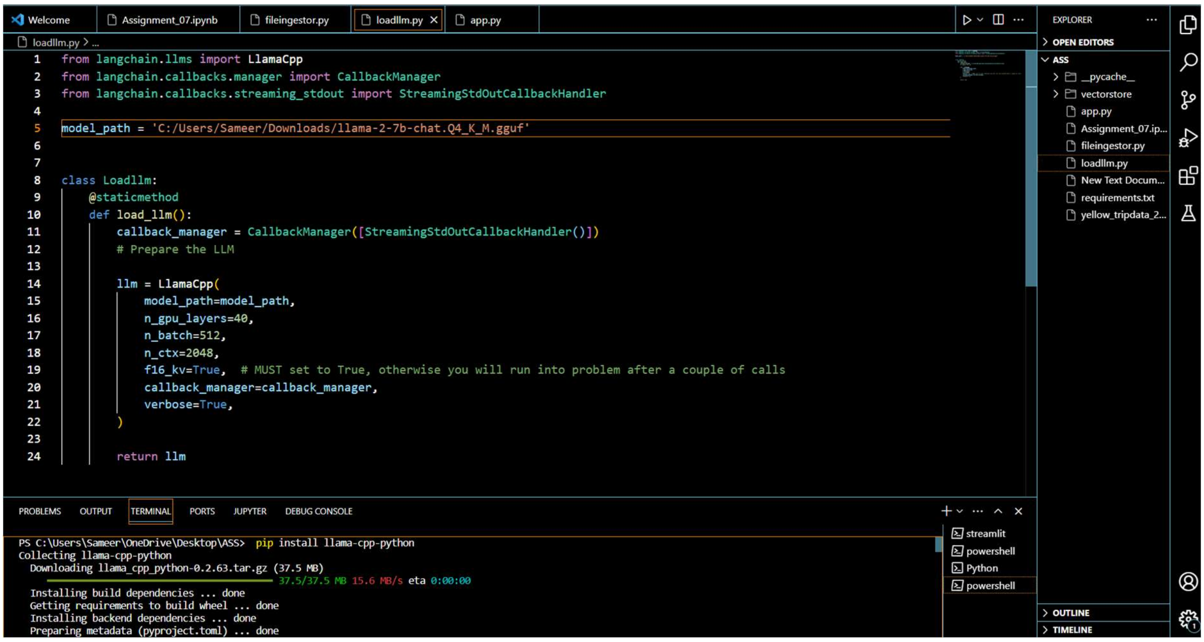
Name: Sameer Shaik

GWID: G49843839

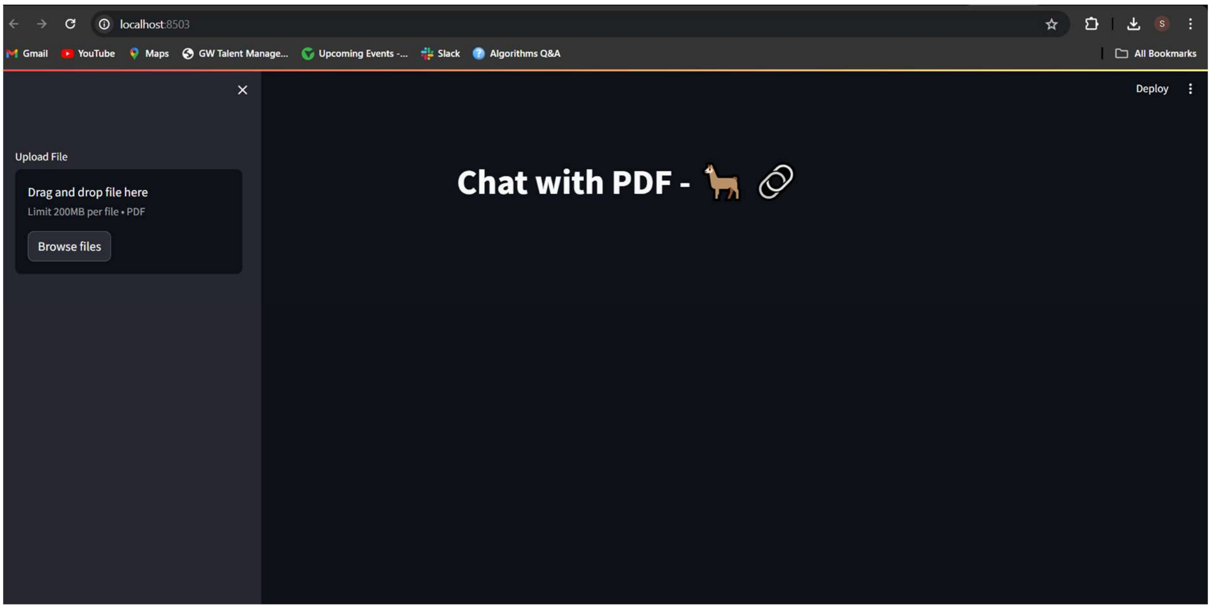
