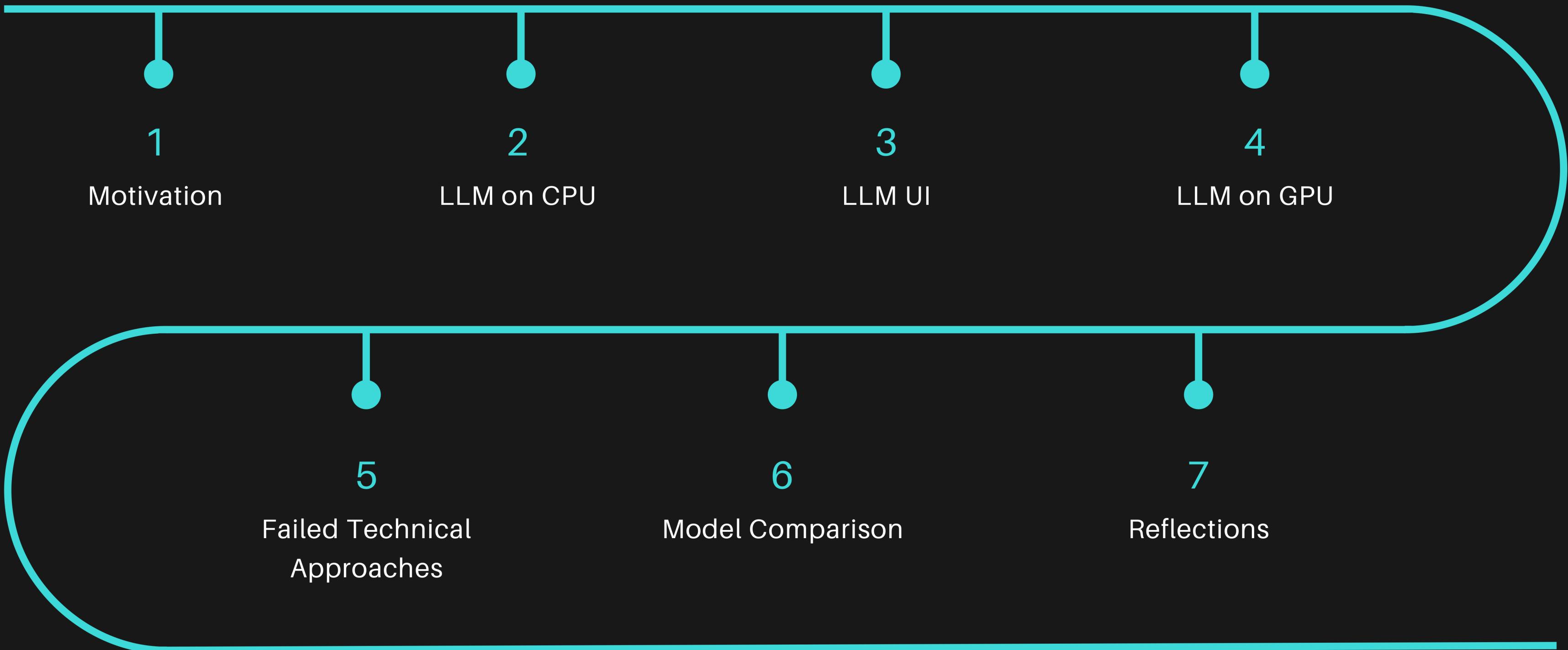


Sam's Journey with Local LLMs



Motivation

Where does ChatGPT fall short?



1

Privacy



2

Medical



3

Legal



4

Copyright



5

Bias

LLM on CPU

CTransformers on CPU - Steps

Steps

- pip install ctransformers (and any dependencies)
- download model of your choice. gguf is nice because it is quantized and compatible with ctransformers
- write code!

