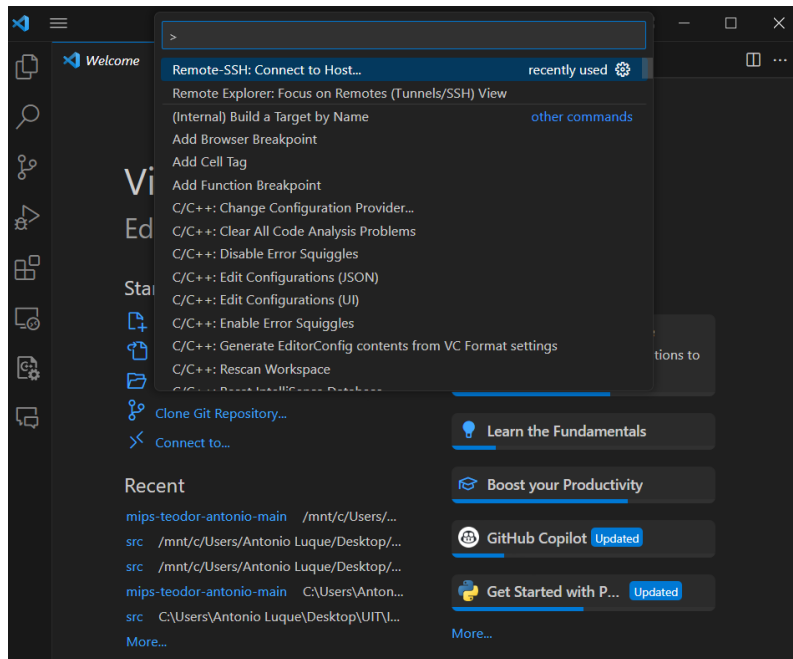
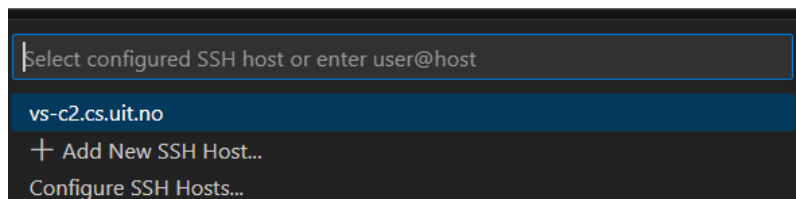


Fetching the public docker file related to privateGPT and executing them and exploring its ability.

We go to vs code and We press F1:

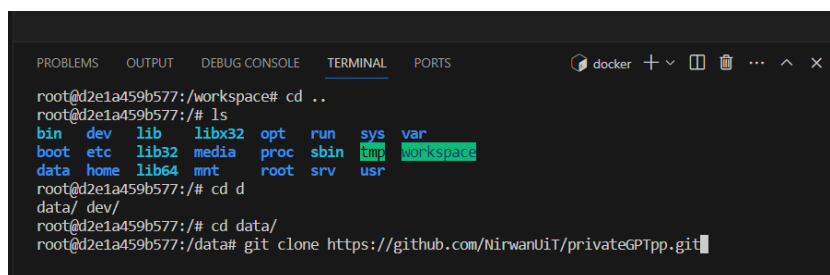
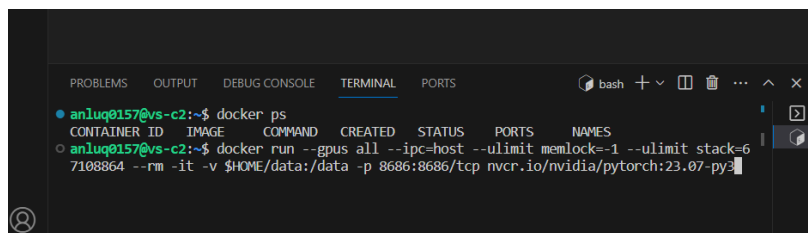


Then We press Remote-SSH: Connect to Host...



Here we add our server and We log in with our passwords.