Sameer.shaik@gwmail.gwu.edu

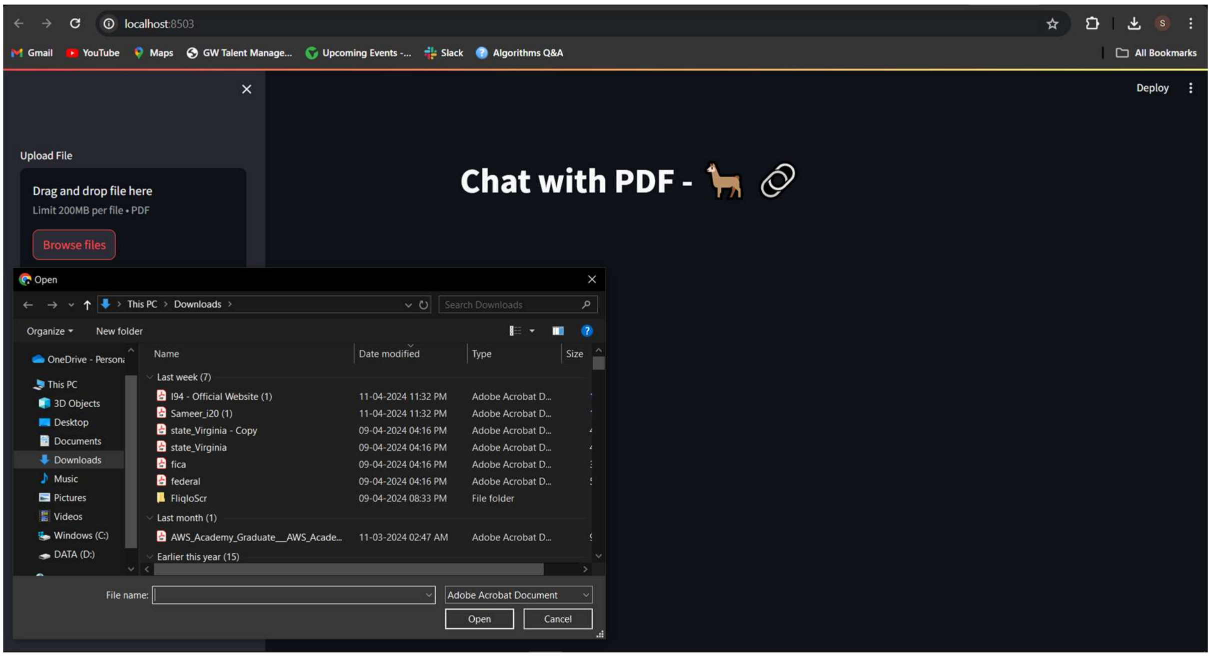
Problem 2:



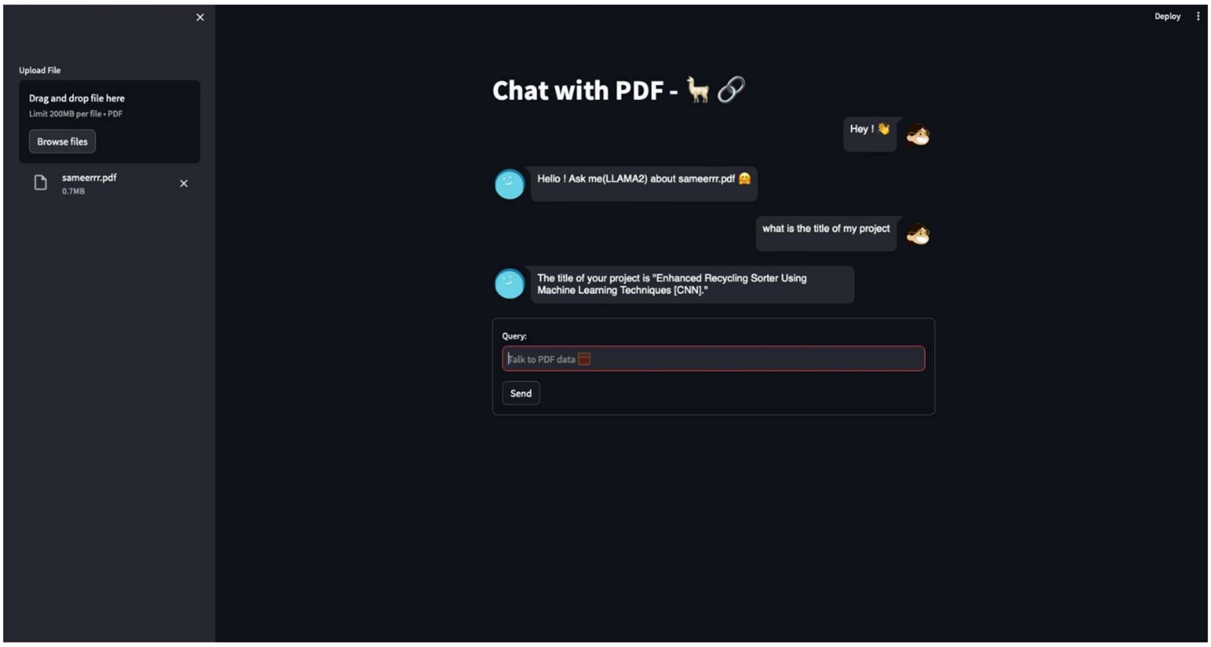
```python
from langchain.llms import LlamaCpp
from langchain.callbacks.manager import CallbackManager
from langchain.callbacks.streaming_stdout import StreamingStdOutCallbackHandler

model_path = 'C:/Users/Sameer/Downloads/llama-2-7b-chat.Q4_K_M.gguf'


class Loadllm:
    @staticmethod
    def load_llm():
        callback_manager = CallbackManager([StreamingStdOutCallbackHandler()])
        # Prepare the LLM

        llm = LlamaCpp(
            model_path=model_path,
            n_gpu_layers=40,
            n_batch=512,
            n_ctx=2048,
            f16_kv=True,  # MUST set to True, otherwise you will run into problem after a couple of calls
            callback_manager=callback_manager,
            verbose=True,
        )

        return llm
```

```
PS C:\Users\Sameer\OneDrive\Desktop\ASS>  pip install llama-cpp-python
Collecting llama-cpp-python
  Downloading llama_cpp_python-0.2.63.tar.gz (37.5 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 37.5/37.5 MB 15.6 MB/s eta 0:00:00
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Installing backend dependencies ... done
  Preparing metadata (pyproject.toml) ... done
```





Name	Date modified	Type	Size
Last week (7)			
I94 - Official Website (1)	11-04-2024 11:32 PM	Adobe Acrobat D...	
Sameer_i20 (1)	11-04-2024 11:32 PM	Adobe Acrobat D...	
state_Virginia - Copy	09-04-2024 04:16 PM	Adobe Acrobat D...	
state_Virginia	09-04-2024 04:16 PM	Adobe Acrobat D...	
fica	09-04-2024 04:16 PM	Adobe Acrobat D...	
federal	09-04-2024 04:16 PM	Adobe Acrobat D...	
FliqloScr	09-04-2024 08:33 PM	File folder	
Last month (1)			
AWS_Academy_Graduate__AWS_Acade...	11-03-2024 02:47 AM	Adobe Acrobat D...	
Earlier this year (15)			



Chat with PDF - 🦙 🔗

Hey ! 👋

Hello ! Ask me(LLAMA2) about sameerrr.pdf 🤗

what is the title of my project

The title of your project is "Enhanced Recycling Sorter Using Machine Learning Techniques [CNN]."

