**Setting up the Gen AI Model**

* pip install vllm
* from vllm import LLM, SamplingParams
* llm = LLM(model="meta-llama/Llama-2-7b-chat-hf") # Load your model on HPC GPU
* sampling\_params = SamplingParams(temperature=0.7, max\_tokens=500)
* prompt = "Explain machine learning to a high school student."
* outputs = llm.generate([prompt], sampling\_params)
* print(outputs[0].outputs[0].text) # Print AI-generated response

**Resource:**

Available models: <https://docs.vllm.ai/en/latest/models/supported_models.html>

Instillation: https://docs.vllm.ai/en/latest/getting\_started/installation/index.html