Short Report on Local LLM

1. Setup Steps

I set up the Local LLM on a **MacBook Air** following the prerequisites:

- Cloned the **CS595 Lab GitHub repository** and pulled the latest updates in VS Code.
- Verified Python 3.10 was installed, then created and activated a **Python virtual environment**.
- Downloaded required model files:
    - `Meta-Llama-3-8B-Instruct-Q4_K_S.gguf`
    - `Phi-3-mini-4k-instruct.Q6_K.gguf`
      and placed them in the specified folder.

Installed dependencies:
`pip install -r requirements.txt`

- in the `/labs/local_llm` directory.
- Implemented functions for **text chunking, embedding, searching similar chunks**, and running the LLM in `local_llm.ipynb`.
- Ran the notebook in VS Code and successfully completed the Local LLM environment setup.

2. Errors Encountered and Solutions

- **Tuple Data Error in PDF Chunks**
    - **Issue:** `TypeError: sequence item 0: expected str instance, tuple found` when joining results from `search_similar_chunks`.
    - **Solution:** Extracted text from each tuple using a list comprehension before joining.
- **Long Query Execution Times**
    - **Issue:** Queries with PDF context took 30+ minutes each.
    - **Solution:** Tested individual components (chunking, embedding, querying) with smaller inputs to debug efficiently before running full queries.

3. Comparison of Results With and Without Context

Below are the six queries I tested to evaluate the model's performance with and without access to the provided context :

Query 1: "What is the average CGM value for my patient?"
With PDF Context: The LLM accurately identified that no Continuous Glucose Monitoring (CGM) data was available in the clinical profile. It referenced relevant metrics like the patient's last HbA1c (8.3%) and fasting glucose (160 mg/dL), explaining the limitations of estimating CGM values without continuous glucose readings.
Without PDF Context: Without the PDF, the LLM provided a generic response, asking for more details about the patient and the specific type of CGM data required. It lacked the ability to reference any actual values from the profile.

Evidence/ Screenshot :
Query 1 :_"Whats is the avg cgm values for my patient?"

```python
new_query = "Whats is the avg cgm values for my patient?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[31] ✓ 34m 24.4s Python

Output With PDF Context:

```
[31] ✓ 34m 24.4s Python
...
    llama_print_timings:        load time =   29712.45 ms
    llama_print_timings:      sample time =      77.38 ms /   138 runs   (    0.56 ms per token,  1783.43 tokens per second)
    llama_print_timings: prompt eval time =   73781.48 ms /  1235 tokens (   59.74 ms per token,    16.74 tokens per second)
    llama_print_timings:        eval time = 1022921.73 ms /   137 runs   ( 7466.58 ms per token,     0.13 tokens per second)
    llama_print_timings:       total time = 1098851.14 ms /  1372 tokens
    Llama.generate: prefix-match hit
    Q: Whats is the avg cgm values for my patient?
    A: Unfortunately, there is no Continuous Glucose Monitoring (CGM) data provided in the given clinical profile. The only available glucose measurements are:

    * Last HbA1c: 8.3%
    * Fasting Glucose: 160 mg/dL

    To estimate average CGM values, we would need more frequent and continuous glucose readings over a longer period. However, based on the provided information, we can infer that the pat

    If you'd like to estimate average CGM values, I can help you with that!
```

Output Without PDF Context:

```
    llama_print_timings:        load time =   29712.45 ms
    llama_print_timings:      sample time =      87.02 ms /   128 runs   (    0.68 ms per token,  1470.88 tokens per second)
    llama_print_timings: prompt eval time =    7686.69 ms /    16 tokens (  480.42 ms per token,     2.08 tokens per second)
    llama_print_timings:        eval time =  948761.25 ms /   127 runs   ( 7470.56 ms per token,     0.13 tokens per second)
    llama_print_timings:       total time =  958446.46 ms /   143 tokens
    Q: Whats is the avg cgm values for my patient?
    A: I'd be happy to help! However, I need more information from you. Could you please provide me with your patient's name or any relevant medical history? Additionally, what type of av

    1. Average blood glucose levels?
    2. Average glucose variability?
    3. Average time-in-range (TIR)?
    4. Average time-above-range (TAR)?
    5. Average time-below-range (TBR)?