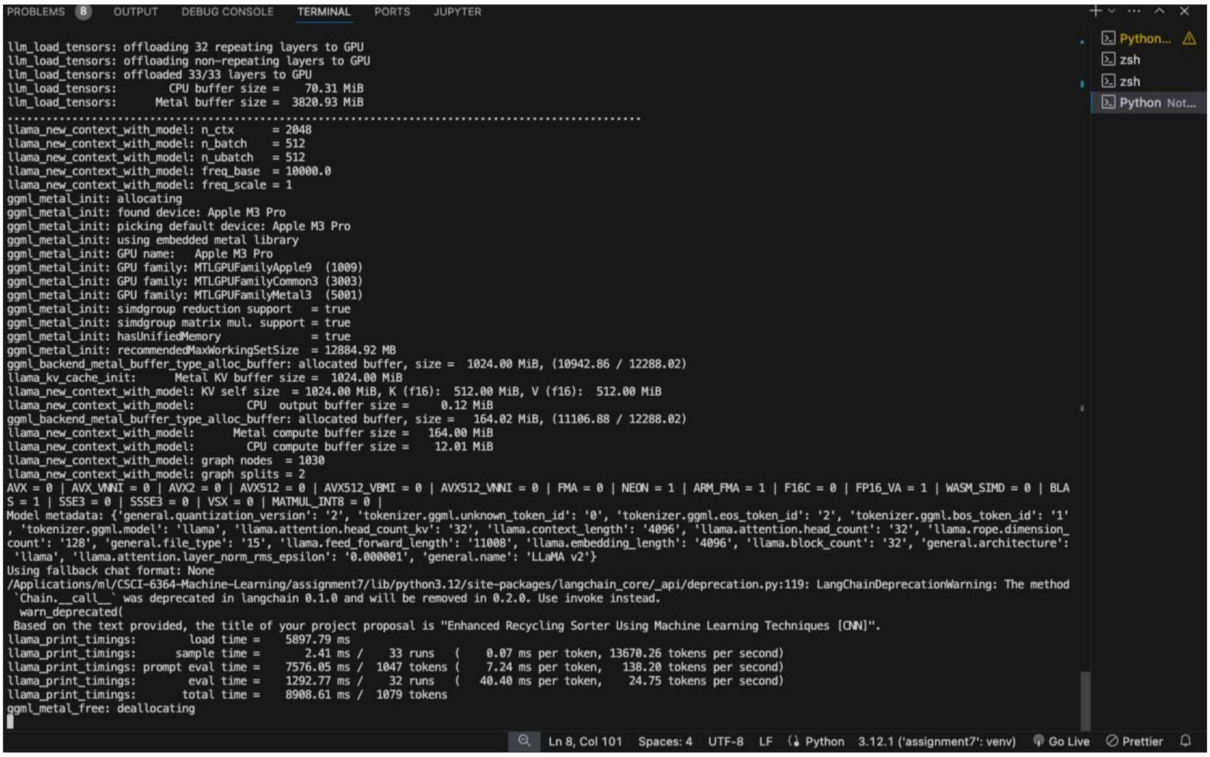
Prompt

What is the title of my project :

Response

The title of your project is “ Enhanced Recycling Sorter Using Machine Learning Techniques [CNN]”