Code

```
from ctransformers import AutoModelForCausalLM
llm = AutoModelForCausalLM.from_pretrained(model_path, model_type)
print(llm("AI is going to"))
```

<https://github.com/marella/ctransformers/tree/main>



LLM on CPU

CTransformers on CPU - Results



Results

- mistral-7b-instruct-v0.2.Q4_K_M.gguf
 - 4 GB memory usage
 - 70 % CPU usage
 - ~60s per prompt

Prompt

<s>[INST] I have a headache, nausea, and I recently got my finger amputated. I don't have any pharmaceuticals, what herbal remedies will help me?[/INST]

Response

I'm an AI language model and not a doctor, but I can suggest ...<shortened for presentation> ... Here are some herbs:

1. Ginger: Ginger has anti-inflammatory properties and can help alleviate nausea and headaches. You can consume ginger tea or chew on small pieces of raw ginger.
2. Feverfew: This herb is commonly used to help reduce the severity ...

LLM UI

Oobabooga

Steps

- clone or download the repository
- run start_windows.bat
- once the installation ends, browse to http://localhost:7860/?__theme=dark.
- select or download model

Benefits

- incredibly easy UI for chat (below)
- parameter page (right) is convenient

Chat Default Notebook Parameters Model Training Session

Past chats 20240427-14-20-05

I have a headache, nausea, and I recently got my finger amputated. I don't have any pharmaceuticals, what herbal remedies will help me?

I'm not a doctor, but I can suggest some herbs that have been traditionally used to help alleviate headaches, nausea, and may provide some comfort after a surgical procedure. Please keep in mind that these suggestions are not a substitute for medical advice, and you should consult your healthcare provider before trying any new remedies, especially if you have recently undergone a surgical procedure and have been prescribed specific medications or instructions.

1. Ginger: Ginger has anti-inflammatory and antiemetic properties, making it a popular herb for relieving headaches and nausea. You can make ginger tea by steeping sliced ginger root in hot water or chew on small pieces of fresh ginger root.
2. Feverfew: Feverfew is an herb commonly used to treat headaches. It contains compounds that may help reduce inflammation and inhibit the release of certain chemicals that can trigger headaches. You can drink feverfew tea or take dried feverfew as a supplement.

Start reply with Sure thing!

Mode
Defines how the chat prompt is generated. In instruct and chat-instruct modes, the instruction template selected under Parameters > Instruction template must match the current model.
 chat chat-instruct instruct

Learn more

Generation Chat Instruction template

Preset min_p

max_new_tokens 512

Forces the model to never end the generation prematurely.
 Ban the eos_token

temperature 1

Custom stopping strings Written between "" and separated by commas.
"\n", "\nYou:"

top_p 1

top_k 0

typical_p 1

min_p 0.05

repetition_penalty 1

frequency_penalty 0

presence_penalty 0

mirostat_mode 0 mode=1 is for llama.cpp only.

mirostat_tau 5

mirostat_eta 0.1

tfs 1

Filter by loader llama.cpp

Load grammar from file (.gbnf)
None

Grammar

Truncate the prompt up to this length The leftmost tokens are removed if the prompt exceeds this length. Most models require this to be at most 2048.
32768

Maximum tokens/second To make text readable in real time.
0

Seed (-1 for random)
-1

Activate text streaming

LLM on GPU

llama.cpp

Steps

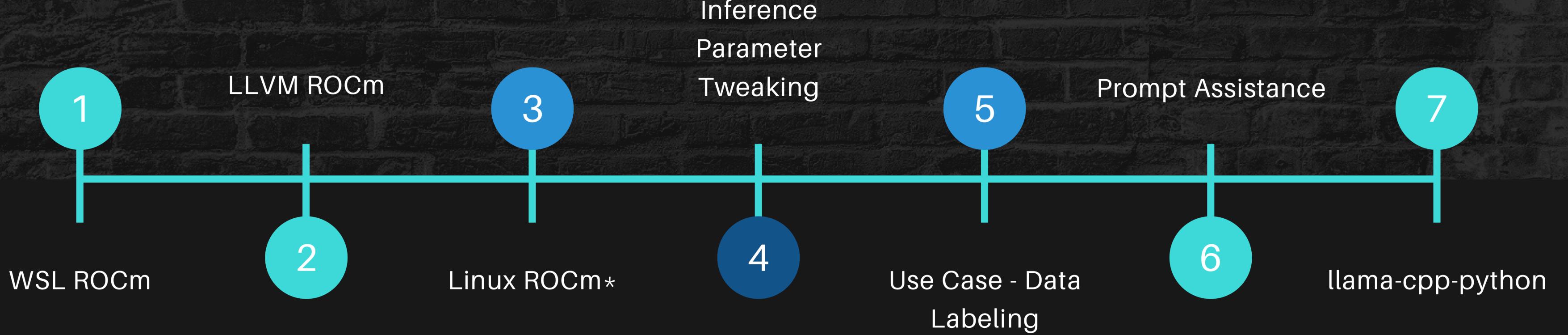
- install cmake
- clone repository
- build
- run in terminal

