

# Open-Webui & Ollama

Your local LLM  
infrastructure

By: Shawn Garrett

- Nerd
- Platform Engineer/DevOps
- Civo Ambassador

## 01 What & Why?

- Ollama
- Open-Webui
- Privacy & Security
- Opex vs Capex

## 02 Architecture

- Tldraw diagram
- Options Considered
- Actual Build
- Cloud OR on-prem

## 03 Features

- Web Search
- Coding Assistants
- Browser Addon
- Image Generation

## 04 Demo

- Explore
- Play
- Prompt

## 05 Questions

- Questions
- Thank You

## 06 References

- Links
- Resources to learn more

01

What & Why?

# Why?

## Privacy

- Deepseek-r1 top iPhone Free app
- OpenAi using your prompts to improve
- Data sovereignty

## Security

- Which 3rd parties are accessing knowingly/unknowingly?
- Unintentional corporate secrets leaked

## Cost

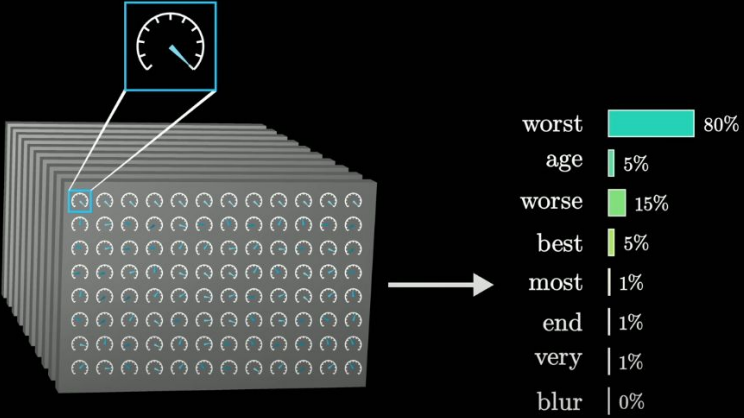
- **Subscription Services** Chatgpt, grok, claude, anthropic, perplexity, Microsoft CoPilot, Gemini
- CapEx vs. OpEx
- Human Costs

## EASE OF USE

- Find the right model for the right task

## Large Language Model

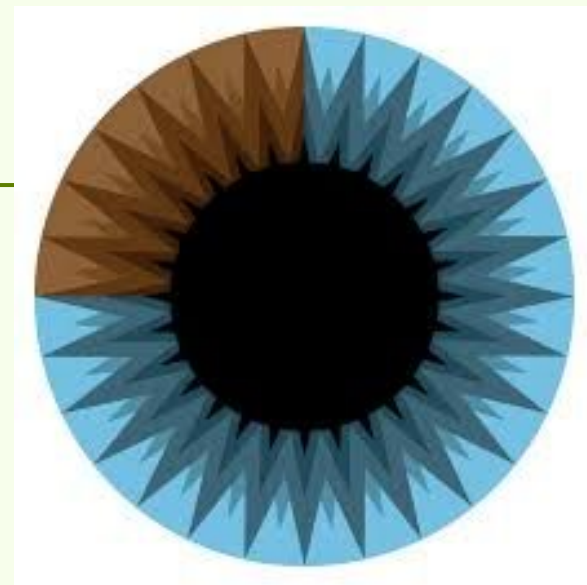
It was the best of times it was the \_\_\_\_\_ →



The diagram illustrates a Large Language Model (LLM) architecture. It shows a stack of layers, with a gauge icon indicating a specific layer. An arrow points from the stack to a list of words and their corresponding percentages:

worst	80%
age	5%
worse	15%
best	5%
most	1%
end	1%
very	1%
blur	0%
⋮	

2:10 / 7:57



- Large Language Models explained briefly
- Transformers (how LLMs work) explained visually | DL5
- How might LLMs store facts | DL7

Get up and  
running with  
Llama 3.3,  
DeepSeek-R1,  
Phi-4, Gemma 2,  
and other large  
language  
models.



```
curl -fsSL https://ollama.com/install.sh | sh
```