Metrics :



```
llm_load_tensors: offloading 32 repeating layers to GPU
llm_load_tensors: offloading non-repeating layers to GPU
llm_load_tensors: offloaded 33/33 layers to GPU
llm_load_tensors:        CPU buffer size =    70.31 MiB
llm_load_tensors:      Metal buffer size =  3820.93 MiB
...................................................................................................
llama_new_context_with_model: n_ctx      = 2048
llama_new_context_with_model: n_batch    = 512
llama_new_context_with_model: n_ubatch   = 512
llama_new_context_with_model: freq_base  = 10000.0
llama_new_context_with_model: freq_scale = 1
ggml_metal_init: allocating
ggml_metal_init: found device: Apple M3 Pro
ggml_metal_init: picking default device: Apple M3 Pro
ggml_metal_init: using embedded metal library
ggml_metal_init: GPU name:   Apple M3 Pro
ggml_metal_init: GPU family: MTLGPUFamilyApple9  (1009)
ggml_metal_init: GPU family: MTLGPUFamilyCommon3 (3003)
ggml_metal_init: GPU family: MTLGPUFamilyMetal3  (5001)
ggml_metal_init: simdgroup reduction support   = true
ggml_metal_init: simdgroup matrix mul. support = true
ggml_metal_init: hasUnifiedMemory              = true
ggml_metal_init: recommendedMaxWorkingSetSize  = 12884.92 MB
ggml_backend_metal_buffer_type_alloc_buffer: allocated buffer, size =  1024.00 MiB, (10942.86 / 12288.02)
llama_kv_cache_init:      Metal KV buffer size =  1024.00 MiB
llama_new_context_with_model: KV self size  = 1024.00 MiB, K (f16):  512.00 MiB, V (f16):  512.00 MiB
llama_new_context_with_model:        CPU  output buffer size =     0.12 MiB
ggml_backend_metal_buffer_type_alloc_buffer: allocated buffer, size =   164.02 MiB, (11106.88 / 12288.02)
llama_new_context_with_model:      Metal compute buffer size =   164.00 MiB
llama_new_context_with_model:        CPU compute buffer size =    12.01 MiB
llama_new_context_with_model: graph nodes  = 1030
llama_new_context_with_model: graph splits = 2
AVX = 0 | AVX_VNNI = 0 | AVX2 = 0 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 0 | NEON = 1 | ARM_FMA = 1 | F16C = 0 | FP16_VA = 1 | WASM_SIMD = 0 | BLA
S = 1 | SSE3 = 0 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.quantization_version': '2', 'tokenizer.ggml.unknown_token_id': '0', 'tokenizer.ggml.eos_token_id': '2', 'tokenizer.ggml.bos_token_id': '1'
, 'tokenizer.ggml.model': 'llama', 'llama.attention.head_count_kv': '32', 'llama.context_length': '4096', 'llama.attention.head_count': '32', 'llama.rope.dimension_
count': '128', 'general.file_type': '15', 'llama.feed_forward_length': '11008', 'llama.embedding_length': '4096', 'llama.block_count': '32', 'general.architecture':
 'llama', 'llama.attention.layer_norm_rms_epsilon': '0.000001', 'general.name': 'LLaMA v2'}
Using fallback chat format: None
/Applications/ml/CSCI-6364-Machine-Learning/assignment7/lib/python3.12/site-packages/langchain_core/_api/deprecation.py:119: LangChainDeprecationWarning: The method
 `Chain.__call__` was deprecated in langchain 0.1.0 and will be removed in 0.2.0. Use invoke instead.
  warn_deprecated(
 Based on the text provided, the title of your project proposal is "Enhanced Recycling Sorter Using Machine Learning Techniques [CNN]".
llama_print_timings:        load time =    5897.79 ms
llama_print_timings:      sample time =       2.41 ms /    33 runs   (    0.07 ms per token, 13670.26 tokens per second)
llama_print_timings: prompt eval time =    7576.05 ms /  1047 tokens (    7.24 ms per token,   138.20 tokens per second)
llama_print_timings:        eval time =    1292.77 ms /    32 runs   (   40.40 ms per token,    24.75 tokens per second)
llama_print_timings:       total time =    8908.61 ms /  1079 tokens
ggml_metal_free: deallocating
```

