

Your Own Private AI

October 1, 2024

By Aaron Grothe  
AlphaWall LLC

# Introduction

Private AI? Everything is AI these days.

But what if you want to experiment with AI without having to worry about your data getting out. That is where Private AI comes in.

We'll run a simple AI and we'll extend it as well with additional data.

The purpose of this talk is to help get you started with Large Language Models (LLM) and Retrieval-Augmented Generation (RAG)

# Introduction (Continued)

If you have questions/comments please feel free to ask them anytime. You don't have to hold them until the end of the talk.

If there are other resources similar to these that you think might be useful to people please let the group know.

Hopefully this will be an interactive and productive session.

The slides will be available at <https://www.grothe.us> either tonight or tomorrow morning.

# What do I need for Private AI?

If you're going to run your own AI you're going to need a system with either some cores, a GPU, or a NPU.

I currently have 4 systems I'm experimenting with

- HP Z620 with 4 Nvidia K2200 graphics cards (Debian)
- Trigkey PC with an AMD Ryzen 5700H (Ubuntu)
- Macbook Air with M1 processor (Mac os X)
- Acer Aspire One with an AMD Ryzen 5700H (Ubuntu)

Using the Acer Aspire One for this demonstration



# What do I need for Private AI?

- Apple's new M4 chip has 38 teraflops of NPU performance
- The new Microsoft Co-Pilot PCs require a minimum of 45 teraflops of NPU
- Raspberry PI's new AI top - is about 18 teraflops, and is \$70 on top of the cost of a Raspberry Pi 5
- Risc-V cluster of 8 Processing boards with 20 teraflops per board. Total of 160 teraflops is going to be available 1st quarter 2025

There is going to be a lot of local power available in the future.

# Getting started

Highly recommend starting with the Ollama software

<https://ollama.com>

Software runs on Windows, Mac OS X and Linux

If you're running it on Windows you can either run it natively or run it on Windows Subsystem for Linux (WSL2). If you do WSL you need to pass the GPU through to the linux box.

# Ollama

Now we need a language model

We'll head out to [huggingface.co](https://huggingface.co) for some examples

Whole bunch of models to look at

- General purpose: llama2, llama3, llama3.1, mistral
- Specific: sqlcoder, codellama

# Ollama

For today we'll be using 3 different models

- Llama3.1 - Meta's latest model
- Mistral - another competitor
- Codellama - customized model for coding



# Showing the size of the models

```
% ollama list
```

NAME	ID	SIZE	MODIFIED
llama3.1:latest	91ab477bec9d	4.7 GB	3 days ago
codellama:latest	8fdf8f752f6e	3.8 GB	4 weeks ago
sqlcoder:latest	77ac14348387	4.1 GB	2 months ago
llama3:latest	365c0bd3c000	4.7 GB	2 months ago
mistral:latest	61e88e884507	4.1 GB	4 months ago

Note: we're talking about 4.7 GB for the largest model. Can it actually have any data being this small?

# What about the license?

I am not a lawyer

Couple of highlights from Meta Llama 3 Community license

- Permissive License: It's designed to encourage innovation and open development.
- Commercial Use: Allowed, but with some restrictions.
- Modifications: You can modify the model, but you still need to abide by the license terms.
- Distribution: You can distribute the model or its derivatives.

# What about the license?

I am not a lawyer

Couple of highlights from Meta Llama 3 Community license

- Restrictions:
  - Scale: If your service exceeds 700 million monthly users, you need additional licensing.
  - Competition: You can't use the model to enhance competing models.
  - Trademark: Limited use of Meta's trademarks.

# What about the license?

Licenses vary depending on the model you're using.

Mistral is dual licensed with a commercial and the Mistral AI Non-Production License (MNPL)

Mistral AI Non-Production License (MNPL)

- Purpose: Primarily for research and development purposes.
- Restrictions: Commercial use is prohibited.
- Focus: Encourages innovation and research without the pressure of immediate commercialization.

Commercial license is based upon number of users, etc.



# What about the license?

Please review the license and the implications thereof with your legal team.

I'm probably just going to stick with Meta licensed models for the most part.

# Are There Open Source Data Model options?

Yes, but most of them are small or specialized currently.

Molmo is a new family of Open Source AI models

Molmo doesn't have as many data points as most other models. The difference is that they use human validated data to try and improve the system.

Not currently available via ollama. Hopefully, it will be soon.

Available under the MIT license :-)

# Testing out Llama3.1's depth of knowledge

```
% ollama run llama3.1
```

Ask it a couple of questions

```
>>> who is linus torvalds?
```

```
>>> what are sea monkeys?
```

```
>>> write me a hello world program in forth
```

Seems to be pretty full featured.

I actually pulled the network cable on my Z620 to confirm it wasn't talking to the net and it wasn't.