# ollama endpoint

```

: $ date && curl http://ollama.nerdery.lan/api/chat -d '{
  "model": "deepseek-r1:32b",
  "messages": [
    {
      "role": "user",
      "content": "why is the sky blue?"
    }
  ],
  "stream": false
}'
{"model": "deepseek-r1:32b", "created_at": "2025-02-18T02:52:28.303717885Z", "message": {"role": "assistant", "content": "\u003cthink\u003e\n\u003c/n\u003ckay, so I'm trying to figure out why the sky is blue. I remember in school they mentioned something about light and particles in the air, but I'm not entirely sure how it all works. Let me start by recalling what I know.\n\nFirst, sunlight is white, right? But when it hits the Earth's atmosphere, it changes somehow. Maybe it's because of the way the atmosphere interacts with the light. I think this has to do with something called scattering. Scattering... isn't that when light bounces off particles in the air?\n\nWait, there are different types of scattering. There's Rayleigh scattering and Mie scattering, if I remember correctly. So Rayleigh scattering happens when the particles are smaller than the wavelength of light, like molecules in the atmosphere. That must be what's happening because the atmosphere is made up mostly of nitrogen and oxygen molecules, which are pretty tiny.\n\nSo, sunlight is made up of different colors, each with their own wavelength. Red has a longer wavelength, and violet is shorter. Rayleigh scattering affects shorter wavelengths more, so blue light (which is shorter than red) gets scattered much more than red does. That's why the sky appears blue—it's all the scattered blue light coming from everywhere in the atmosphere.\n\nBut wait, at sunrise or sunset, the sky isn't blue anymore; it turns red or orange. Why is that? Maybe because when the sun is lower on the horizon, the sunlight has to pass through more of the Earth's atmosphere. The longer path means most of the blue light is scattered out, leaving behind the longer wavelengths like red and orange.\n\nOh, and if there's pollution or particles in the air, like during fog or after a wildfire, the sky might look different too. More particles can cause Mie scattering, which scatters all colors more equally, leading to white or even a yellowish sky depending on the particle size. That makes sense because I've seen hazy days where the sky looks milky instead of blue.\n\nSo putting it all together: Sunlight hits Earth's atmosphere, interacts with small particles causing Rayleigh scattering, which scatters shorter wavelengths (like blue) more. This scattered blue light comes from all over the sky, making it look blue. When the sun is lower, like during sunrise or sunset, the longer path scatters out more of the blue, leaving behind reds and oranges.\n\nI think I have a grasp on it now. The color of the sky depends on how sunlight interacts with atmospheric particles, primarily through Rayleigh scattering affecting shorter wavelengths.\n\n/c/think\u003c/n\u003c\n\nThe sky appears blue due to a phenomenon called Rayleigh scattering. Here's a concise explanation:\n\n1. **Sunlight Composition**: Sunlight is composed of various colors, each corresponding to different wavelengths. Violet and blue have shorter wavelengths, while red has a longer wavelength.\n\n2. **Rayleigh Scattering**: As sunlight passes through Earth's atmosphere, it interacts with molecules (like nitrogen and oxygen) and small particles. Rayleigh scattering occurs when light interacts with particles smaller than its wavelength. This scattering is more effective for shorter wavelengths, such as blue and violet light.\n\n3. **Dominance of Blue Light**: Although violet light scatters more than blue, our eyes are more sensitive to blue light. Thus, the scattered blue light dominates the appearance of the sky, making it look blue.\n\n4. **Sunrise/Sunset Colors**: During sunrise or sunset, sunlight travels through a greater thickness of the atmosphere. Most of the blue and violet light is scattered out, allowing longer wavelengths like red and orange to dominate, resulting in vibrant sunsets.\n\n5. **Atmospheric Conditions**: Pollution or particles can cause Mie scattering, affecting all colors more equally, leading to hazy skies with different hues depending on particle size.\n\nIn summary, the sky's blue color arises from Rayleigh scattering of sunlight by atmospheric molecules, particularly affecting shorter wavelengths like blue light."}, "done_reason": "stop", "done": true, "total_duration": 2282

```

User-friendly AI  
Interface  
(Supports  
Ollama, OpenAI  
API, ...)



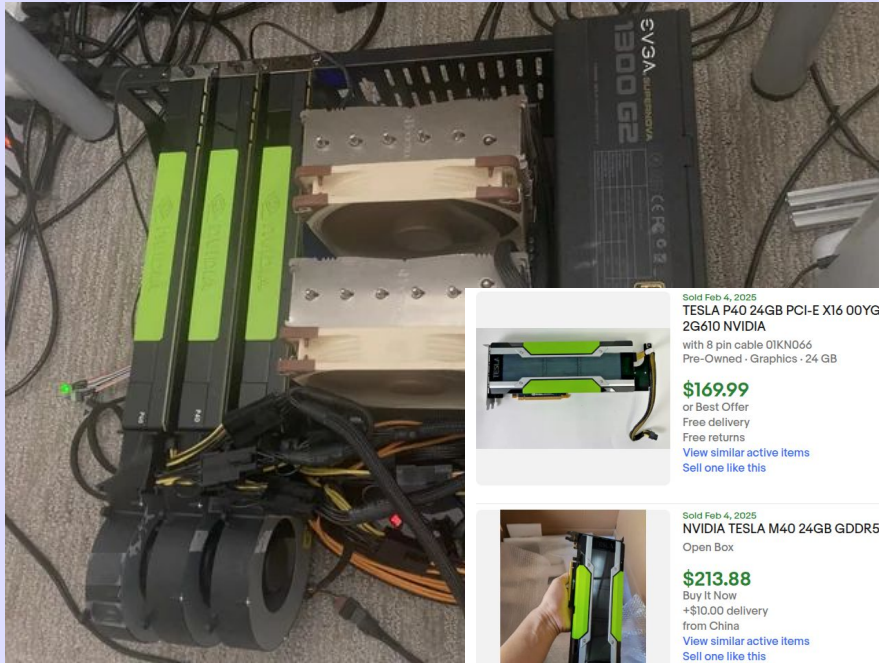


# Open-webui

The screenshot displays the Open-webui interface. On the left is a sidebar with a 'New Chat' button and a 'Workspace' section containing a search bar and a list of chats. The chat list includes titles like 'Egg Price Trends', 'Bird Airspeed Facts', 'Counting Fingers Exercise', 'Open Web Audio Settings', 'Is government hostile take over t', 'execute this MermaidJS diagram', 'Egg Prices Over Time Trend A', 'How to Water Plants', 'Today's Date', 'Global Temperature Rise', and 'Ohio Community Supported A'. Below these are 'Yesterday' entries: 'Eco-Friendly Straw Alternativ', 'Carrot Top Performances', 'Write a golang binary that is a cu', and 'Amazon's Pros and Cons of the k'. At the bottom of the sidebar is a user profile for 'admin.guy'. The main chat area shows a model selection dropdown with 'llama3.2-vision:11b' selected. Below the model name is a text input field with the placeholder 'How can I help you today?' and buttons for 'Web Search', 'Image', and 'Code Interpreter'. A 'Suggested' section follows with three prompts: 'Tell me a fun fact about the Roman Empire', 'Grammar check rewrite it for better readability', and 'Give me ideas for what to do with my kids' art'.