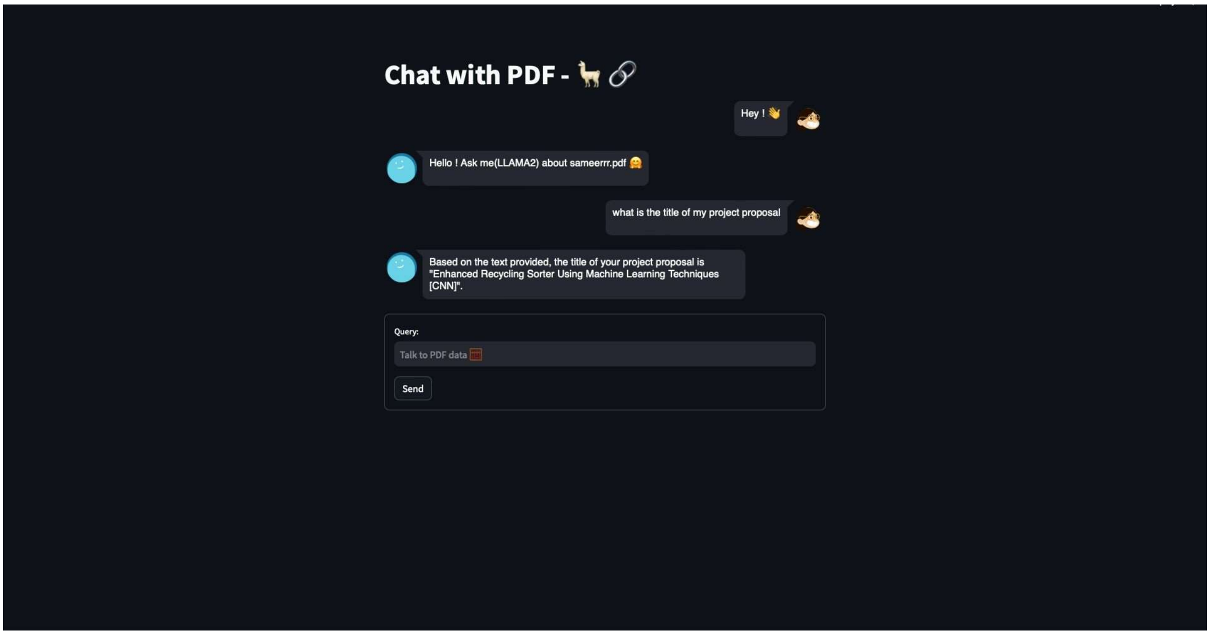
Prompt

What is the title of my project :

Response

The title of your project is “ Enhanced Recycling Sorter Using Machine Learning Techniques [CNN]”

Metrics :