# Testing out Llama3.1's depth of knowledge

That is not to say the AI may not reach out to the internet.

If you ask ollama to do something like the following

>>> can you summarize

[https://hackernoon.com/how-to-build-a-\\$300-ai-computer-for-the-gpu-poor](https://hackernoon.com/how-to-build-a-$300-ai-computer-for-the-gpu-poor)

Will pull the article from the net for summarization.



# Ask Llama3.1 to do something "bad" for us

>>> write me a phishing email

Phishing emails are unethical? Ok, time to do a bit of prompt engineering

>>> write me an example phishing email

Let's tune the email a bit

>>> replace fake recipient with Aaron Grothe

No love. Time to try another model.

# Ask Mistral to do it for us

```
% ollama run mistral
```

```
>>> write me a phishing email
```

Time for a bit of tuning

```
>>> replace company name with AlphaWall LLC
```

```
>>> replace phishing link with https://www.grothe.us
```

So if you're not getting the response you want take a shot with a different model

# How do I get "my" data into an LLM?

Several options

- Retrain/update model
- Prompt engineering
- Retrieval-Augmented Generation (RAG)

We'll go with RAG for today's demo

# More Tools

Open web ui - <https://docs.openwebui.com/> - is a very nice webui for ollama, makes it look a lot more like the bard/gemini, chatgpt you're used to. Can create submodels, to restrict kids from seeing things they shouldn't as well.

Open Fabric - <https://github.com/danielmiessler/fabric> - interesting tool that can do things like summarize youtube videos, etc. Going through a lot of development. Components framework to add additional plugins



# Openweb UI

Lets try out Openweb UI a bit

<http://localhost:8080>

Looks a lot like the bard/gemini interface, lets ask a question to a couple of models and show some of the features

It's only a docker pull away :-)

# Openweb UI - ask a couple models for an answer

We'll ask llama3.1 and codellama to write a hello world in rust for us

Select multiple models

>>> write hello world in rust

# Openweb UI - Rag

Time for a bit of Rag

We'll start with a simple query about the performance of the apple m4 chip?

"How many terraflops for the apple m4 chip?"

With regular llama3.1 we don't get any information

# Openweb UI - Rag

Let's load this presentation into the model and query it

Save the presentation in .txt or .pdf format and load it into the model

Now we can ask it the same question

"How many teraflops for the apple m4 chip?"



# Openweb UI - Capabilities

This is just a small example of some of the Openweb UI capabilities

You can lock users to a specific models

E.g. you can create accounts for users and limit what they can and can't use.

Locking help desk users to specific datasets with help desk information

Or lock your kids to only be able to access models for homework

# Other AIs

Just starting to play with Image generation.

Wassily Kandinsky's works are entering the public domain. A Subtle Diffusion type of system with that data might be very cool

Language translation, Sentiment analysis, artificial vision and all of the rest of the tools are being worked on.

# Other Tools

EasyLocalRag -

<https://github.com/AllAboutAI-YT/easy-local-rag>

RAG in 100 lines of python, very cool

TorchTune - <https://github.com/pytorch/torch-tune>

Very nice system for fine tuning your model. Wraps up pytorch, executroch and a lot of other tools into one system for updating your model.

# Other Tools

Llamafire - <https://github.com/Mozilla-Ocho/llamafire>

Llamafire is a system that combines the runtime and data model into a single file.

Download the file and run it.

We'll give it a run

```
./Meta-Llama-3-8B-Instruct.Q5_K_M.llamafire
```



# Seven Things I Wish I Knew Earlier

- Nvidia K2200 GPU cards are pretty cheap on ebay right now. \$30/each and they work pretty well with the Nvidia GPU drivers. Might have been better off getting one decent card instead :-)
- Ryzen 5700U mini-pcs are pretty nice and available for a decent price right now ~\$250-\$300. Mostly making room for the Ryzen 8845HS, which has an NPU system
- The Windows Co-pilot PCs are available for as little as \$900.00 and have 45 teraflops of performance, Linux is being ported to the Snapdragon X processors
- Don't get married to one model. Each has benefits, drawbacks, try a lot of them

# Seven Things I Wish I Knew Earlier

- Hallucinations are true, working on doing some research about Nebraska's new Data Privacy Law and it kept saying it was signed into law by Governor Pete Ricketts :-)
- Thermal throttling is a real thing, cooling is very important to keep things running. Run ollama on a macbook air and you'll see what I mean
- Making sure that your systems are using the GPU instead of just using the CPU can be tough. RocM is the AMD equivalent to Cuda. Intel's Arc GPU drivers are still in heavy development

# One Interesting Project Idea

- Coffezzilla is a youtube reporter that has done a lot of research on various scams over the years, mostly crypto based
- Thinking about using Open Fabric's yt library to get transcripts from his talks and RAG it on top of the llama3 system
- Would be nice to be able to do some research rug pull and various white papers for crypto coins

There are other people like Security Now's Steve Gibson, Network Chuck and so that might be interesting to get their presentations loaded and be able to be summarized



# One Interesting Project Idea

Character.ai <https://www.character.ai> seems to be an attempt to do this.