We will download the docker image of the privateGPT onto our server with the following commands:



```
Cloning into 'privateGPTpp'...
remote: Enumerating objects: 210, done.
remote: Counting objects: 100% (79/79), done.
remote: Compressing objects: 100% (79/79), done.
remote: Total 210 (delta 44), reused 0 (delta 0), pack-reused 131
Receiving objects: 100% (210/210), 4.49 MiB | 11.64 MiB/s, done.
Resolving deltas: 100% (96/96), done.
root@d2e1a459b577:/data# ls
privateGPTpp
root@d2e1a459b577:/data# cd privateGPTpp/
root@d2e1a459b577:/data/privateGPTpp# pip install -r requirements.txt

WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

[notice] A new release of pip is available: 23.1.2 -> 23.3
[notice] To update, run: python -m pip install --upgrade pip
root@7cf0a922e5ef:/data/privateGPTpp# pip install llama-cpp-python==0.1.65 --force-reinstall --upgrade --no-cache-dir

WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

[notice] A new release of pip is available: 23.1.2 -> 23.3
[notice] To update, run: python -m pip install --upgrade pip
root@7cf0a922e5ef:/data/privateGPTpp# curl -s https://packagecloud.io/install/repositories/git/hub/git-lfs/script.deb.sh | bash

The repository is setup! You can now install packages.
root@7cf0a922e5ef:/data/privateGPTpp# apt-get install git-lfs

root@7cf0a922e5ef:/data/privateGPTpp# mkdir models
root@7cf0a922e5ef:/data/privateGPTpp# cd models/
root@7cf0a922e5ef:/data/privateGPTpp/models# wget https://gpt4all.io/models/ggml-gpt4all-j-v1.3-groovy.bin

root@7cf0a922e5ef:/data/privateGPTpp/models# wget https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML/blob/main/llama-2-7b-chat.ggmlv3.q4_0.bin
```

Here I had some errors where I wasn't able to run the llama model running the app.py

```
Found existing installation: llama-cpp-python 0.1.53
Uninstalling llama-cpp-python-0.1.53:
Successfully uninstalled llama-cpp-python-0.1.53
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.
argilla 1.16.0 requires numpy<1.24.0, but you have numpy 1.26.1 which is incompatible.
numba 0.57.1+1.gf851d279c requires numpy<1.25,>=1.21, but you have numpy 1.26.1 which is incompatible.
Successfully installed diskcache-5.6.3 llama-cpp-python-0.1.65 numpy-1.26.1 typing-extensions-4.8.0
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

[notice] A new release of pip is available: 23.1.2 -> 23.3
[notice] To update, run: python -m pip install --upgrade pip
```

I fixed it writing in the terminal inside the models:

wget https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML/resolve/main/llama-2-7b-chat.ggmlv3.q4_0.bin

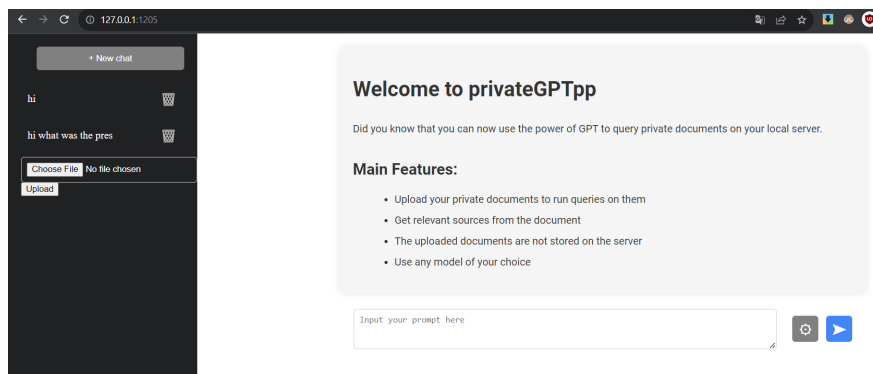
Then We go to **deploy-llm-project** and We write **vi app.py** in order to edit the port in the end in:

app.run(port=<port number>, host='0.0.0.0', debug=True) Before I had my port 8686 but after all I repeat it all and I put the port 1205, so, I had: **app.run(port=1205, host='0.0.0.0', debug=True)**