Chat with PDF - 🦙 🔗

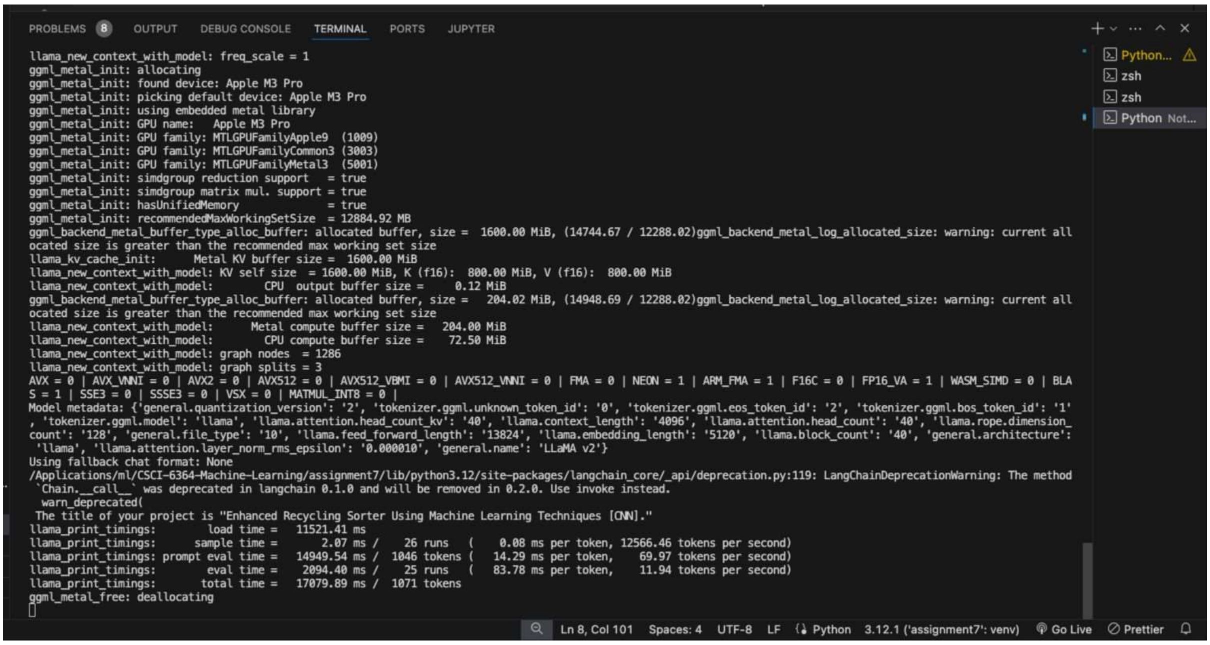
Hey ! 👋

Hello ! Ask me(LLAMA2) about sameerrr.pdf 🤗

what is the title of my project proposal

Based on the text provided, the title of your project proposal is "Enhanced Recycling Sorter Using Machine Learning Techniques [CNN]".

Metrics



```
llama_new_context_with_model: freq_scale = 1
ggml_metal_init: allocating
ggml_metal_init: found device: Apple M3 Pro
ggml_metal_init: picking default device: Apple M3 Pro
ggml_metal_init: using embedded metal library
ggml_metal_init: GPU name:   Apple M3 Pro
ggml_metal_init: GPU family: MTLGPUFamilyApple9  (1009)
ggml_metal_init: GPU family: MTLGPUFamilyCommon3 (3003)
ggml_metal_init: GPU family: MTLGPUFamilyMetal3  (5001)
ggml_metal_init: simdgroup reduction support   = true
ggml_metal_init: simdgroup matrix mul. support = true
ggml_metal_init: hasUnifiedMemory              = true
ggml_metal_init: recommendedMaxWorkingSetSize  = 12884.92 MB
ggml_backend_metal_buffer_type_alloc_buffer: allocated buffer, size =  1600.00 MiB, (14744.67 / 12288.02)ggml_backend_metal_log_allocated_size: warning: current all
ocated size is greater than the recommended max working set size
llama_kv_cache_init:      Metal KV buffer size =  1600.00 MiB
llama_new_context_with_model: KV self size  = 1600.00 MiB, K (f16):  800.00 MiB, V (f16):  800.00 MiB
llama_new_context_with_model:        CPU  output buffer size =     0.12 MiB
ggml_backend_metal_buffer_type_alloc_buffer: allocated buffer, size =   204.02 MiB, (14948.69 / 12288.02)ggml_backend_metal_log_allocated_size: warning: current all
ocated size is greater than the recommended max working set size
llama_new_context_with_model:      Metal compute buffer size =   204.00 MiB
llama_new_context_with_model:        CPU compute buffer size =    72.50 MiB
llama_new_context_with_model: graph nodes  = 1286
llama_new_context_with_model: graph splits = 3
AVX = 0 | AVX_VNNI = 0 | AVX2 = 0 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 0 | NEON = 1 | ARM_FMA = 1 | F16C = 0 | FP16_VA = 1 | WASM_SIMD = 0 | BLA
S = 1 | SSE3 = 0 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.quantization_version': '2', 'tokenizer.ggml.unknown_token_id': '0', 'tokenizer.ggml.eos_token_id': '2', 'tokenizer.ggml.bos_token_id': '1'
, 'tokenizer.ggml.model': 'llama', 'llama.attention.head_count_kv': '40', 'llama.context_length': '4096', 'llama.attention.head_count': '40', 'llama.rope.dimension_
count': '128', 'general.file_type': '10', 'llama.feed_forward_length': '13824', 'llama.embedding_length': '5120', 'llama.block_count': '40', 'general.architecture':
 'llama', 'llama.attention.layer_norm_rms_epsilon': '0.000010', 'general.name': 'LLaMA v2'}
Using fallback chat format: None
/Applications/ml/CSCI-6364-Machine-Learning/assignment7/lib/python3.12/site-packages/langchain_core/_api/deprecation.py:119: LangChainDeprecationWarning: The method
 `Chain.__call__` was deprecated in langchain 0.1.0 and will be removed in 0.2.0. Use invoke instead.
  warn_deprecated(
 The title of your project is "Enhanced Recycling Sorter Using Machine Learning Techniques [CNN]."
llama_print_timings:        load time =   11521.41 ms
llama_print_timings:      sample time =       2.07 ms /    26 runs   (    0.08 ms per token, 12566.46 tokens per second)
llama_print_timings: prompt eval time =   14949.54 ms /  1046 tokens (   14.29 ms per token,    69.97 tokens per second)
llama_print_timings:        eval time =    2094.40 ms /    25 runs   (   83.78 ms per token,    11.94 tokens per second)
llama_print_timings:       total time =   17079.89 ms /  1071 tokens
ggml_metal_free: deallocating
```