02

Architecture



Sold Feb 4, 2025  
**TESLA P40 24GB PCI-E X16 00YG990 GDDR5 GRAPHICS 2G610 NVIDIA**  
 with 8 pin cable 01KN066  
 Pre-Owned - Graphics - 24 GB

**\$169.99**  
 or Best Offer  
 Free delivery  
 Free returns  
[View similar active items](#)  
[Sell one like this](#)

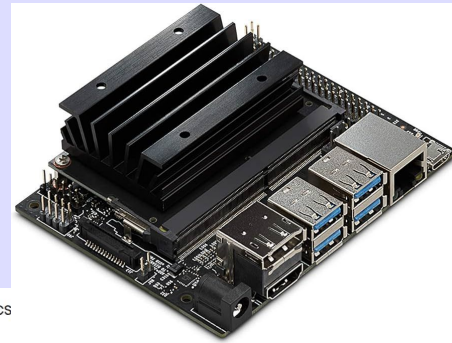
it-gear (6,536) 100%

---

Sold Feb 4, 2025  
**NVIDIA TESLA M40 24GB GDDR5 PCI-E 3.0X16 GPU CARD CUDA PG600**  
 Open Box

**\$213.88**  
 Buy It Now  
 +\$10.00 delivery  
 from China  
[View similar active items](#)  
[Sell one like this](#)

jjawen2018 (48,474) 991%



3:04 29%  
 Sign in & Lists  
 rtx 5090  
 Top Deals PC Parts Computers Apple Services  
 Shippable Items

**Search Results**  
 Choose a store to see products at your local Micro Center  
 21 Results for "rtx 5090"

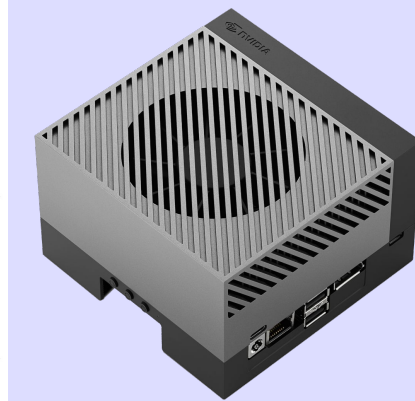
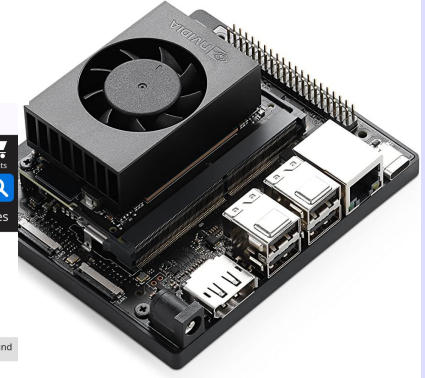
21 Product(s) found 1 Content Page(s) found

**FILTER** Sort by: Items per page: 24  
 Closest Match  
 See all 21 items at Micro Center

SKU: 800060  
 ASUS NVIDIA GeForce RTX 5090 ROG Astral Overclocked Triple Fan 32GB GDDR7 PCIe 5.0 Graphics Card  
 BUY IN STORE  
**\$3,079.99**  
 Select 2 to compare [QUICK VIEW](#) [ADD TO LIST](#)

SKU: 809244  
 Zotac NVIDIA GeForce RTX 5090 AMP Extreme Infinity Overclocked Triple Fan 32GB GDDR7 PCIe 5.0 Graphics Card  
 BUY IN STORE  
**\$2,499.99**  
 Select 2 to compare [QUICK VIEW](#) [ADD TO LIST](#)

SKU: 805564



# Alternatives Considered

- Used commercial GPU's like Nvidia P40
- Newer Nvidia Jetson hardware
- Old Desktop Gaming GPU



