index

October 26, 2023

— title: "Fractal-U LLM Course" author: "Chris Kroenke" date: "9/25/2023" toc: true badges: true categories: [fractal, python, LLM] image: fractal-llm-mascot.png jupyter: python3 —

1 Fractal-U LLM Course

Overview and motivations for the Fractal-U LLM Course.

2 Intro

In this course we will fine-tune a Large Language Model (LLM) to create a simple AI Agent that is specialized for a certain task. What exactly do we mean by Agent, and why would we focus on a simple one?

There is a lot of talk and hype around Agents at the moment. Folks are thinking about a future where we all have personalized AI assistants, aka Agents, at our fingertips. These Agents promise to make our lives easier and more comfortable. They are the AI assistants from Science Fiction made material: TARS from Interstellar, Jarvis from Iron Man, HAL 9000 from Space Odyssey, Samantha from Her, etc.

2.1 Realisitic Agent Expectations

However, despite the breakneck speed of progress in AI and LLMs, Agents as powerful as those are still a ways off. It is hard if not impossible to predict the exact timelines. For now it is safe to say that Agents of that caliber won't be here anytime "soon". But, barring some force majeure, they *will* exist at some point.

The gap, then, is that folks are promising these advanced Agents while on the ground we are still dealing with sycophantic LLMs, their hallucinations, and prohibitive compute requirements.

2.2 Fine-Tuning Reasonable Agents

Where does that leave us? Well, as a recent announcement from OpenAI shows, fine-tuning GPT-3.5 on small, clean datasets surpasses GPT-4 on certain tasks.

That's what we are aiming for. In other words, we already have the ability to develop outrageously powerful tools by fine-tuning LLMs on small, clean datasets.

So we won't build Iron Man's Jarvis, but we will aim for far more than a simple chatbot. If it helps, think about our simple Agent like an Intelligent Rubber Duck. A Rubber Duck is anything (actual yellow rubber duckie optional) that you keep around your desk and talk to about your work. It is

a physical tool for thought, since it's so often helpful to speak out loud the swirl of thoughts in our head.

Our simple Agent will be a Rubber Duck that can talk with us. When we ask it a question about our work, it will respond given what it knows about the project as a whole. Or, if we are simply verbalizing a thought to untangle it, the Agent will give us feedback or suggest other approaches.

If we can be so bold: our Agent will be a mini-Jarvis laser-focused on a specific topic. Then, as both the tools and tech progresses, we will have a full development stack to unlock even more capabilities from our Intelligent Rubber Duck.

> Course Summary: We will fine-tune an LLM on a small, clean dataset of our choosing to build an Intelligent Rubber Duck to better help us create.

2.3 Course Overview

This course is meant to be public-facing. Students should write about their journey to both chronicle and track their progress. Publishing a piece of work lets us see it under a new critical light, and that is a powerful tool for learning. Sharing each other's work will also crystalize your understanding of a topic by seeing it from different angles. Blogging and writing can be scary, but here we aim to make it as smooth and helpful as possible.

Each lesson is inside a Jupyter Notebook. The notebooks build on each other, but they are also self-contained and reviewable on their own.

Here is an outline of the notebook lessons:

1. Python environments for LLMs
> Using mamba and pip to create isolated and reproducible python environments.

2. Blogging with nbdev
> Setting up a blog with nbdev and github pages.

3. Running HuggingFace models
> Using the HuggingFace API to run ML models, and get familiar with weight files.

4. Running LLMs with llama.cpp
> Running quantized, optimized LLMs optimized for deployment.

5. Processing text documents for LLMs
> Preparing text data for fine-tuning and Retrieval Augmented Generation (RAG) tasks

6. Fine-tuning LLMs on one or few GPUs
> Scripts and setups to efficiently fine-tune LLMs

7. Running fine-tuned LLMs locally on your phone
> Deploying fine-tuned, quantized LLMs on mobile devices

Thank you for coming on this journey! I promise that we will learn a ton and have lots of fun along the way.