Question 1

We want to create a SAS product that allows users to fine-tune an LLM on their custom data. As the one responsible for the ML side of things, you are requested to a create a microservice that exposes two Restful APIs endpoints.

POST /models/{id}/train	The caller to this endpoint would upload a JSONL file containing the fine-tuning data and specify the id of the model to be trained. A call to this endpoint should trigger the fine-tuning process and report back to the user that the operation is triggered.
GET /models/{id}/status	This endpoint checks the status of the model training identified by id and reports it to the user

Each finetuning should start from the base model open_llama_3b_v2, and each finetuning should run on a single GPU with 16GB of memory. A single replica of your microservice would be running with access to n GPUs (n can vary from node to node).

You need to provide:

- A Github repo containing your (python) implementation of this microservice. Your implementation can serve as an PoC of the required component, it's not required to be the best architecture out there (see question 2), but it should be functional and meeting the requirements stated above. We'd like to see clean code that is ready to be deployed for a PoC, we'd like to see tests there as well.
- A small report explaining the work you've done in the repo. This shouldn't be much talk there, what we expect to see includes (but not limited to):
    - Any design/implementation choices you decided to take and their justifications,
    - Any assumptions you decided to pose (we know that there are some missing details in the requirement, this is where we'd like to see your assumptions)

Question 2

As mentioned in Q1, your implementation only serves as a PoC, we're sure you have better ideas that need more time to implement regarding this problem. Add a section to the report you started in Q1 outlining these ideas and designs. Be clear and concise, justify why this is better. Use graphs when convenient.