto-fixing Output Parser
Google Gemini Chat Model1
Structured Output Parser
HTTP methods
GET Retrieve resources Fetch VM status, list nodes, get logs.
POST Create or trigger actions Start/stop VMs, create backups.
PUT Update/replace entire resource configuration Modify VM configurations.
DELETE Delete resources Remove VMs, delete users, remove files.
OPTIONS Fetch supported methods for an endpoint Check available operations for an API.
PATCH Apply partial updates Update specific fields in VM settings.
Switch
mode: Rules
API Key for Proxmox
** Create Credentails *** ensure to create credentials in Proxmox Data Center as API Key and then create credentails.
** Add Credentials to n8n ** Click on Credentails, add new Credentails and Chose Header Auth
** In Header Auth Below will be used
Name : Authorization
Value : PVEAPIToken=<user>@<realm>!<token-id>=<token-value>
Suppose my token id is n8n and key is 1234 so value will be as below
Value : PVEAPIToken=root@pam!n8n=1234
Structure Response
HTTP Request
If
If1
HTTP Request1
HTTP Request2
HTTP Request3
HTTP Request4
AI Agent
Execute workflow from Webhook
Open chat
Chat
Session: 544cb...
Logs

IDNOG 10


Chat with local LLMs using
Not Secure
n8n.salamahsystems.com:5678/workflow/MVrC8a67s9XiedbD
n8n
Overview
Personal
Templates
Variables
Insights
Help
What's New
Affan Basalamah
Personal / Chat with local LLMs using ...
+ Add tag
Inactive
Share
Saved
Star 121,987
Editor
Executions
Evaluations
Chat with local LLMs using n8n and Ollama
This n8n workflow allows you to seamlessly interact with your self-hosted Large Language Models (LLMs) through a user-friendly chat interface. By connecting to Ollama, a powerful tool for managing local LLMs, you can send prompts and receive AI-generated responses directly within n8n.
How it works
1. When chat message received: Captures the user's input from the chat interface.
2. Chat LLM Chain: Sends the input to the Ollama server and receives the AI-generated response.
3. Delivers the LLM's response back to the chat interface.
Set up steps
• Make sure Ollama is installed and running on your machine before executing this workflow.
• Edit the Ollama address if different from the default.
Connect to remote Ollama
When chat message received
Chat LLM Chain
Model*
Model
Ollama Chat Model
Ollama setup
• Connect to your local Ollama, usually on http://localhost:11434
• If running in Docker, make sure that the n8n container has access to the host's network in order to connect to Ollama. You can do this by passing --net=host option when starting the n8n Docker container
Open chat
Chat
Session: f6ff7...
Logs

IDNOG 10

The Journey Continues

Countless potential opportunities in local AI

- Lots of promises of network automation can be done with local AI help
- Automation tools (n8n) doesn’t yet have connector for popular network OS/appliances
- Opportunities for opensource centralized SoT (source of truth) (e.g. Netbox, Nautobot)
- Lots more…

Do you want to join me in this journey?

- Join me at https://s.id/salamahsystems.com
  - Fill me a little survey
  - Maybe join a group
  - Maybe you want to receive newsletter

Thanks!