    Please provide me with more context or clarify what specific information you're looking for, and I'll do my best to assist you!
```

Query 2: "What is the patient's ID?"
With PDF Context: The LLM responded by stating that the clinical profile does not include a patient ID. It demonstrated awareness of the document's structure and clearly communicated the absence of the requestedinformation. **Without PDF Context:** The LLM gave a vague response, suggesting that it couldn't find any specific patient information and offering to create a new patient record.

Evidence/ Screenshot :

Query 2 : "What is the patient's ID?"

```python
new_query = "What is the patient's ID?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[32] ✓ 26m 12.3s Python

```
... Llama.generate: prefix-match hit
```

Output With PDF Context:

```
new_query = "What is the patient's ID?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[32] ✓ 26m 12.3s Python

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      44.32 ms /    88 runs   (    0.50 ms per token,  1985.65 tokens per second)
llama_print_timings: prompt eval time =   77575.92 ms /  1214 tokens (   63.90 ms per token,    15.65 tokens per second)
llama_print_timings:        eval time =  655105.17 ms /    87 runs   ( 7529.94 ms per token,     0.13 tokens per second)
llama_print_timings:       total time =  734213.15 ms /  1301 tokens
Llama.generate: prefix-match hit
Q: What is the patient's ID?
A: I'm happy to help! However, I must clarify that there is no "ID" mentioned in the provided clinical profile. The information provided appears to be a comprehensive medical history
```

Output Without PDF Context:

```
llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      57.52 ms /   113 runs   (    0.51 ms per token,  1964.50 tokens per second)
llama_print_timings: prompt eval time =    9260.89 ms /    12 tokens (  771.74 ms per token,     1.30 tokens per second)
llama_print_timings:        eval time =  819704.54 ms /   112 runs   ( 7318.79 ms per token,     0.14 tokens per second)
llama_print_timings:       total time =  830691.72 ms /   124 tokens
Q: What is the patient's ID?
A: I'm happy to help! However, I'm a large language model, I don't have access to any specific patient information or medical records. Each time you interact with me, it's a new conve

If you're looking for assistance with a specific patient's care or need help with a particular medical issue, please feel free to provide more context or details about the patient, su
```

Query 3: "What is the patient's last recorded blood pressure?"

With PDF Context: The LLM noted that the clinical profile does not include a specific blood pressure reading. However, it highlighted that the patient has hypertension, which is controlled with medication (Lisinopril 10 mg daily). This response was aligned with the information provided in the profile. Without PDF Context: Without the PDF, the LLM gave a generic answer, stating that it doesn't have access to the patient's records and suggesting checking medical charts for the information. However, it also guided the user on how to calculate or track the blood pressure .

Evidence/Screenshot:

Query 3 : What is the patient's last recorded blood pressure?

```
new_query = "What is the patient's last recorded blood pressure?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[33] ✓ 14m 16.2s Python

```
Llama.generate: prefix-match hit
```

Output With PDF Context:

```
new_query = "What is the patient's last recorded blood pressure?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[33] ✓ 14m 16.2s Pytho

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      20.14 ms /    41 runs   (    0.49 ms per token,  2036.05 tokens per second)
llama_print_timings: prompt eval time =   74247.31 ms /  1217 tokens (   61.01 ms per token,    16.39 tokens per second)
llama_print_timings:        eval time =  294330.49 ms /    40 runs   ( 7358.26 ms per token,     0.14 tokens per second)
llama_print_timings:       total time =  369252.04 ms /  1257 tokens
Llama.generate: prefix-match hit
Q: What is the patient's last recorded blood pressure?
A: The patient's medical history does not mention their last recorded blood pressure. However, it does mention that they have hypertension, which is controlled with medication (Lisi
```

Output Without PDF Context:

```
llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      32.38 ms /    65 runs   (    0.50 ms per token,  2007.60 tokens per second)
llama_print_timings: prompt eval time =    7311.43 ms /    15 tokens (  487.43 ms per token,     2.05 tokens per second)
llama_print_timings:        eval time =  470069.19 ms /    64 runs   ( 7344.83 ms per token,     0.14 tokens per second)
llama_print_timings:       total time =  478424.94 ms /    79 tokens
Q: What is the patient's last recorded blood pressure?
A: I apologize, but I'm a large language model, I don't have access to any specific patient information, including their blood pressure readings. If you're looking for a patient's med
```

Query 4: "What is the patient's BMI?"
With PDF Context: The LLM correctly retrieved the patient's BMI from the clinical profile (31 kg/m²) and recognized it as indicative of obesity. This demonstrates the LLM's ability to extract precise data when context is available.Without PDF Context: The LLM asked for additional details, such as the patient's height and weight, to calculate the BMI. It was unable to provide an accurate response without access to the document.

Evidence/Screenshot:
Query 4 : What is the patient's BMI?

```
new_query = "What is the patient's BMI?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[34] ✓ 9m 14.1s Python

```
Llama.generate: prefix-match hit
```

Output With PDF Context:

```
new_query = "What is the patient's BMI?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[34] ✓ 9m 14.1s Pytho

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      13.48 ms /    24 runs   (    0.56 ms per token,  1780.68 tokens per second)
llama_print_timings: prompt eval time =   68876.60 ms /  1214 tokens (   56.74 ms per token,    17.63 tokens per second)
llama_print_timings:        eval time =  169817.40 ms /    23 runs   ( 7383.37 ms per token,     0.14 tokens per second)
llama_print_timings:       total time =  239128.61 ms /  1237 tokens
Llama.generate: prefix-match hit
Q: What is the patient's BMI?
A: According to the clinical profile, the patient's BMI is 31 kg/m², which is classified as obese.
```

Output Without PDF Context:

```
llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      20.97 ms /    39 runs   (    0.54 ms per token,  1859.53 tokens per second)
llama_print_timings: prompt eval time =    7772.99 ms /    12 tokens (  647.75 ms per token,     1.54 tokens per second)
llama_print_timings:        eval time =  299777.04 ms /    38 runs   ( 7888.87 ms per token,     0.13 tokens per second)
llama_print_timings:       total time =  308118.83 ms /    50 tokens
Q: What is the patient's BMI?
A: I'm happy to help! However, I need more information from you. Can you please provide me with the patient's height and weight so I can calculate their Body Mass Index (BMI)?
```

+ Code + Markdown

Query 5: "What is today's date?"
With PDF Context: The LLM indicated that there was no date mentioned in the provided clinical profile. It explained that the profile contains medical data but lacks any specific time or date references.Without PDF Context: The LLM provided a general response, acknowledging its lack of real-time access to the current date. It suggested using alternative methods such as checking a phone, computer, or online search to determine today's Date.
Evidence/Screenshot:
Query 5 : what is today's date? (Testing irrelevant question - non medical questions)

```
new_query = "what is today's date?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[35] ✓ 23m 29.4s Python

```
Llama.generate: prefix-match hit
```

Output With PDF Context:

```
new_query = "what is today's date?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

[35] ✓ 23m 29.4s Python

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      33.19 ms /    69 runs   (    0.48 ms per token,  2078.69 tokens per second)
llama_print_timings: prompt eval time =   76777.27 ms /  1213 tokens (   63.30 ms per token,    15.80 tokens per second)
llama_print_timings:        eval time =  513725.66 ms /    68 runs   ( 7554.79 ms per token,     0.13 tokens per second)
llama_print_timings:       total time =  591668.12 ms /  1281 tokens
Llama.generate: prefix-match hit
Q: what is today's date?
A: I apologize, but there is no information provided about the current date in the given clinical profile. The patient's medical history, demographics, and lifestyle factors are all
```

Output Without PDF Context:

