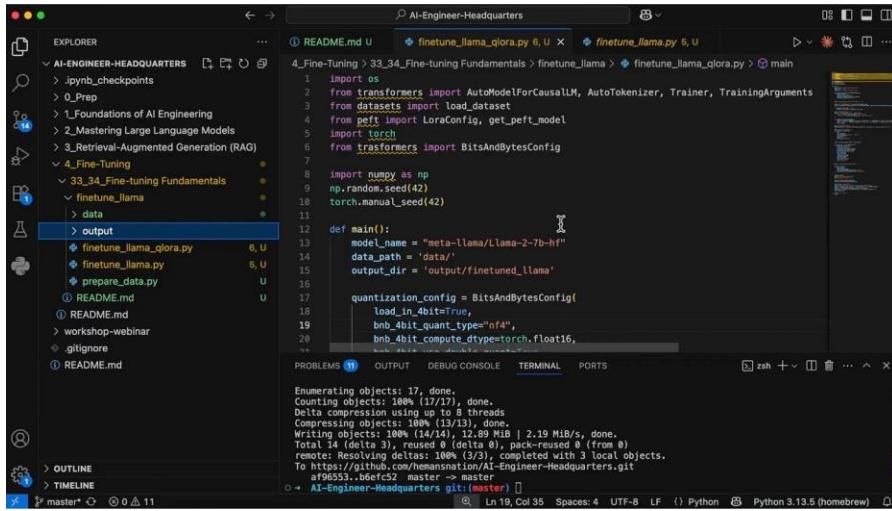


## Synthetic Data Generation & Fine-Tuning Specialist

**Author:** Santhi Maddipudi



The screenshot shows a code editor interface with the title bar "AI-Engineer-Headquarters". The left sidebar displays a file tree for a project named "AI-ENGINEER-HEADQUARTERS". The main editor area contains a Python script named "finetune\_llama\_qlora.py". The script imports os, transformers, datasets, peft, torch, and numpy. It sets up a main function to load a dataset, initialize a LLaMA model with a BitsAndBytesConfig, and perform 4-bit quantization. The terminal below shows the command "git pull" being run, indicating the latest code changes.

```
import os
from transformers import AutoModelForCausalLM, AutoTokenizer, Trainer, TrainingArguments
from datasets import load_dataset
from peft import LoraConfig, get_peft_model
import torch
from transformers import BitsAndBytesConfig

import numpy as np
np.random.seed(42)
torch.manual_seed(42)

def main():
    model_name = "meta-llama/Llama-2-7b-hf"
    data_path = 'data'
    output_dir = 'output/finetuned_llama'

    quantization_config = BitsAndBytesConfig(
        load_in_4bit=True,
        bnb_4bit_quant_types="nf4",
        bnb_4bit_compute_dtype=torch.float16,
        ...

Enumerating objects: 17, done.
Counting objects: 100% (17/17), done.
Delta compression using up to 5 threads.
Compressing objects: 100% (13/13), done.
Writing objects: 100% (14/14), 12.89 MiB | 2.19 MiB/s, done.
Total 14 (delta 3), reused 0 (delta 0), pack-reused 0 (from 0).
remote: Resolving deltas: 100% (3/3), completed with 3 local objects.
To https://github.com/llm-finetuning/AI-Engineer-Headquarters.git
  a96553...06defc52 master -> master

```

### Problem we are solving

GPT-4 costs \$30 per million tokens.

For high-volume tasks like generating SQL queries, writing product descriptions, or answering support tickets, these costs destroy budgets.

Meanwhile, GPT-4 is overkill for most tasks.

A small 8B model fine-tuned for one specific job can match GPT-4 quality at 1% of the cost.

### Tech stack

Unsloth, Hugging Face TRL, WandB, Ragas, GPT-4 API, Python.

### Proof of work :

In progress