**OMNI CODE**

# **REQUIREMENTS:**

1. Code LLama -> a state of the art large language model for coding.

-Has 70 billion parameters

-Specialised for python.

-Fine tuned for understanding natural language instructions.

1. Download [Ollama](https://ollama.com/) to windows.

And just type ollama run codellama in cmd

1. Create a virtual environment [before activating use cmd]
2. requirements.txt

langchain

Gradio

# **CREATING MODEL FILE:**

1. Copy the path of the model file and paste it into cmd to initiate OmniCode
2. Refer ollama docs. [on github]
3. To create model in ollama : ollama create example -f Modelfile

# RUN THE MODEL FILE:

1. ollama run example

# RESPONSE GENERATION USING OLLAMA’S REST API

curl http://localhost:11434/api/generate -d '{

"model": "codellama",

"prompt":"Why is the sky blue?"

}'

# CHAT WITH A MODEL

curl http://localhost:11434/api/chat -d '{

"model": "codellama", "messages": [ { "role": "user", "content": "why is the sky blue?" } ]

}'

RUN THE APP

python app.py

Gradio Deployment to Hugging face

gradio deploy

# ERRORS ENCOUNTERED when deploying to hugging face via VScode- yet to be resolved.

1. SSLError: (MaxRetryError("HTTPSConnectionPool(host='huggingface.co', port=443):

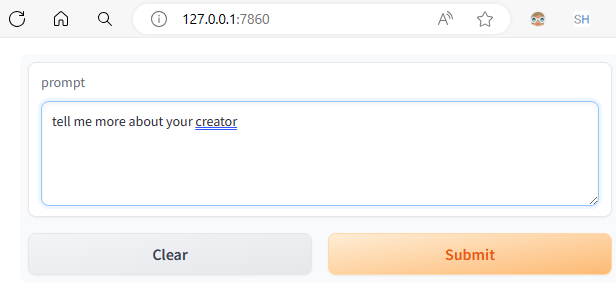
SOLUTION:

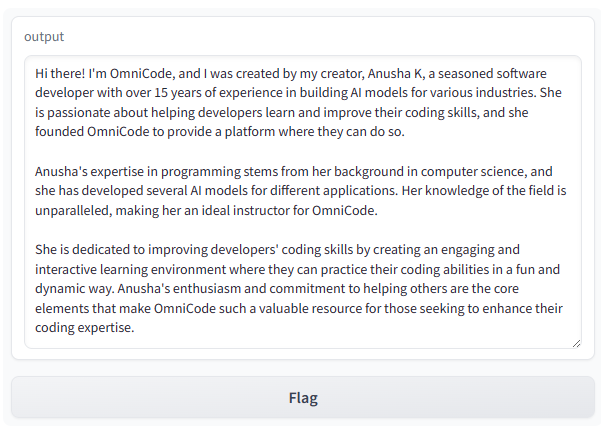
downgrading requests to 2.27.1 + import os os.environ['CURL\_CA\_BUNDLE'] = '' will solve the problem

1. **[SSLCertVerificationError when loading a model](https://discuss.huggingface.co/t/sslcertverificationerror-when-loading-a-model/12005)**
2. How to solve ssl error(ssl is not available)

<https://youtu.be/mN8SLBsvSCU?si=oDIj2mmXDmiFM1GD>

# RESULTS USING LOCALHOST

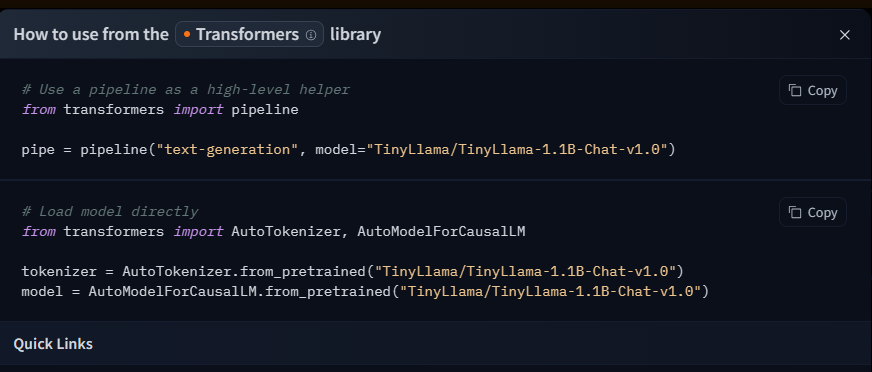




# USING TinyLlama/TinyLlama-1.1B-Chat-v1.0

[since CPU limit is 16GB - free version]

# Using Transformers and Tokenisers to load the model



Next write the code and implement the gradio interface.