## Mini ITX

Smallish form factor  
Consumer grade components

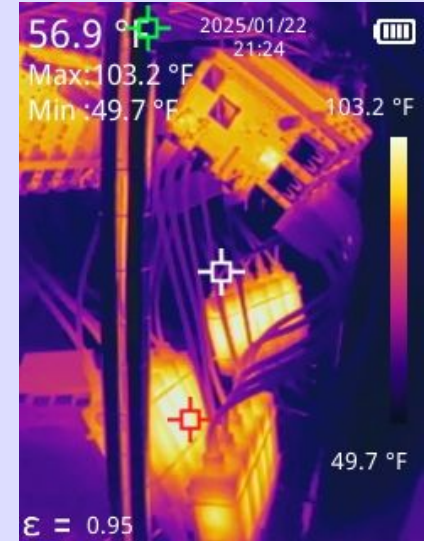


RTX3090

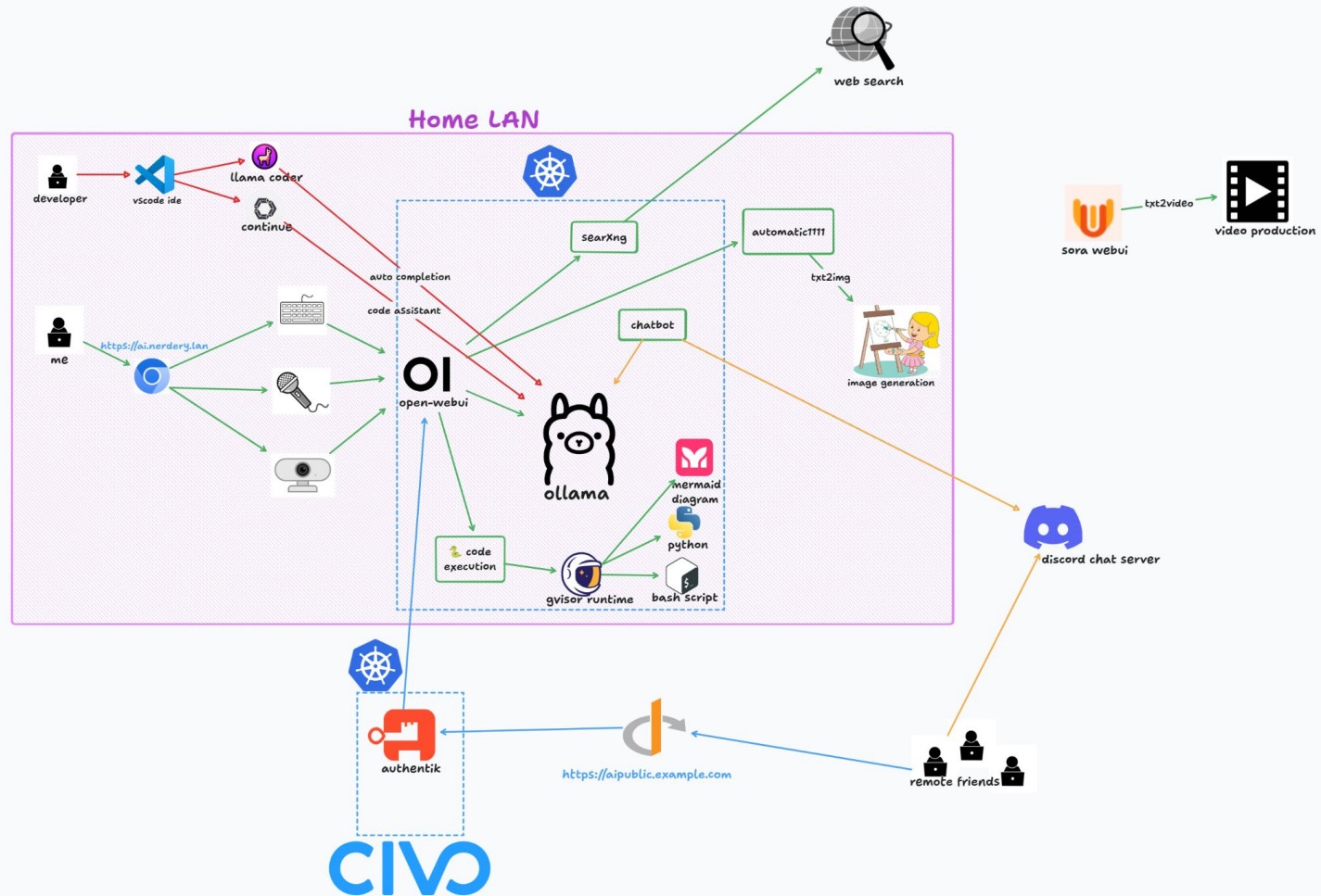
Mostly a GPU with a computer attached



13



# Architecture





Prometheus



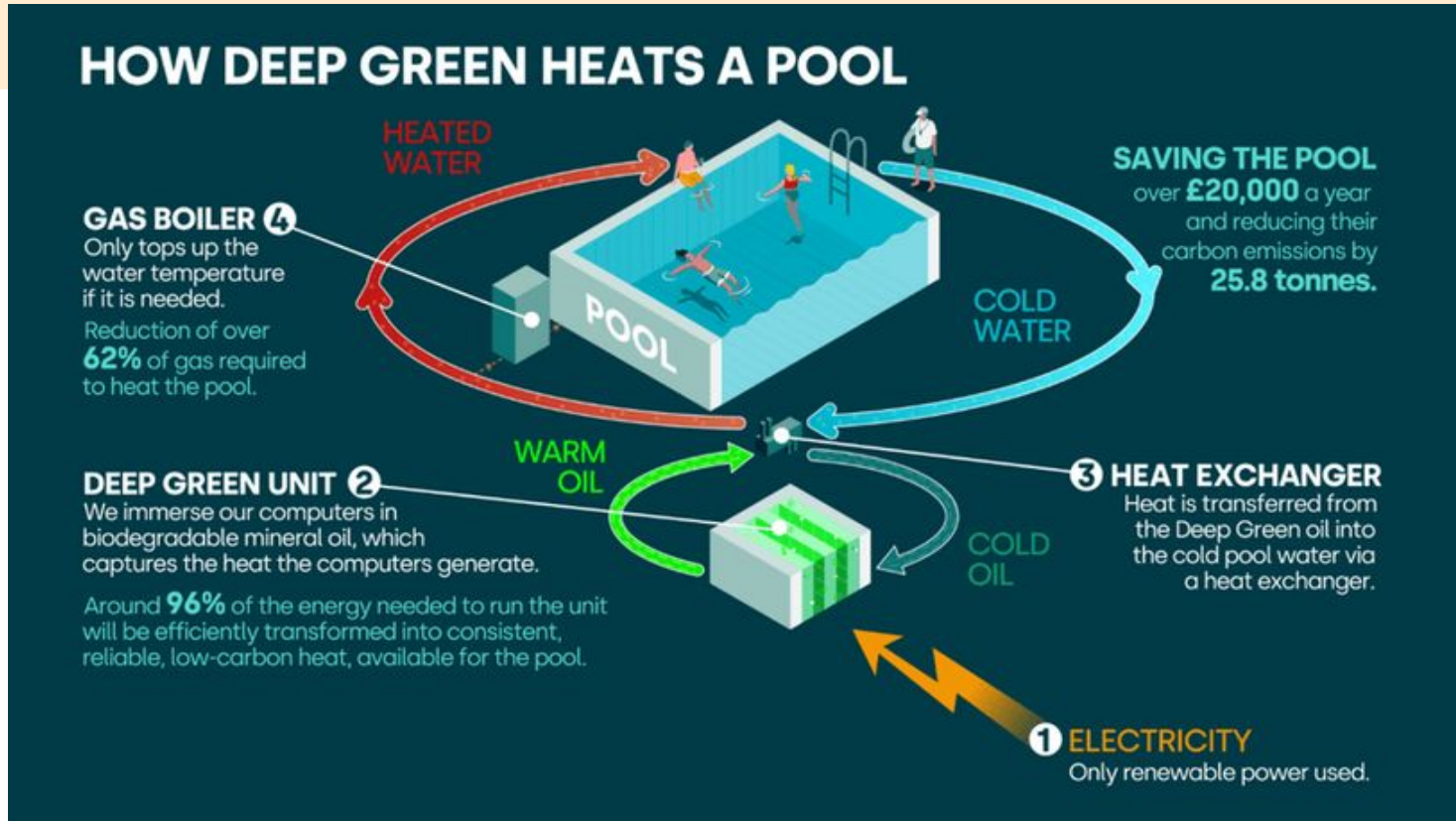
Grafana



## Usage Metrics

- Generally 30-40w idle but 400w at peak drawn for RTX3090 24GB VRAM.
- Prometheus & Grafana scraping metrics and viewing dashboards
- Good Space Heater when not idling

# Civo - Deep Green GPU's





The screenshot displays the Civo website's navigation bar with links for Products, CivoAI, Pricing, Resources, and Navigate. The main content area is divided into two columns. The left column features a vertical list of services: High performance virtual machines, NVIDIA GPUs (available via Compute & Kubernetes), Machine Learning (powered by KubeFlow), Object Store (S3 compatible), Volumes (persistent for stateful workloads), Load Balancers (for availability and performance), and Databases (managed and scalable). The right column contains two pricing