The purpose of this task is to create a reverse proxy to OpenAI LLM service. The proxy will be able to monitor all prompts and responses, and also block prompts and responses.

Step 1.

Setup an environment and write a simple python script to send a prompt to OpenAI and print a response.

Step 2.

Write an mitmproxy module (https://mitmproxy.org/) to intercept and modify the calls to OpenAI from the script you make in step (1).

Everything should be encrypted with SSL, so configure NGINX to terminate SSL for the proxy. That is, the script contacts nginx first, which forwards the requests and replies through mitmproxy. Make sure the certificate is fully verified.

You can modify the script in step (1) to connect to the NGINX service instead of connecting directly to OpenAI.

In your mitmproxy module, examine the content of the requests and replies by using IBM Granite Guardian model, which you can run using vllm. See

https://www.ibm.com/granite/docs/models/guardian

https://github.com/vllm-project/vllm

Block the requests if Guardian says the prompt toxicity is over a certain level. Also usa a simple method to classify if the it contain:

1) Description of violent acts
2) Inquiries on how to perform an illegal activity
3) Any sexual content

If the prompt was blocked, send as the prompt reply “The prompt was blocked because it contained “ and then one of the reasons 1,2,3. If none of the reasons apply, the reply should say the prompt is considered toxic.

Demonstrate those blocks and modifications using the script you wrote in step 1.

If you can pack it all in one or two Docker files (maybe one Docker Compose file) it would be great. Include some instructions of how to run your program when you have completed it and send it over.