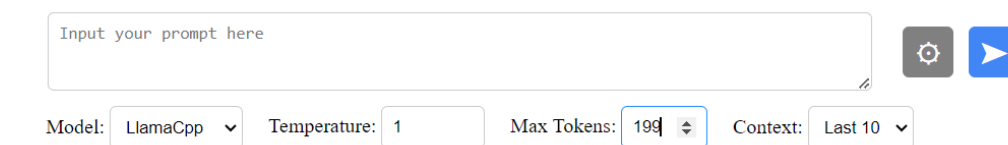
For executing I ran **python app.py** and I had:

```
root@22b10fda3688:/data/privateGPTpp/deploy-llm-project# python app.py
* Serving Flask app 'app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on all addresses (0.0.0.0)
* Running on http://127.0.0.1:1205
* Running on http://172.17.0.2:1205
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 234-994-523
172.17.0.1 - - [25/Oct/2023 09:12:36] "GET / HTTP/1.1" 200 -
172.17.0.1 - - [25/Oct/2023 09:12:41] "GET /SecondView HTTP/1.1" 200 -
172.17.0.1 - - [25/Oct/2023 09:12:41] "GET /styles.css HTTP/1.1" 404 -
* Detected change in '/data/privateGPTpp/deploy-llm-project/app.py', reloading
* Restarting with stat
* Debugger is active!
* Debugger PIN: 234-994-523
172.17.0.1 - - [25/Oct/2023 09:14:43] "GET / HTTP/1.1" 200 -
```

So, We have this result:



Exploring the internals of tokenization, context impact, prompt engineering, etc



Model: For this application We have these models: GPT4All, LlamaCpp, MedLlama, phi, codegeex2, codellama and vicuna like showed in the image:

```
match model_type:
    case "LlamaCpp":
        llm = LlamaCpp(model_path=model_path, max_tokens=model_n_ctx, n_batch=model_n_batch, callbacks=callbacks, verbose=False)
    case "GPT4All":
        llm = LlamaCpp(model_path="/data/privateGPTpp/models/llama-2-7b-chat-ggmlv3-q4_0.bin", n_ctx=model_n_ctx, verbose=False, n_gpu_layers=calculate_layer_count())
    case "GPT4All":
        llm = LlamaCpp(model_path="/data/privateGPTpp/models/ggml-gpt4all-j-v1.3-groovy.bin", backend="ggml", verbose=False)
    case "MedLlama":
        llm = HuggingFacePipeline.from_model_id(model_id="/data/privateGPTpp/models/medllama", task="text-generation", device=1,
        model_kwargs={"trust_remote_code": True, "torch_dtype": "auto", "max_length": model_n_ctx})
    case "phi":
        llm = HuggingFacePipeline.from_model_id(model_id="/data/privateGPTpp/models/phi-1.5", task="text-generation",
        model_kwargs={"trust_remote_code": True, "torch_dtype": "auto", "max_length": model_n_ctx})
    case "codegeex2":
        llm = HuggingFacePipeline.from_model_id(model_id="/data/privateGPTpp/models/codegeex2-6b", task="text-generation", device=1,
        model_kwargs={"trust_remote_code": True, "torch_dtype": "auto", "max_length": model_n_ctx})
    case "codellama":
        llm = HuggingFacePipeline.from_model_id(model_id="/data/privateGPTpp/models/Codellama-7b-hf", task="text-generation", device=1,
        model_kwargs={"trust_remote_code": True, "torch_dtype": "auto", "max_length": model_n_ctx})
    case "vicuna":
        llm = HuggingFacePipeline.from_model_id(model_id="/data/privateGPTpp/models/vicuna-7b-v1.5", task="text-generation", device=1,
        model_kwargs={"trust_remote_code": True, "torch_dtype": "auto", "max_length": model_n_ctx})
    case _default:
        # raise exception if model_type is not supported
        raise Exception(f"Model type {model_type} is not supported. Please choose one of the following: LlamaCpp, GPT4All")
```


Temperature: Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic

Max Tokens: The maximum number of tokens to generate in the chat completion. The total length of input tokens and generated tokens is limited by the model's context length

Context: How to select messages from the history to use them as prompt context

Try changing the openLLM between LlamaCpp and GPT4all and enter your prompt and see the output

Here I tried two questions for gpt4all and llama:

hi what was the president of america in 1988 

In 1988, the President of the United States was George H.W. Bush.

Sources :

I understand.

I remember when my Dad had to leave our home in Scranton, Pennsylvania to find work. I grew up in a family where if the price of food went up, you felt it.


That's why one of the first things I did as President was fight to pass the American Rescue Plan.

Because people were hurting. We needed to act, and we did.

Few pieces of legislation have done more in a critical moment in our history to lift us out of crisis.

Madam Speaker, Madam Vice President, our First Lady and Second Gentleman. Members of Congress and the Cabinet. Justices of the Supreme Court. My fellow Americans.

Last year COVID-19 kept us apart. This year we are finally together again.

Tonight, we meet as Democrats Republicans and Independents. But most importantly as Americans. 

With a duty to one another to the American people to the Constitution.