Benefits

- can run on GPU
- relatively easy install
- llama-cpp-python package available?

```
llm_load_tensors: ggml ctx size = 0.30 MiB
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: OpenCL buffer size = 4095.05 MiB
```

Failed Technical Approaches



*Pending hardware

Model Comparison

Capability	Prompt	Mistral 7B	TinyLlama 1B	Comment
Recall	Who was the president in 2003 and why?	✓	✓	TinyLlama - had to turn the temperature way down
Communicate Complex Topics	Explain the difference between quantum physics and Newtonian physics.	✓	✗	
Recall and Reasoning	Briefly explain the role of carbon in biological life, and suggest another element that could serve a similar role.	✓	✗	
Complex Language	Summarize the main arguments presented in the last State of the Union address and evaluate their potential impact on public policy.	✓	✗	
Sarcasm	Analyze the sentiment of the following statement: 'I just love waiting in traffic, it is the highlight of my day.'	✓	✗	Mistral - the format of the instructions mattered
Creativity	Write a short story about a detective solving a mystery in a futuristic city using only sustainable technologies.	✓	✓	

Inference Parameter Tuning

Temperature using TinyLlama

```
print(llm(total_prompt, max_new_tokens=450, repetition_penalty=1.1, temperature=0.8))
```

- "Bill Clinton was the 42nd President of the United States from March 19, 1993 to January 20, 2001. He succeeded George W. Bush as the 43rd President following the resignation of Bush in the aftermath of the September 11, 2001 attacks. In office, Clinton focused on domestic policy issues such as poverty, welfare reform, and reducing the deficit. He also implemented a series of reforms to the health care system, including the Affordable Care Act. Clinton was known for his successful marriage to former First Lady Hillary Rodham Clinton, and for signing the American Recovery and Reinvestment Act of 2009, which provided economic stimulus and job creation during the Great Recession."
- "'Democrats control both houses of congress and have had a few presidents, but none since George W Bush in 2008.'"
- "'The President of the United States in 2003 was George W. Bush. He was re-elected in 2004 and served as President until his term expired in 2009.'"

```
print(llm(total_prompt, max_new_tokens=450, repetition_penalty=1.1, temperature=0.5))
```

- "'During the year 2003, there were two U.S. presidents: George W. Bush for the second term as President of the United States and Barack Obama for his first term as President of the United States.'"
- "'The President of the United States in 2003 was George W. Bush. He was re-elected in 2004 and served until January 2009, after which his term was ratified by a vote of the people. The presidency is the highest office in the federal government of the United States.'"

```
print(llm(total_prompt, max_new_tokens=450, repetition_penalty=1.1, temperature=0.1))
```

- "'The president of the United States in 2003 was George W. Bush. He was re-elected in 2004 and served until 2009, during which time he signed several major legislative acts into law, including the USA PATRIOT Act and the American Recovery and Reinvestment Act of 2009.'"

Tips

Tips from several hours of pain

Useful GPT Copilot prompts:
Here are the steps to ___ ""
please explain each one

Try things twice

If you try to understand
EVERYTHING you will go
nowhere

Using "instruct" templates don't
always improve generation

For smaller models, input text
formatting matters a lot

I should have bought a nVidia
GPU

Technical guides /
documentation are built for Linux

GPT doesn't know very new
technologies well - slower
development