# LLM Focus Group - Spring 2024

ML@B Research Committee



## General Overview

In this focus group, we’ll walk through the training, fine-tuning, and deployment of an LLM from scratch. We’ll start from the basics and then learn more about new techniques, tips, and tricks. We’ll also explore a lot of related topics, such as multi-node processing, quantization, and deployment.

## Deliverables

* A working LLM
* (Multiple) blog posts
* Deploying our LLM on an inference server

## Potential Topics/Schedule

### **Module 1: Implementing a Transformer**

* Understand the Transformer architecture
* Understand and implement self-attention
* Implement the GPT architecture

### **Module 2: Data**

* Understand tokenization and tokenizers
* Learn about dataset curation for LLM training
* Read about data scaling laws
* Experiment with different dataset makeups (code, Wikipedia, general internet percentages, etc)

### **Module 3: Model Training**

* Write training code
* Implement Dataloaders, Dataset
* Learn more about multi-node processing / DDP

### Module 4: Tricks and Tips

* Review most common / newest transformer tricks and techniques
* Implement tweaks

### Module 5: Fine Tuning

* Fine-tuning process, fine-tuning for specific tasks
* RLHF?

### Module 6: Deployment

* Explore quantization techniques?
* Deploy on server for inference

## Structure

* Meetings every week
* Each week has “discussion leaders” (aka the person or group in charge that week). Discussion leaders are in charge of a mini-presentation. This often includes giving a code walkthrough, which would require writing the code / code outline for that week.
* A different version of the code with some parts blocked out can be given to non-presenters that week.
* Normal members should read the assigned readings / research the assigned topics. They will complete the given notebook/code during the walkthrough.

## What’s in it for me?

* Lots of learning
* Lots of experience with LLMs
* Gain experience doing a lot of important things / learning crucial skills
  + If you ever do DL research you will encounter having to learn most of this (esp. Multi-node processing, writing good data loaders, etc).
* Be able to talk about how you coded an LLM from scratch on your resume or during interviews

## Signup

| **Interested? Fill out this When2Meet and list your name below:** <https://www.when2meet.com/?23562308-unvr3> |
| --- |

**IF YOU’RE INTERESTED, LIST YOUR NAME BELOW AFTER FILLING OUT THE W2M**

* Chuyi Shang
* Charlie cheng
* Samarth Jajoo
* Seyone
* Naz Col
* Michael Lutz
* luca manolache
* Ria doshi
* Eren
* Maggie Dong
* Tim
* Mehdi
* Alishba
* Nyx
* Arvind
* Parth
* Andrew
* Terry Kim (chuyi = 🐐)
* Evan
* Skyla Ma
* Ria Jain
* Stephen
* Timothy Gao
* Ayushi
* Anisha
* Sara Eginova
* Harbani Jaggi
* Eric Wang
* Aryan Goyal
* Rohan
* Narisam
* Erfan
* Tejas
* Austen Liao
* Riteka