➢ Title

I built an AI supercomputer with 5 Mac Studios

➢ Source

https://www.youtube.com/watch?v=Ju0ndy2kwlw&t=142s

➢ Introduction

Network Chuck first discusses how he connected 5 Mac Studios together to give himself a Private AI supercomputer at home. He then shares his motivation to be able to run large AI models locally rather than using a cloud platform, such as Google Colab, or Open AI APIs.

➢ -Why we use Private AI

✓ **Privacy**:

Your data never leaves your machine, so there are no more cloud services to trust.

✓ **Control**:

You control how, when, and where the model runs.

✓ **No restrictions**:

No limits on API calls, no limits on data, and no waiting time.

✓ **Faster access:**

There are have options available and if set up correctly, local only models respond quickly and stay offline.

➢ Tools used in video

- ✓ 5 Mac Studios (with Apple M-series chips, each powerful individually).
- ✓ LAN (w/ Ethernet cables & an Ethernet switch) for connecting each Mac Studio machines.
- ✓ Ray & PyTorch - tools to run machine learning jobs across other devices.
- ✓ PrivateGPT / Ollama / LM Studio - examples of running LLMs (like Llama2) on local hardware.
- ✓ Terminal/CLI - to command and script the entire AI system.

➢ Step by step process explained in the video

- ✓ **Set up the Hardware**
  - Unboxed and set up 5 Mac Studios.
  - Connected them via Ethernet to a switch (network hub).

- ✓ **Installed Useful Tools**
  - Installed Homebrew (package manager) and set up PyTorch.
  - Installed Ray (distributed computing tools) to connect all the devices together.
  - Set up Ollama so I would be able to run LLMs like LLaMA, Mistral, or others locally.

- ✓ **Link the Macs**
  - Used Ray to make all the Mac Studios work together (computing cluster).
  - One Mac was the "head node", and the others are "workers".

- ✓ **Run a Model Locally**
  - Produced a language model using Ollama.
  - Tested how fast and efficiently the cluster could process AI task.

- ✓ **Compare with Cloud**
  - Demonstrated that this was faster and more private than using OpenAI or any other external tools.

➢ Benefits of this setup

- ✓ **Privacy**
  - All data is offline and secure.

- ✓ **Speed**
  - Distributed power from 5 machines = faster AI processing

- ✓ **Cost-saving**
  - One-time hardware cost; no more cloud bills every month

- ✓ **Learning**
  - Helps with understanding system design, networking, and AI tools.

- ✓ **Customizable**
  - You can modify models, enhance security, or scale.

➢ What I learned

- ✓ How to configure and connect multiple systems as an AI cluster.
- ✓ Why Ray is powerful for managing distributed computing.
- ✓ That Macs with M-series chips are capable of substantial AI if used together.
- ✓ That products like Ollama and LM Studio help you to run LLMs without being connected to the internet.
- ✓ Importance of privacy, control, and speed with AI.

➢ Screenshot's:

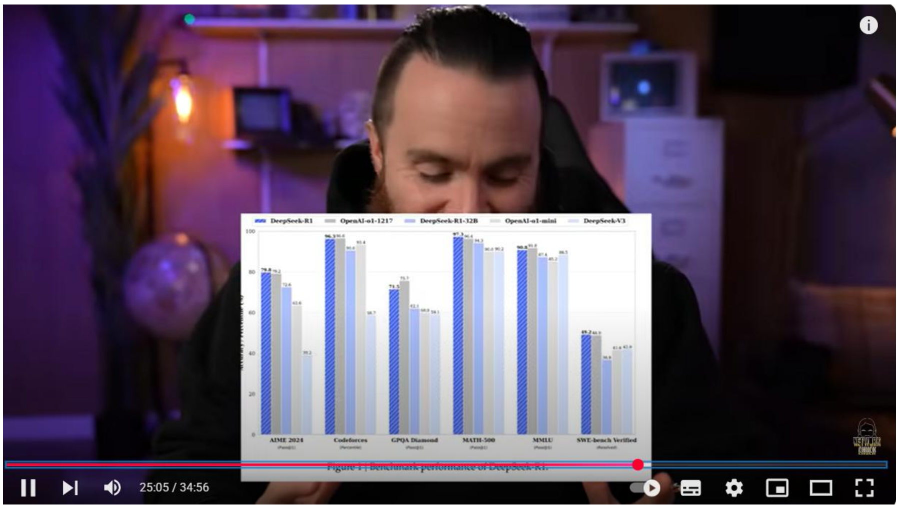
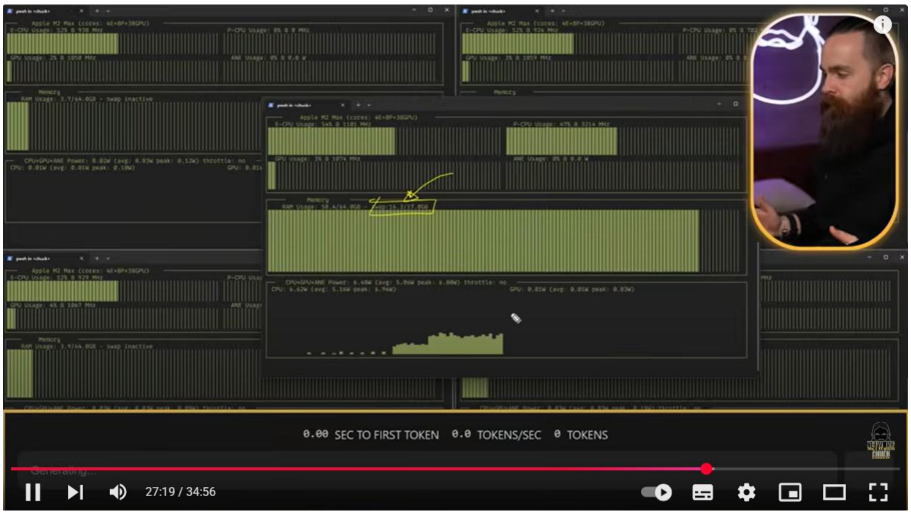


Figure 1 | Benchmark performance of DeepSeek-R1.



```
Apple M2 Max (cores: 4E+8P+38GPU)
E-CPU Usage: 54% @ 1101 MHz                P-CPU Usage: 47% @ 3214 MHz
GPU Usage: 3% @ 1074 MHz                   ANE Usage: 0% @ 0.0 W
Memory
RAM Usage: 58.4/64.0GB - swap: 16.3/17.0GB
CPU+GPU+ANE Power: 6.48W (avg: 5.06W peak: 6.80W) throttle: no
CPU: 6.62W (avg: 5.14W peak: 6.96W)        GPU: 0.01W (avg: 0.01W peak: 0.03W)
```

```
0.00 SEC TO FIRST TOKEN   0.0 TOKENS/SEC   0 TOKENS
Generating...
```