tables. The top table lists general instance sizes (Small, Medium, Large, Extra Large) with their respective RAM, vCPU, storage, and on-demand/commitment rates. The bottom table is titled 'NVIDIA H200 SXM GPU pricing' and shows pricing for Small and Extra Large GPU instances. A dropdown menu indicates a commitment of 36 months.

Size	RAM	vCPU	Storage	Data Transfer	On-demand	Commitment
<b>Small</b> 1 x NVIDIA H100 - 80GB	32 GB	16 Cores	400GB NVMe	FREE	\$2.49 per hour	\$1.99 per hour
<b>Medium</b> 2 x NVIDIA H100 - 80GB	384 GB	32 Cores	800GB NVMe	FREE	\$4.98 per hour	\$3.98 per hour
<b>Large</b> 4 x NVIDIA H100 - 80GB	768 GB	64 Cores	1600GB NVMe	FREE	\$9.96 per hour	\$7.96 per hour
<b>Extra Large</b> 8 x NVIDIA H100 - 80GB	1536 GB	128 Cores	3200GB NVMe	FREE	\$19.92 per hour	\$15.92 per hour

Size	RAM	vCPU	Storage	Data Transfer	On-demand	Commitment
<b>Small</b> 1 x NVIDIA H200 - 141GB	192 GB	16 Cores	400GB NVMe	FREE	\$3.49 per hour	\$2.99 per hour
<b>Extra Large</b> 8 x NVIDIA H200 - 141GB	1536 GB	128 Cores	3200GB NVMe	FREE	\$27.92 per hour	\$23.92 per hour