Problem 3:

Prompt - Write about Neural Networks

##### top_k=40, top_p=0.95, temperature=0.8, max_tokens=25

Neural Networks and Their Applications

Introduction

Artificial neural networks have been a rapidly evolving field of research in recent years, with applications in various industries such as computer vision, natural language processing, speech recognition, and predictive analytics. This article will provide an overview of neural networks, their architecture, types, and applications.

What are Neural Networks?

Neural networks are a computational model inspired by the structure and function of the human brain. They consist of interconnected nodes or neurons that process information and learn from data. Each neuron receives input from other neurons or external sources, performs computations on that input, and then sends the output to other neurons or to the final output layer.

The architecture of a neural network typically includes three layers: input, hidden, and output. The input layer takes in the raw data, the hidden layer performs complex computations and feature extraction, and the output layer generates the final predictions or decisions. The connections between neurons are adjusted during training to minimize the error between the predicted output and the actual output.

Prompt - Write about Neural Networks

#### top_k=70, top_p=0.55, temperature=0.6, max_tokens=500

Neural Networks and Their Applications

Introduction

Neural networks have revolutionized the field of artificial intelligence and machine learning. These networks are modeled after the human brain and are capable of learning and making decisions based on data. In this article, we will explore the basics of neural networks and their applications in various fields.

Basics of Neural Networks

A neural network consists of interconnected nodes or neurons that process information. Each neuron receives input from other neurons or external sources, performs a computation on that input, and then sends the output to other neurons or to the output layer. The network learns by adjusting the weights and biases of the connections between neurons based on the data it is trained on.

There are several types of neural networks, including:

1. Feedforward Networks: These networks process information in only one direction, from input to output, without any feedback loops.
2. Recurrent Networks: These networks have feedback connections, allowing them to process information in a loop. This allows the network to keep track of information over time.
3. Convolutional Networks: These networks are designed to process data with grid-like topology, such as images.

Applications of Neural Networks

Neural networks have numerous applications in various fields, including:

1. Image Recognition: Neural networks can be trained to recognize objects in images, such as faces, animals, or vehicles.
2. Natural Language Processing: Neural networks can be used to process and analyze natural language data, such as text or speech.
3. Predictive Modeling: Neural networks can be used to predict future events or outcomes based on past data.
4. Robotics: Neural networks can be used to control robots and enable them to learn from experience.
5. Healthcare: Neural networks can be used in medical imaging, disease diagnosis, and drug discovery.
6. Finance: Neural networks can be used for fraud detection, credit risk assessment, and portfolio optimization.
7. Recommendation Systems: Neural networks can be used to recommend products or services based on user behavior and preferences.
8. Autonomous Vehicles: Neural networks can be used to control autonomous vehicles and