```
llama_print_timings:        load time =   29712.45 ms
llama_print_timings:      sample time =      42.25 ms /    82 runs   (    0.52 ms per token,  1940.78 tokens per second)
llama_print_timings: prompt eval time =    7022.53 ms /    11 tokens (  638.41 ms per token,     1.57 tokens per second)
llama_print_timings:        eval time =  802040.81 ms /    81 runs   ( 9901.74 ms per token,     0.10 tokens per second)
llama_print_timings:       total time =  810441.89 ms /    92 tokens
Q: what is today's date?
A: I'm happy to help! However, I'm a large language model, I don't have real-time access to the current date and time. But I can tell you that as of our conversation starting, my trai
```

Query 6: "Is the patient feeling numbness?"
With PDF Context: The LLM confirmed that the patient is experiencing tingling and numbness in their feet, attributing it to diabetic neuropathy, which is consistent with the clinical profile's information.Without PDF Context: Without access to the PDF, the LLM gave a more general response, explaining that numbness could be caused by various conditions and suggesting consultation with a healthcare professional. It asked for more details about the symptoms.

Evidence/Screenshot:
Query 6 : Is the patient feeling numbness?

```
new_query = "Is the patient feeling numbness?"
# First call: Uses PDF context
new_answer = run_llm_with_pdf_knowledge(new_query)
print("Q:", new_query)
print("A:", new_answer)
# Second call: Direct LLM query without context
new_answer = run_llm(new_query)
print("Q:", new_query)
print("A:", new_answer)
```

✓ 17m 18.8s Python

```
Raw token embeddings shape: (1, 8, 4096)
```

Output With PDF Context:

```
Raw token embeddings shape: (1, 8, 4096)
Llama.generate: prefix-match hit

llama_print_timings:        load time =   35326.74 ms
llama_print_timings:      sample time =       8.23 ms /    27 runs   (    0.30 ms per token,  3281.08 tokens per second)
llama_print_timings: prompt eval time =   55951.65 ms /  1214 tokens (   46.09 ms per token,    21.70 tokens per second)
llama_print_timings:        eval time =  168981.65 ms /    26 runs   ( 6499.29 ms per token,     0.15 tokens per second)
llama_print_timings:       total time =  225215.13 ms /  1240 tokens
Llama.generate: prefix-match hit
Q: Is the patient feeling numbness?
A: According to the clinical profile, yes, the patient is experiencing tingling and numbness in their feet due to diabetic neuropathy.
```

Output Without PDF Context:

```
llama_print_timings:        load time =   35326.74 ms
llama_print_timings:      sample time =      41.12 ms /   123 runs   (    0.33 ms per token,  2991.03 tokens per second)
llama_print_timings: prompt eval time =    6269.05 ms /    12 tokens (  522.42 ms per token,     1.91 tokens per second)
llama_print_timings:        eval time =  798644.63 ms /   122 runs   ( 6546.27 ms per token,     0.15 tokens per second)
llama_print_timings:       total time =  806132.46 ms /   134 tokens
Q: Is the patient feeling numbness?
A: I'm happy to help! As a helpful assistant, I don't have direct access to the patient's medical information or physical presence. However, I can provide general information and guid

Numbness is a common symptom that can be caused by various factors such as nerve damage, compression, or inflammation. If you're concerned about someone's numbness, it's essential to

Can you please provide more context or information about the patient's symptoms? For instance, where exactly are they feeling numbness, and when did it start?
```

Observation: Context significantly improves LLM responses, making them accurate and patient-specific. Without context, responses are vague and less reliable.

4. Reflection: LLMs in Healthcare

- With context, LLMs can provide **accurate, patient-specific answers** from clinical data.
- Without structured context, responses are **generic and limited**.
- LLMs are promising for **summarizing patient information, answering queries, and assisting clinical decisions**, but require high-quality input data for reliability.