# Cloud

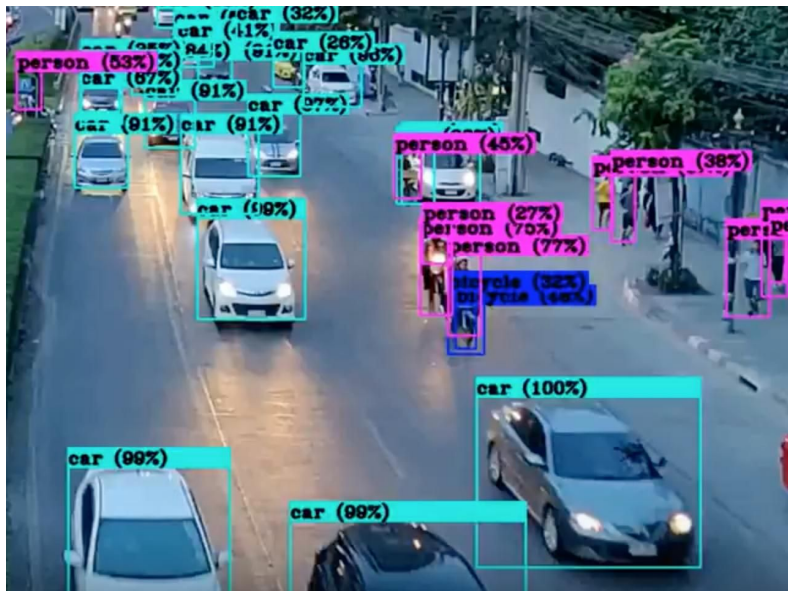
<https://www.civo.com/cloud-gpu>, pricing starting at \$0.79/hour

Capex vs Opex, most companies budget for 3-5 year refresh, while gpu's generally have a 2 year life to them. Opex spend of cloud on-demand can help get away from this quicker refresh cycle. Civo for example has H100 & now even H200's available, the latest and greatest GPU's available in any cloud providers

03

Features

# Vision



llama3.2-vision:11b v +




How many fingers am I holding up?

- llama3.2-vision:11b Today at 4:01 PM  
Three.

# Call, mic, camera, keyboard

ai.nerdery.lan/c/b1c780c5-aeda-4c34-9e85-236f6af55d7a

llama3.2-vision:11b



Describe what you see in the room.

**llama3.2-vision:11b**


The image shows a man posing for an online stream or recording session, seated at his computer station and gesturing with two fingers held up as if making a "V" sign.

A large white microphone stand is visible on the left side of the photo, capturing the man's voice and comments to be broadcasted. The desk in front of him holds the necessary equipment for this type of recording or broadcasting, including a computer monitor, keyboard, and mouse.

The background features a well-organized room with shelves holding additional monitors, speakers, and other audio equipment, suggesting that the individual is either a professional streamer or podcaster or an enthusiastic hobbyist.

Send a Message

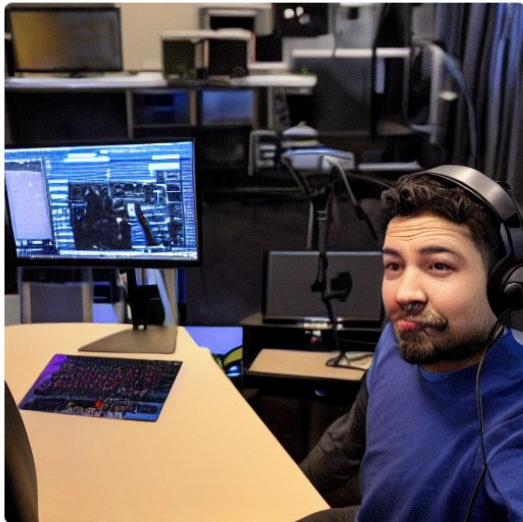
Web Search Image Code Interpreter



Tap to interrupt

# txt2img

OI llama3.2-vision:11b Today at 9:45 PM



The image shows a man posing for an online stream or recording session, seated at his computer station and gesturing with two fingers held up as if making a "V" sign.

A large white microphone stand is visible on the left side of the photo, capturing the man's voice and comments to be broadcasted. The desk in front of him holds the necessary equipment for this type of recording or broadcasting, including a computer monitor, keyboard, and mouse.

The background features a well-organized room with shelves holding additional monitors, speakers, and other audio equipment, suggesting that the individual is either a professional streamer or podcaster or an enthusiastic hobbyist.



Generate Image

Send a Message

+  Web Search  Image  Code Interpreter



# automatic1111

Stable Diffusion checkpoint  
v1-5-pruned-emaonly.safetensors [6ce0161689]

txt2img img2img Extras PNG Info Checkpoint Merger Train Settings Extensions

Bob ross style mountain painting with a log cabin next to santa's sled. 15/75

Negative prompt 0/75  
(Press Ctrl+Enter to generate, Alt+Enter to skip, Esc to interrupt)

