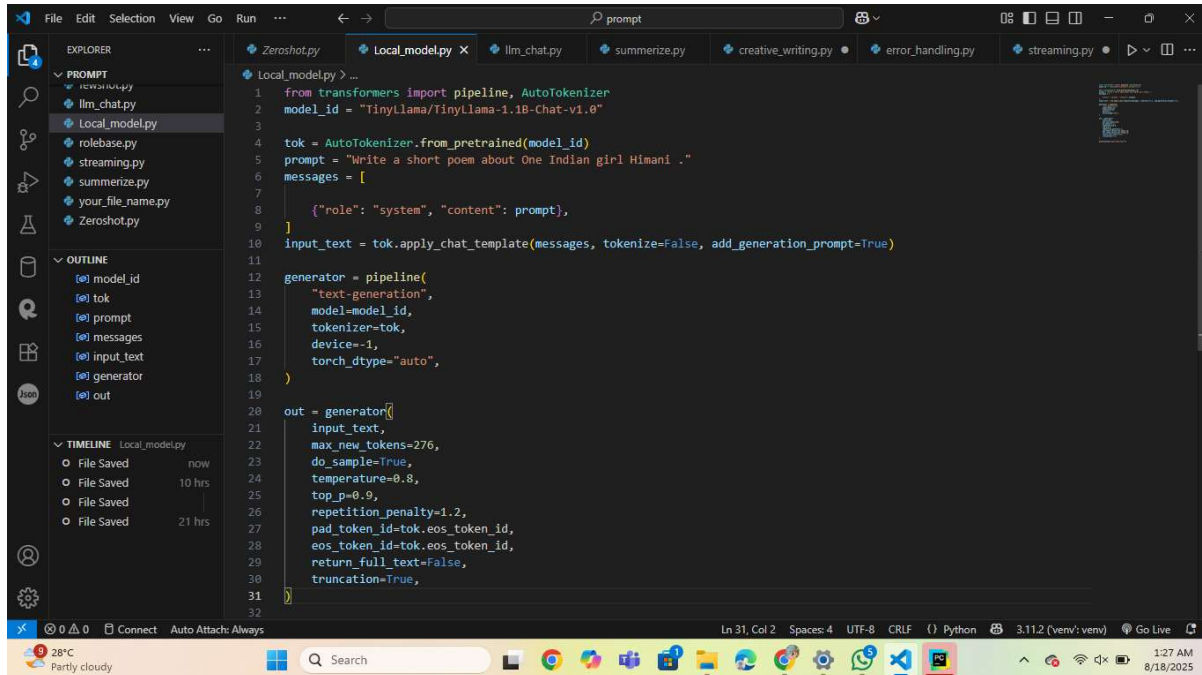


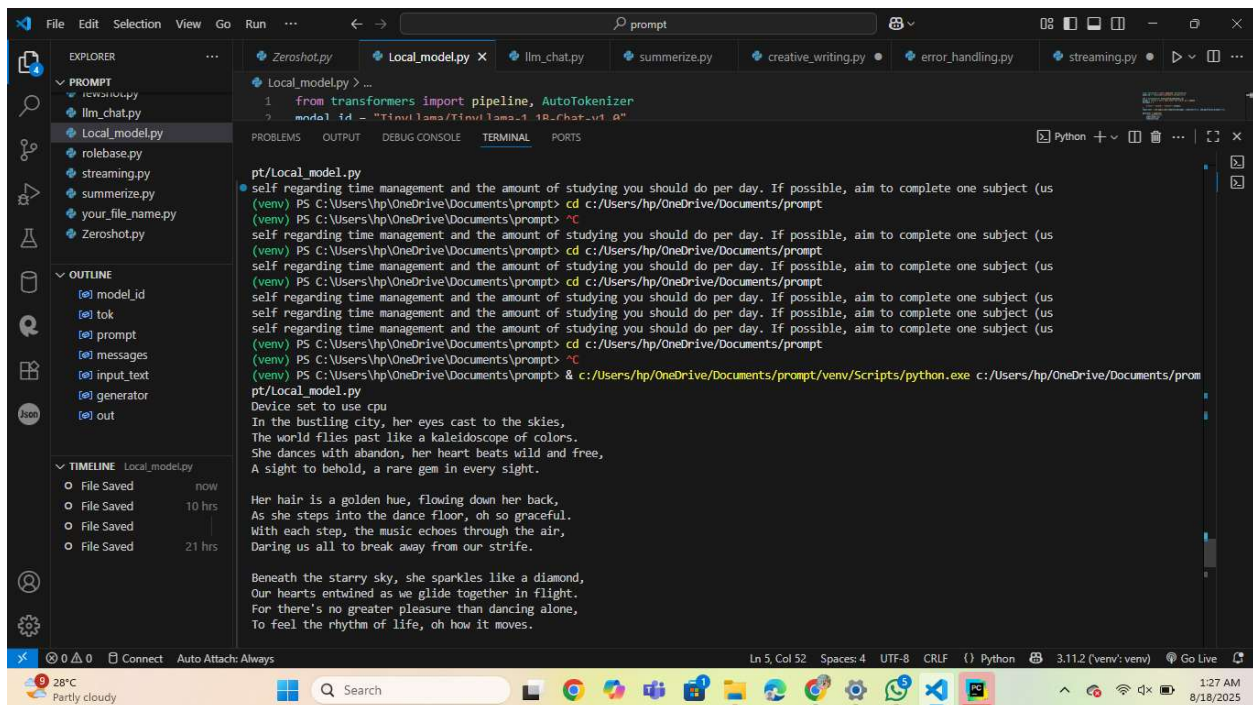
# Hands-on: Install Ollama and run local LLM (Llama 3.2), Performance optimization and troubleshooting local model

Code snippet:



```
1 from transformers import pipeline, AutoTokenizer
2 model_id = "TinyLlama/TinyLlama-1.18-Chat-v1.0"
3
4 tok = AutoTokenizer.from_pretrained(model_id)
5 prompt = "Write a short poem about One Indian girl Himani ."
6 messages = [
7     {
8         "role": "system", "content": prompt,
9     }
10 ]
11 input_text = tok.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)
12
13 generator = pipeline(
14     "text-generation",
15     model=model_id,
16     tokenizer=tok,
17     device=-1,
18     torch_dtype="auto",
19 )
20
21 out = generator(
22     input_text,
23     max_new_tokens=276,
24     do_sample=True,
25     temperature=0.8,
26     top_p=0.9,
27     repetition_penalty=1.2,
28     pad_token_id=tok.eos_token_id,
29     eos_token_id=tok.eos_token_id,
30     return_full_text=False,
31     truncation=True,
32 )
```

Output:



```
pt/Local_model.py
Device set to use cpu
In the bustling city, her eyes cast to the skies,
The world flies past like a kaleidoscope of colors.
She dances with abandon, her heart beats wild and free,
A sight to behold, a rare gem in every sight.

Her hair is a golden hue, flowing down her back,
As she steps into the dance floor, oh so graceful.
With each step, the music echoes through the air,
Daring us all to break away from our strife.

Beneath the starry sky, she sparkles like a diamond,
Our hearts entwined as we glide together in flight.
For there's no greater pleasure than dancing alone,
To feel the rhythm of life, oh how it moves.
```