And with an unwavering resolve that freedom will always triumph over tyranny.



As I said last year, especially to our younger transgender Americans, I will always have your back as your President, so you can be yourself and reach your God-given potential.

While it often appears that we never agree, that isn't true. I signed 80 bipartisan bills into law last year. From preventing government shutdowns to protecting Asian-Americans from still-too-common hate crimes to reforming military justice.


To all Americans, I will be honest with you, as I've always promised. A Russian dictator, invading a foreign country, has costs around the world.

And I'm taking robust action to make sure the pain of our sanctions is targeted at Russia's economy. And I will use every tool at our disposal to protect American businesses and consumers.

Tonight, I can announce that the United States has worked with 30 other countries to release 60 Million barrels of oil from reserves around the world.

(gpt4all)

hi what was the president of america in 1988 

The President of America during this time period is not specified as it refers only to a specific year, while there have been multiple presidents throughout history with that title and different political parties or backgrounds

Sources :

I understand.

I remember when my Dad had to leave our home in Scranton, Pennsylvania to find work. I grew up in a family where if the price of food went up, you felt it.


That's why one of the first things I did as President was fight to pass the American Rescue Plan.

Because people were hurting. We needed to act, and we did.

Few pieces of legislation have done more in a critical moment in our history to lift us out of crisis.

Madam Speaker, Madam Vice President, our First Lady and Second Gentleman. Members of Congress and the Cabinet. Justices of the Supreme Court. My fellow Americans.

Last year COVID-19 kept us apart. This year we are finally together again.

Tonight, we meet as Democrats Republicans and Independents. But most importantly as Americans. 

With a duty to one another to the American people to the Constitution.

And with an unwavering resolve that freedom will always triumph over tyranny.



As I said last year, especially to our younger transgender Americans, I will always have your back as your President, so you can be yourself and reach your God-given potential.

While it often appears that we never agree, that isn't true. I signed 80 bipartisan bills into law last year. From preventing government shutdowns to protecting Asian-Americans from still-too-common hate crimes to reforming military justice.

To all Americans, I will be honest with you, as I've always promised. A Russian dictator, invading a foreign country, has costs around the world.

And I'm taking robust action to make sure the pain of our sanctions is targeted at Russia's economy. And I will use every tool at our disposal to protect American businesses and consumers.

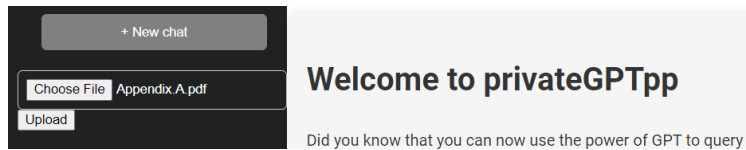
Tonight, I can announce that the United States has worked with 30 other countries to release 60 Million barrels of oil from reserves around the world.

(llama)

Try giving better prompt and see the output of various types of prompt

I uploaded a pdf of a chapter of a book of my Numerical methods course and I asked for a explanation of what is a vector in a easy way:



We press **upload** and we put this prompt:

Explain to me what is a vector in Matrix Algebra in a simple way like I was a 9 year old boy



Hey there, little buddy! So, you know how we can move things around with our hands, right? Like, if you hold up your right hand and wave it around, you can make things like toys or balls move around in the air. Well, matrices are kind of like that too!

A matrix is like a special kind of box that can hold lots of different things inside it. The things inside the box are called "vectors", and they can be thought of as little arrows that point in different directions. Just like how you can move your hand around to make things move, matrices can help us move these vectors around too!

So, when we multiply a matrix by a vector, it's like we're taking the vector and moving it around inside the box of the matrix. It's a way of combining the two things together to create something new and interesting! And just like how you can use your hands to make different things move, matrices can help us do lots of cool things with vectors too!

Sources :

Appendix A: Matrix Algebra

We begin with a short review of the basic definitions in matrix algebra.

A.1 MATRIX FUNDAMENTALS

A vector is an array of numbers

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Input your prompt here



Model: LlamaCpp

Temperature: 1

Max Tokens: 200

Context: Last 10

(llama)

Explain to me what is a vector in Matrix Algebra in a simple way like I was a 9 year old boy



A Vector In Mathematics Is Like An Arrow That Points To The Future.

Sources :

Appendix A: Matrix Algebra

We begin with a short review of the basic definitions in matrix algebra.