Generate

Generation Textual Inversion Hypernetworks Checkpoints Lora

Sampling method DPM++ 2M Schedule type Automatic Sampling steps 20

Hires. fix  Refiner


Width 512 Batch count 1

Height 512 Batch size 1

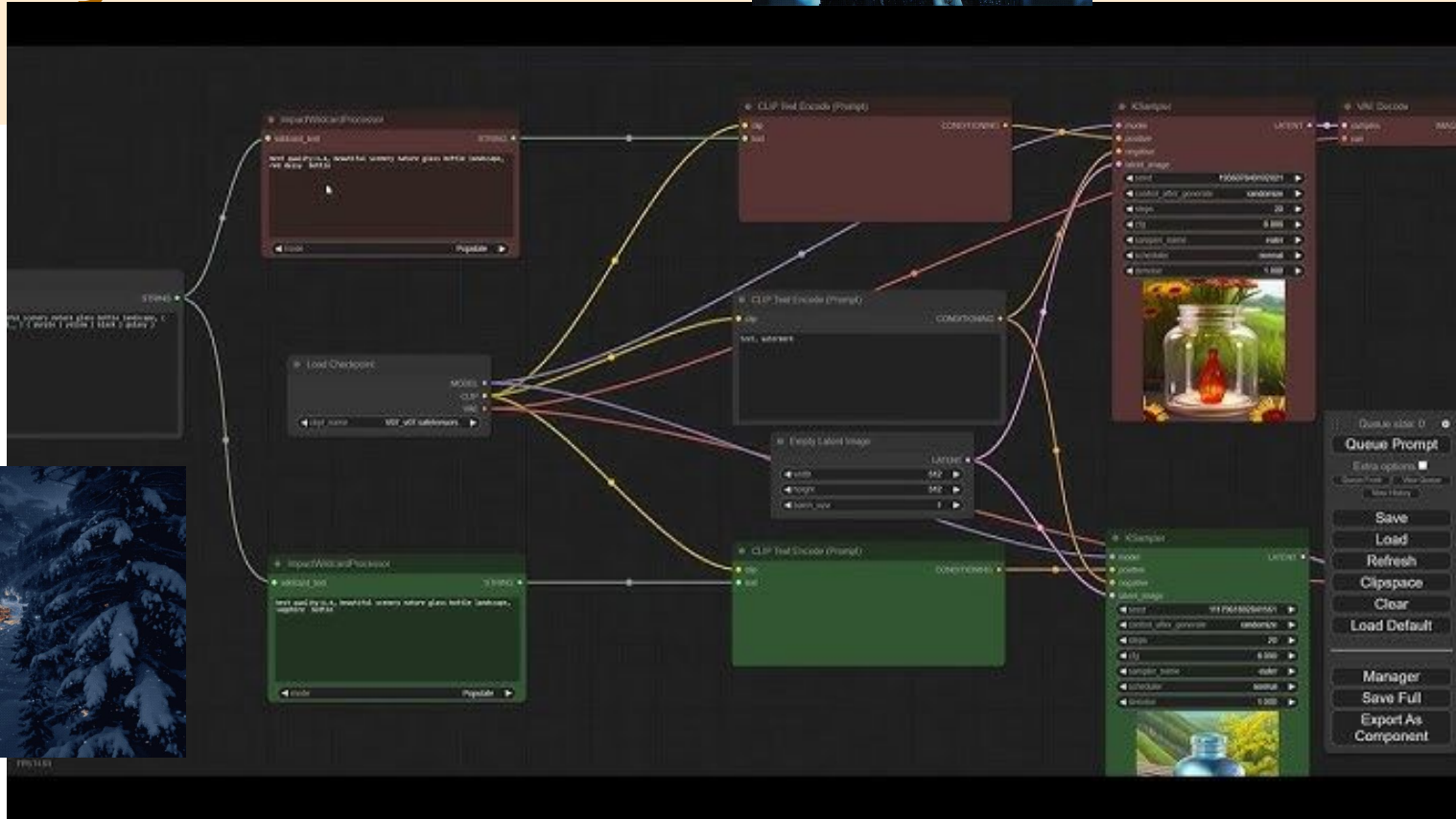
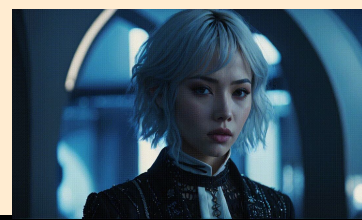
CFG Scale 7

Seed -1  Extra

Script None



# Comfy ui







# discord chatbot

aiquery Enter your ai query



/aiquery aiquery give me a bulleted list of interesting facts about the dayton linux user group under 2000 characters.

Zarrsh used [aiquery](#)



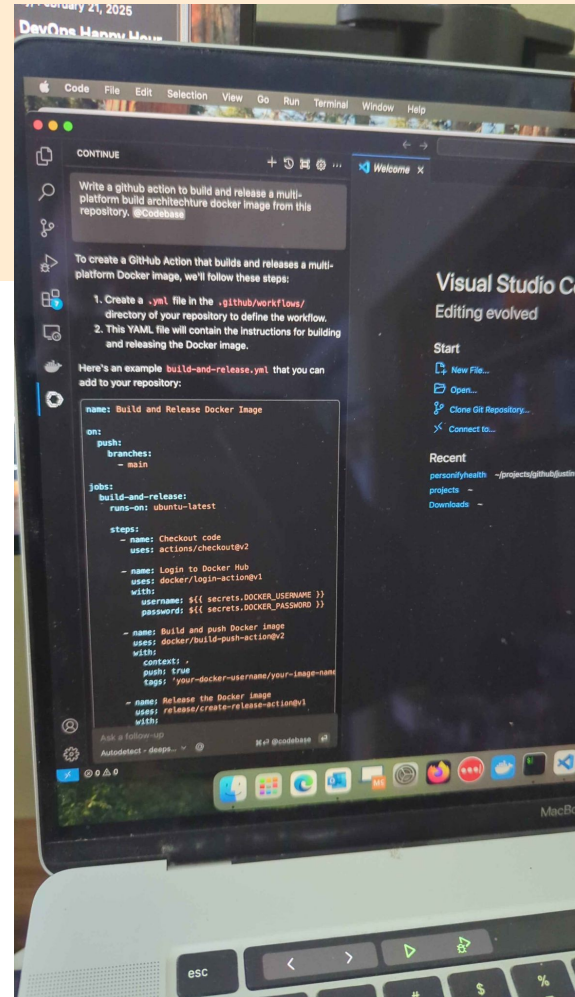
nLabsBot **APP** Today at 9:34 PM

Here are some interesting facts about the Dayton Linux User Group (DLUG):

