Best Practices for using local Large Language Models with the UM High-Performance Computing cluster

Instructors: Sean R. Meyer, MBA, PhD and Mark Nuppnau

Transforming Your Research with Generative AI tutorial series

January 23, 2025

MICHIGAN INSTITUTE FOR DATA & AI IN SOCIETY
UNIVERSITY OF MICHIGAN

MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN

1. Llama Setup Demo
2. Advanced Research Computing
3. MM GenAI Policy Overview
4. Llama Usage Demo
5. Use Cases
6. GenAI in Business Operations

Llama Setup

Prerequisites

- Access to an ARMIS2 Slurm account (for billing purposes)
  - UM Research Computing Package (UMRCP)
  - Precision Health membership
- Some experience with Python and Shell (command line interface)

Objectives

- Provide an overview of how to prepare a Python environment for LLM instantiation
- Interact with the LLM via command line interface
- Start an API server that can be queried in Python similar to UMGPT/OpenAI
- Chain together conversations between multiple LLMs with LangChain

Documentation

- https://github.com/sjodingLabs/armis2-llama-demo

Advanced Research Computing

- ARMIS2 High Performance Computing Cluster
- Turbo Research Storage
  - Can be used for sensitive data with approval
- Data Den
  - “cold storage” backup for archived projects
- Open OnDemand
  - Web-based interface for interactive with applications like JupyterLab, Visual Studio Code, Rstudio, etc.

Michigan Medicine Policies and Guidelines

Michigan Medicine Appropriate Clinical Use of Generative Artificial Intelligence Tools Policy
https://michmed-administration.policystat.com/policy/15777590/latest

Guideline on Generative Artificial Intelligence Tools
https://michmed-administration.policystat.com/policy/14811127/latest

Environment Setup

- Start ARMIS2 JupyterLab session with Open OnDemand
- Install python packages (requires GPU access)
- Clone llama.cpp GitHub repository
- Build llama.cpp (requires GPU access)

Llama Usage

Local vs. Web-based LLMs

Local

Local GAI tools (non-internet-accessing) are open-source tools that do not require an internet connection. Local tools are those that can be downloaded and run entirely locally in a computing environment without accessing the internet (an example is Meta's Llama 2 language model).

Web-based

Web-based GAI tools (internet-accessing) are public tools that access or require an internet connection (some examples are: ChatGPT for text accessed through the public webpage, Midjourney for images) and enterprise tools that access or require an internet connection (an example is Epic-integrated GPT-4).

Source: Guideline on Generative Artificial Intelligence Tools

Use Cases

- De-identification
- Synthetic data
- Note Summarization
- Programming/Coding
- Converting Unstructured to Structured Data
- Helper Chat Bots

Useful Definitions

- **Llama.cpp framework**
  - Enable LLM inference with minimal setup and state-of-the-art performance on a wide range of hardware.
- **Quantization**
  - Neural networks that have been compressed by converting their weights and activations from high-precision floating-point numbers to lower-precision numbers
- **Parameters**
  - Larger models have more parameters, which allows them to handle more complex language relationships
- **Retrieval Augmented Generation (RAG)**
  - Provides more accurate responses by cross-referencing knowledge sources

Other Models

- Qwen 2.5 Coder 32B
- DeepSeek-R1

GenAI Research Translation

- UMGPT
  - Exploratory Research
  - Requires exception to policy for limited PHI usage
- ARMIS2/Turbo
  - Scaled Research
  - Can be used with PHI
- Workflow Integration (Future State)
  - Cloud Provider

Download and Test Llama-3.2-3B

Llama-3.2-3B

- 3 billion parameters
- downloading the model requires ~3GB of storage space
  - GPU VRAM required ~num_params * 0.8 (i.e., 3B*0.8 = 2.4GB)
  - The number of digits stored for each model weight will determine the file size of the model.
  - Space and GPU memory requirements can be reduced by storing a lower number of digits for each model weight, but there is a tradeoff with the quality of the output.

Filename	Quant type	File Size	Description
Llama-3.2-3B-Instruct-Q8_0.gguf	Q8_0	3.42GB	Extremely high quality, generally unneeded but max available quant.
Llama-3.2-3B-Instruct-Q6_K.gguf	Q6_K	2.64GB	Very high quality, near perfect, recommended.
Llama-3.2-3B-Instruct-Q5_K_M.gguf	Q5_K_M	2.32GB	High quality, recommended.
Llama-3.2-3B-Instruct-Q4_K_M.gguf	Q4_K_M	2.02GB	Good quality, default size for must use cases, recommended.
Llama-3.2-3B-Instruct-Q4_K_S.gguf	Q4_K_S	1.93GB	Slightly lower quality with more space savings, recommended.
Llama-3.2-3B-Instruct-IQ4_XS.gguf	IQ4_XS	1.83GB	Decent quality, smaller than Q4_K_S with similar performance, recommended.

Demo: OpenAI API

OpenAI API

-

Demo: LangChain

LangChain

Usage

- allows for simplified interactions with LLMs
- chains together various elements of an application

Types of chains

- LLM Chain
  - user input is passed into a PromptTemplate to transform the input into a coherent prompt
  - the prompt is passed into an LLM to generate an LLM output
  - output is passed to an OutputParser to format the results
- Sequential Chain
  - combines individual chains, creating a continuous sequence of chains

References

Llama.cpp
https://github.com/ggerganov/llama.cpp?tab=readme-ov-file#llama-cli

https://github.com/ggerganov/llama.cpp/blob/master/docs/build.md#cuda

Create VS Code setup file (ARMIS2)
https://documentation.its.umich.edu/arc-hpc/open-ondemand/vs-code

ARMIS2 Overview and Access Guide for Researchers
ARC Tools - RCP.MMRCI & Armis 2 details.docx

Building effective agents (article)
https://www.anthropic.com/research/building-effective-agents

University of Michigan Sensitive Data Guide
https://safecomputing.umich.edu/dataguide/data/19

Extending Llama.cpp using llama-cpp-python library
https://github.com/abetlen/llama-cpp-python

Scan the QR code or go to this link:

https://myumi.ch/23dxR

to provide feedback on the session

Thank you!

For past sessions, see videos at:
midas.umich.edu/generative-ai-tutorial-series/

MICHIGAN INSTITUTE FOR DATA & AI IN SOCIETY
UNIVERSITY OF MICHIGAN

MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN