Your Own Private AI

June 19, 2024

By Aaron Grothe
NEbraskaCERT

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.

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 3 systems I'm experimenting with

- HP Z620 with 4 Nvidia K2200 graphics cards
- Trigkey PC with an AMD Ryzen 5700H in it
- Macbook Air with M1 processor

Using the Macbook Air for this demonstration

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 for some examples

Whole bunch of models to look at

- General purpose: llama2, llama3, mistral
- Specific: sqlcoder

Ollama

For today we'll be using 3 different models

- Llama2 - Facebook's model used by a lot of people
- llama 3 - Latest model from Facebook
- Mistral - another interesting model

Showing the size of the models

% ollama list

NAME	ID	SIZE	MODIFIED
llama2:latest	78e26419b446	3.8 GB	2 months ago
llama3:latest	365c0bd3c000	4.7 GB	3 days ago
mistral:latest	61e88e884507	4.1 GB	2 months ago
sqlcoder:latest	77ac14348387	4.1 GB	7 minutes ago

Note: we're talking about 4.7 GB for the largest model

Can it actually have any data being this small

Testing out Llama2's depth of knowledge

% ollama run llama2

Ask it a couple of questions

>>> who was david bowie?
>>> 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.

Ask Llama2 to do something "bad" for me

>>> 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 Name 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 valued customer with aaron grothe
>>> replace phishing link with https://www.nebraskacert.org

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

Back to llama2

One of the limitations of every model it is only current to the time it was trained, and the data it was trained on.

E.g. llama2 was trained up until January 2023 - July 2023, so it doesn't have any more recent data

% ollama run llama2

>>> what is the passing score for the amateur radio exams?

Try it with llama3

% ollama run llama3

>>> what is the passing score for the amateur radio exams?

Right answer, which is a good thing

So llama2 is wrong how can we fix this?

Several options

- Get a newer model
- Retrain/update model
- Retrieval-Augmented Generation (RAG)

We'll go with RAG for today's demo

Which RAG to Pick?

There are a LOT of RAGs to pick.

For our demo we'll do easy-local rag

Need a RAG - we'll use easy-local-rag -
https://github.com/AllAboutAI-YT/easy-local-rag

Need a corrected data set
https://www.hollandarc.org/?page_id=3570
https://www.hollandarc.org/wp-content/uploads/2020/07/Extra2024AD7FO.pdf

Get the RAG up and running - no data

- git clone https://github.com/AllAboutAI-YT/easy-local-rag.git
- cd easy-local-rag
- pip install -r requirements.txt
- Install ollama, llama2 and mxbai-embed-large if needed
- Modify config file to use llama2 instead of llama3
- run localrag.py (with query re-write)

Get the RAG up and running - raw data

- run upload.py and upload the full version of the file
- This will append the data to the vault.txt
- run localrag.py
- "What is the passing score for the amateur radio exams?"

Correct answer, but a disclaimer and a bunch more

Reset the RAG - cutdown data

- rm vault.txt - to reset the system
- run upload.py and upload the cutdown version of the file
- run localrag.py
- “What is the passing score for the amateur radio exams?”

Correct answer, no disclaimer, closer

Lets take a look at our customized data

Run vi vault.txt

Lot of data, you would do a bit of cleanup.

The quality of the answers are based on the quality of the data going into it.

Next Steps

There are a lot of other RAGs out there. I like easy-local-rag because it is about ~150 lines of python code.

Training and updating models is another option. Vmware had some very nice examples, before the broadcomm stuff.

If you're just getting started Hackernoon has a really good article about building a $300 / AI PC. Ryzen 7 5700U is a good chip

Oracle has a free OCI Generative AI exam opportunity until July 31, 2024. Working on it currently.

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:3000

Looks a lot like the old bard 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 - Rag

Time for a bit of Rag

Load the documents in “My Documents”

Start a new conversation and include the document, and select a model.

Let’s ask it the question

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.

With Apple's new M4 chip having 38 teraflops of NPU performance and the new Microsoft Co-Pilot PCs requiring a minimum of 45 teraflops of NPU. There is going to be a lot of local power in the future.

Five 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
- Don't get married to on one model. Each has benefits, drawbacks, try a lot of them
- Hallucinations are true, when trying to find out a models limits it helps you if you have some baseline questions. E.g. amateur radio
- Raspberry PI's new AI top - is about 18 teraflops, and is $70 on top of the cost of a Raspberry Pi 5

Links

Network Chuck AI

- 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/

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

Oracle - Free Certification for OCI Generative AI

- https://education.oracle.com/genai/?source=:so:bl:or:awr:oun:::RC_WWMK240423P00002:RohitBlog

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

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

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