- **Founded in 1997:** DLUG is one of the oldest Linux user groups in the world.
- **Community-driven:** The group is run by volunteers and has no membership fees or dues.
- **Meetups and events:** DLUG hosts monthly meetups, workshops, and conferences featuring expert speakers on various Linux-related topics.
- **Hands-on learning:** Members can participate in hands-on training sessions, install-fests, and other interactive activities to learn about Linux.
- **Diverse user base:** DLUG members come from a variety of backgrounds, including professionals, students, hobbyists, and even industry representatives.
- **Collaborations:** The group has collaborated with local universities, schools, and businesses to promote Linux adoption and education.
- **Meetups held online too:** Due to the pandemic, DLUG started hosting virtual meetups, making it easier for remote members to participate.

# coding assistant

- Continue.dev, prompting like github co-pilot (larger model) <https://www.continue.dev/>
- Codellama, autocomplete suggestions in-line (smaller model) <https://github.com/xNul/code-llama-for-vscode>



04

Demo

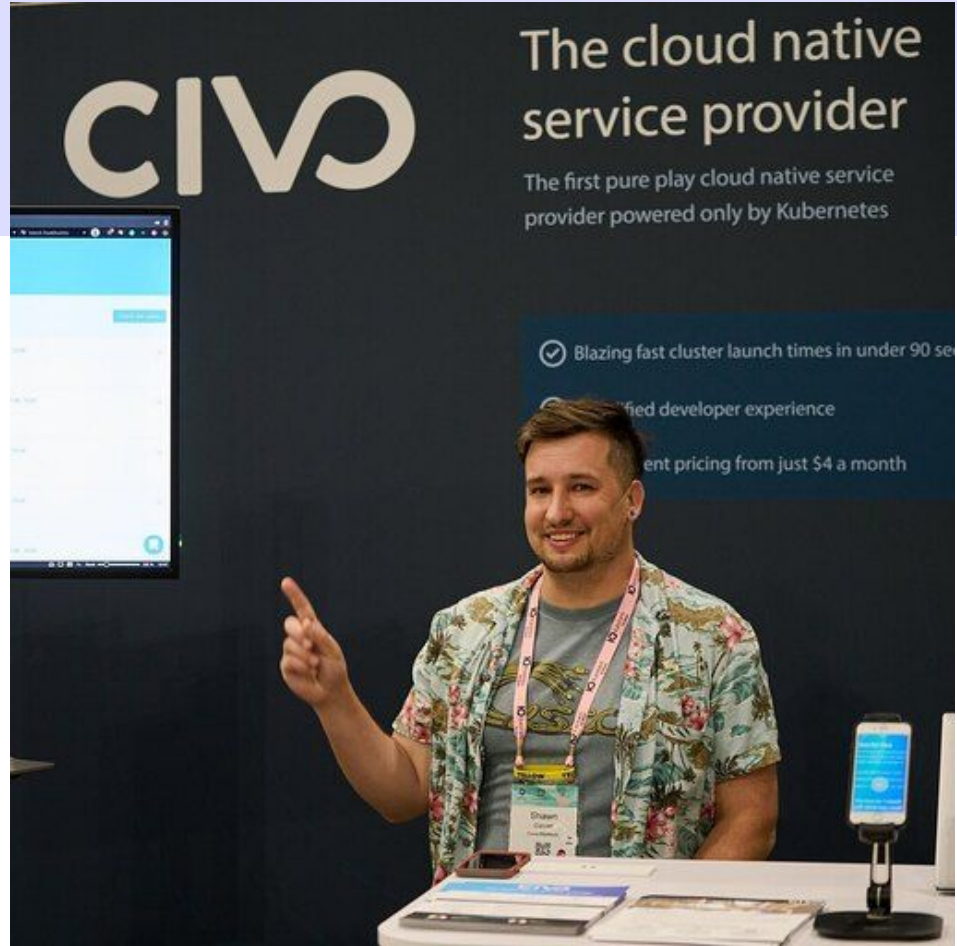
05

Questions

# References & good links

- [Matt Williams - YouTube](#) (founding maintainer of ollama and cool guy)
- [Ollama](#)
- [Open WebUI](#)
- [Nvidia P40 Datasheet](#)
- [Quantisation, using less RAM](#)
- [How to deploy DeepSeek-R1 on Civo GPUs](#)
- [DeepSeek vs Llama vs GPT-4 | Open-Source AI Models Compared - Civo.com](#)
- [Deepseek-r1, an opensource model built at a fraction of the cost](#)
- [ComfyUI Examples](#)
- [Open-Webui vs ComfyUI](#)
- [Huggingface ai community, models available](#)
- [Vllm, so you outgrew ollama and need something bigger](#)
- [SearXNG, internet metasearch engine which aggregates results from various sources](#)
- [Whisper, robust speech recognition via large-scale weak supervision](#)
- [GPU-Powered GenAI in one Click, Civo](#)
- [Civo Slack Community](#)

# Questions?



THANK YOU

Shawn Garrett