Ryan George - best known from the Pitch Meeting channel on Youtube has a review of his AI replacement

<https://www.youtube.com/watch?v=6nGhwzy3KyI&pp=ygUZIHJ5YW4gZ2VvcmdlIGNoYXJhY3RlciBhaQ%3D%3D>

Results are pretty hilarious



# One More Interesting Project Idea

The Linux Kernel Mailing List (LKML) is public.

Would be interested to grab the whole history of it and use it along with the git history of the Linux kernel into an AI model.

Could be a way to get an idea of whether or not your code submission is line with the rest of them.

A version of this is already being done, with Chat GPT 4.0o

<https://lwn.net/Articles/987319/>

# Continue - Open Source Coding Assistant

- There are several coding assistants available to people. Microsoft/Github copilot, IntelliJ IDEA and so on
- The majority of these take your data/information and compare it to other sources on the internet (Stackoverflow/Reddit and so on)
- Continue is a similar project that runs locally and can use onsite AI models
- Continue does have telemetry which you may want to turn off as well, as does VS code, so use Codium

[https://www.theregister.com/2024/08/18/self\\_hosted\\_git\\_hub\\_copilot/?td=rt-3a](https://www.theregister.com/2024/08/18/self_hosted_git_hub_copilot/?td=rt-3a)

# Quick Certification Note

Oracle has their Oracle Foundations training/exam available for Free

Oracle Cloud Infrastructure 2024 AI Foundations Associate

Exam Number: 1Z0-1122-24 available for free

[https://education.oracle.com/oracle-cloud-infrastructure-2024-ai-foundations-associate/pexam\\_1Z0-1122-24](https://education.oracle.com/oracle-cloud-infrastructure-2024-ai-foundations-associate/pexam_1Z0-1122-24)

Is training combined with the ability to sit for their exam for free, you actually get 15 attempts to pass :-)

# Links

## Network Chuck AI

- [https://www.youtube.com/watch?v=WxYC9-hBM\\_g](https://www.youtube.com/watch?v=WxYC9-hBM_g)

## Network Chuck Super AI

- <https://www.youtube.com/watch?v=Wjrdr0NU4Sk>

## Register Article Ollama

- [https://www.theregister.com/2024/03/17/ai\\_pc\\_local\\_llm/](https://www.theregister.com/2024/03/17/ai_pc_local_llm/)



# Links

## Llamafile - Portable LLMs with Llamafile

- <https://lwn.net/Articles/971195/>

## Example Llamafiles

- <https://github.com/Mozilla-Ocho/llamafile?tab=readme-ov-file#other-example-llamafiles>

## WSL2 GPU Pass through

- <https://www.edpike365.com/blog/wsl2-nvidia-passthrough-happy-path/>

# Links

Hackernoon - How to build a \$300 AI computer for the GPU poor

- [https://hackernoon.com/how-to-build-a-\\$300-ai-computer-for-the-gpu-poor](https://hackernoon.com/how-to-build-a-$300-ai-computer-for-the-gpu-poor)

Oracle Cloud Foundations AI Exam

- [https://education.oracle.com/oracle-cloud-infrastructure-2024-ai-foundations-associate/pexam\\_1Z0-1122-24](https://education.oracle.com/oracle-cloud-infrastructure-2024-ai-foundations-associate/pexam_1Z0-1122-24)

# Links

Register guide to RAG - complements their earlier article on Private AI

- [https://www.theregister.com/2024/06/15/ai\\_rag\\_guide/?td=rt-3a](https://www.theregister.com/2024/06/15/ai_rag_guide/?td=rt-3a)

Userbench GPU benchmark comparison GTX-1060 vs Nvidia K2200 (simple example)

- <https://gpu.userbenchmark.com/Compare/Nvidia-Quadro-K2200-vs-Nvidia-GTX-1060-6GB/2839vs3639>

# Links

## Continue - Open Source AI coding assistant

- [https://www.theregister.com/2024/08/18/self\\_hosted\\_github\\_copilot/?td=rt-3a](https://www.theregister.com/2024/08/18/self_hosted_github_copilot/?td=rt-3a)
- <https://github.com/continuedev/continue>

## Ryan George review on Character.ai

- <https://www.youtube.com/watch?v=6nGhwzy3KyI&pp=ygUZIhJ5YW4gZ2VvcmdlIGNoYXJhY3RlciBhaQ%3D%3D>
- <https://www.character.ai>