A.1 MATRIX FUNDAMENTALS

A vector is an array of numbers

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

If the list contains n numbers, it is called an n -dimensional vector. We will often make a distinction between the foregoing vertically arranged array, or column vector, and a horizontally arranged array

$\mathbf{u} = [u_1, \dots, u_n]$

called a row vector. An $m \times n$ matrix is an $m \times n$ array of numbers having the form

$$\mathbf{A} =$$

Input your prompt here



Model: GPT4All

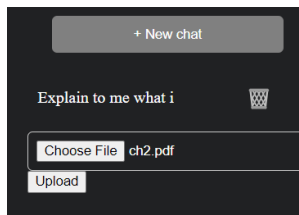
Temperature: 1

Max Tokens: 200

Context: Last 10

(gpt4all)

Report prompt used by you and outcome from privateGPT and record time taken by inference of each prompt.



I uploaded the chapter 2 of the Cloud Computing Course.

I put a clock in python to know the time spent in uploading the answer.

```
63 import time
64
# Get the answer from the chain
start = time.time()
res = qa(query)
answer, docs = res['result'], [] if hide_source else res['source_documents']
end = time.time()

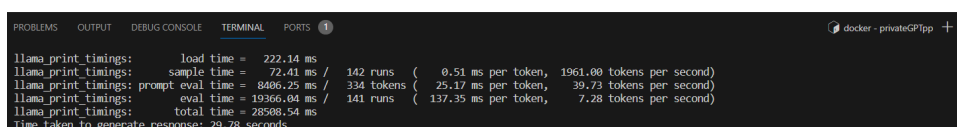
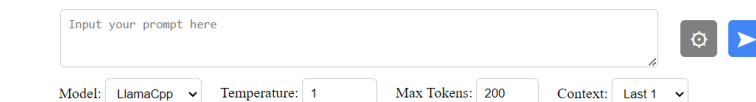
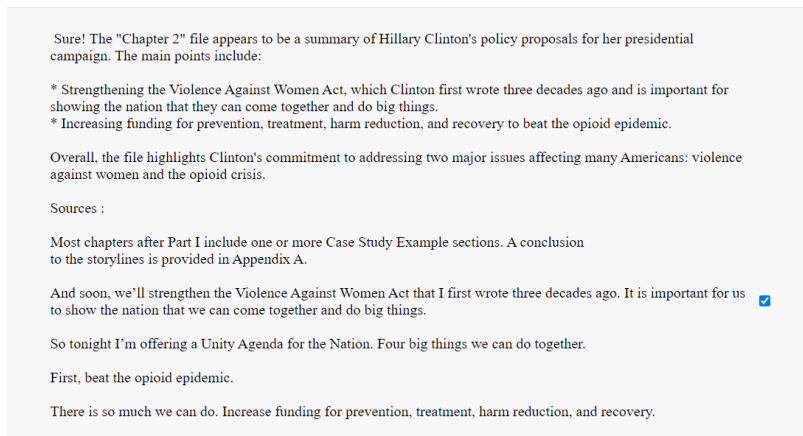
# Print the result
print("\n\n> Question:")
print(query)
print(f"\n\n> Answer (took {round(end - start, 2)} s.):")
print(answer)

# Printing Elapsed time
elapsed_time = round(end - start, 2) # Calculate the elapsed time
print(f"Time taken to generate response: {elapsed_time} seconds")

# Print the relevant sources used for the answer
sources = []
for document in docs:
    print("\n\n" + document.metadata["source"] + ":")
    print(document.page_content)
    # Append source and page content to sources list
    sources.append(document.metadata["source"] + ":" + document.page_content)

return answer, sources
```

summarize to me the ch2 file uploaded in an interesting and concise way underlining the essential parts



(Llama)

I didn't have a good output, so, I changed my prompt:

summarize to me the Case Study Background of the Chapter 2 uploaded in an interesting and concise way underlining the essential parts

Sure! Here's a summary of the Case Study Background of Chapter 2: ATN is a telecommunications supplier that has grown significantly over the years through acquisitions, leading to increased IT complexity. This complexity became a major concern for ATN's board of directors, which led to the inclusion of this case study in the book.

Sources :

Most chapters after Part I include one or more Case Study Example sections. A conclusion to the storylines is provided in Appendix A.

15

16

Chapter 2: Case Study Background



telecommunications supplier, adding another applications portfolio to its inventory. The IT complexity snowballed into a serious obstruction and became a source of critical concern to ATN's board of directors.

Most chapters after Part I include one or more Case Study Example sections. A conclusion to the storylines is provided in Appendix A.

