**17 May 2025**

Started the day by searching for a suitable dataset, but couldn’t find one that met all the criteria. Decided to move forward with web scraping instead. Initially attempted to scrape 100 questions and answers from five Stack Overflow pages tagged with pytorch. However, I soon realized that some questions had no answers, so I updated the script to skip those and store only the highest-voted answer per question.

Later, I revised the goal to extract 1,000 top-rated questions and answers based on vote count. Eventually, I finalized a script that scrapes top-voted Q&A pairs across various machine learning and deep learning frameworks.

Upon running the script, I encountered a limitation: Stack Overflow's API enforces strict rate limits on user calls. As a workaround, I plan to run the script incrementally over time. In the meantime, I moved on to the next task in the pipeline.

Created a preprocessing script that concatenates all the JSON files and formats them into the structure:

{

"instruction": "How do I fix 'command not found' in Ubuntu?",

"input": "",

"output": "Ensure the package is installed. Use `sudo apt install <package>`."

}

This part went smoothly and worked well on the data I had collected so far.

Next, I began working on the fine-tuning script. Attempted to use Unsloth with LLaMA 3–8B in 4-bit precision, but it exceeded my laptop’s GPU capacity. That’s enough progress for today. Tomorrow, I’ll continue scraping and try using Hugging Face’s transformers and peft libraries for fine-tuning instead.