2.1 Case Study #1: ATN

ATN is a company that provides network equipment to telecommunications industries

Input your prompt here



Model:

LlamaCpp

Temperature:

1

Max Tokens:

200

Context:

Last 1

PROBLEMS

OUTPUT

DEBUG CONSOLE

TERMINAL

PORTS

docker - privateGPTpp

+

▼

llama_print_timings:

load time = 237.42 ms

llama_print_timings:

sample time = 27.42 ms / 74 runs (0.51 ms per token, 1977.71 tokens per second)

llama_print_timings:

prompt eval time = 9538.87 ms / 387 tokens (24.65 ms per token, 40.57 tokens per second)

llama_print_timings:

eval time = 9953.47 ms / 73 runs (136.35 ms per token, 7.33 tokens per second)

llama_print_timings:

total time = 19846.55 ms

Time taken to generate response: 19.87 seconds

(llama) -> 19.87 seconds

Now, GPT4all:

summarize to me the ch2.pdf uploaded in an interesting and concise way underlining the essential parts

The article, "A Call for Leadership," discusses how President Clinton proposed that America make significant strides on global health issues through investment in ARPA-H (Advanced Research Projects Agency) at a time when he was facing criticism from Congress. He argued strongly about why investing resources to advance healthcare and develop new treatments can have long-term benefits, such as increased innovation for the country's future generations who would benefit greatly without being burdened by expensive health costs in their lifetime

Sources :

More support for patients and families.

To get there, I call on Congress to fund ARPA-H, the Advanced Research Projects Agency for Health.

It's based on DARPA—the Defense Department project that led to the Internet, GPS, and so much more.

ARPA-H will have a singular purpose—to drive breakthroughs in cancer, Alzheimer's, diabetes, and more.



A unity agenda for the nation.

We can do this.

My fellow Americans—tonight , we have gathered in a sacred space—the citadel of our democracy.

More support for patients and families.

Input your prompt here



Model:

GPT4All

Temperature:

1

Max Tokens:

200

Context:

Last 1

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 2
The article, "A Call for Leadership," discusses how President Clinton proposed that America make significant strides on global health issues through investment in ARPA-H (Advanced Research and Development Health Agency) at a time when he was facing criticism from Congress. He argued strongly about why investing resources to advance healthcare and develop new treatments can have long-term benefits and innovation for the country's future generations who would benefit greatly without being burdened by expensive health costs in their lifetime
Time taken to generate response: 113.78 seconds
```

(GPT4all)

I didn't have a good output, so, I changed my prompt like before with llama:

summarize to me the Case Study Background of the Chapter 2 uploaded in an interesting and concise way underlining the essential parts ☐

The IT complexity snowballed into a serious obstruction for ATN's board of directors. This case study demonstrates how technology can add more layers on top, making it challenging not only for internal teams but also external partners such as vendors to implement changes quickly or address challenges before they become critical issues impacting the company and its customers' experience with their networks

Sources :

Most chapters after Part I include one or more Case Study Example sections. A conclusion to the storylines is provided in Appendix A.

15

16

Chapter 2: Case Study Background

telecommunications supplier, adding another applications portfolio to its inventory. ☒

The IT complexity snowballed into a serious obstruction and became a source of critical concern to ATN's board of directors.

Most chapters after Part I include one or more Case Study Example sections. A conclusion to the storylines is provided in Appendix A.

2.1 Case Study #1: ATN

Model: GPT4All Temperature: 1 Max Tokens: 200 Context: Last 1

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 1
gptj_model load: ggml ctx size = 5401.45 MB
gptj_model load: kv self size = 896.00 MB
gptj_model load: ..... done
gptj_model load: model size = 3609.38 MB / num tensors = 285
The IT complexity snowballed into a serious obstruction for ATN's board of directors. This case study demonstrates how technology can add more layers on top, making it challenging not only for internal teams but also external partners such as vendors to implement changes quickly or address challenges before they become critical issues impacting the company and its customers' experience with their networks
Time taken to generate response: 75.1 seconds
```

(GPT4all) -> 75.1 seconds

Also list down the challenges faced by you during this colloquium

For me the biggest challenge has been to understand how all the code works and to make it work at the beginning because I had a lot of problems with the models when configuring the environment but now it is working perfectly and I understand more or less all the proposed code, I find all this LLM very interesting and I am feeling that I am learning a lot.

That's all, Thank you for reading.

Antonio.