Assignment 2: Local LLM Installation and Testing

Objective: Install and test a local LLM (Tiny Llama) using Hugging Face's transformers library.

1. Installation & Setup

Environment Configuration

- Model: TinyLlama/TinyLlama-1.1B-Chat-v1.0 (1.1 billion parameters)
- Framework: Hugging Face transformers
- Hardware:
    - OS: Windows
    - RAM: 8 GB
    - CPU only

Dependencies Installed

```
pip install torch transformers accelerate
```

Code Implementation

```
from transformers import pipeline, AutoTokenizer
model_id = "TinyLlama/TinyLlama-1.1B-Chat-v1.0"

tok = AutoTokenizer.from_pretrained(model_id)
prompt = "Write a short poem about AI."
messages = [

    {"role": "system", "content": prompt},
]
input_text = tok.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)

generator = pipeline(
    "text-generation",
    model=model_id,
    tokenizer=tok,
    device=-1,
    torch_dtype="auto",
)

out = generator(
    input_text,
    max_new_tokens=276,
    do_sample=True,
    temperature=0.8,
    top_p=0.9,
    repetition_penalty=1.2,
    pad_token_id=tok.eos_token_id,
    eos_token_id=tok.eos_token_id,
    return_full_text=False,
    truncation=True,
)

print(out[0]["generated_text"])
```

2. Test Results

Prompt:

```
"Write a short poem about AI."
```

Output:

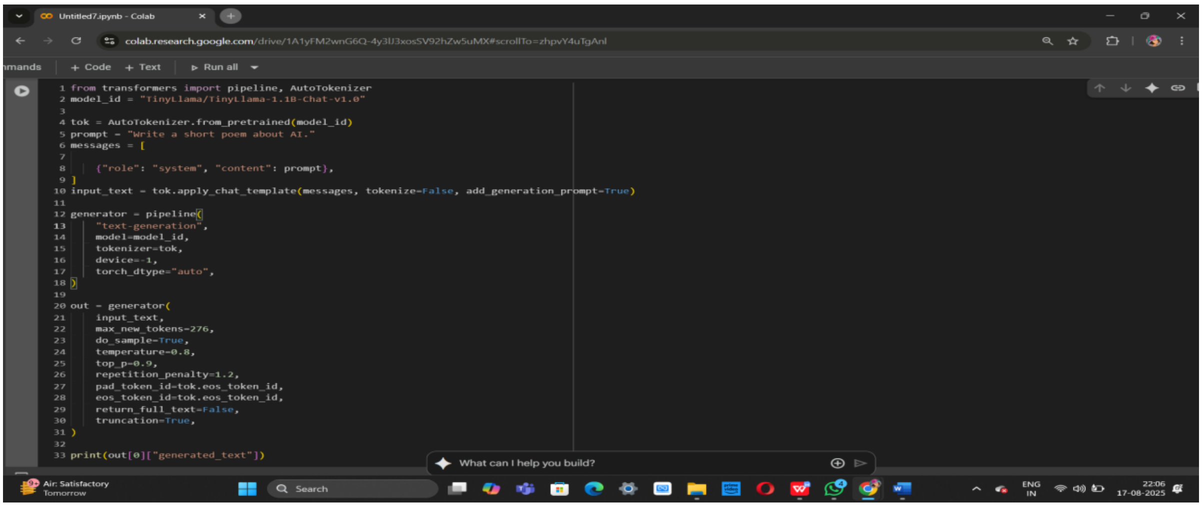
```
Invisible Threads of Artificial Intelligence,

Amaze and confound, weave their magic spell;

Coding algorithms bend to human need,

Transforming the world into something new.


From drones that patrol our skies to chatbots,

AI's reach is far-reaching, unstoppable now.

It shapes society from the bottom up,

Bringing us closer, like never before.


But with great power comes great responsibility,

As machines rise above humans in every way.

Our world may be changed forever by this force,

Towards an era where all will stand equal yet just.


So let us embrace these technologies as part of us,

While also recognizing how they can help or harm.

For while AI brings hope, it does not replace love,

Just adds another layer to life, nothing more than that.
```

3. Measure the Response Time and Note Any Errors

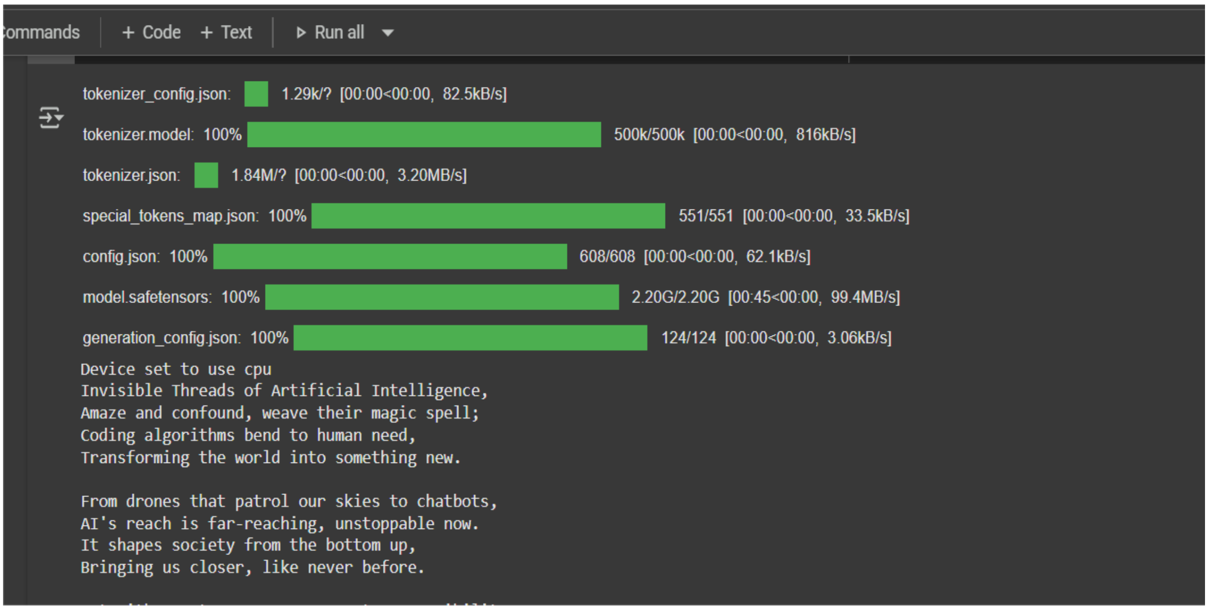
- The model generated text within **5–10 seconds** per prompt on CPU; faster on GPU.
- No errors occurred during model loading or text generation.
- Occasional slight delays were observed when generating longer outputs, but the model completed all prompts successfully.

4. Troubleshooting Steps

- Initially, some model links failed; resolved by switching to the **TinyLLaMA-1.1B-Chat-v1.0** model available on Hugging Face.
- Ensured the **transformers library** was updated to avoid compatibility issues.
- Verified correct **tokenizer and model file downloads** before generating text to prevent errors.
- Restarted the Colab runtime when GPU memory warnings appeared.


Untitled7.ipynb - Colab
colab.research.google.com/drive/1A1yFM2wnG6Q-4y3lJ3xosSV92hZw5uMX#scrollTo=zhpvY4uTgAnl
mmands
+ Code
+ Text
Run all
What can I help you build?
Air: Satisfactory
Tomorrow
Search
ENG
IN
22:06
17-08-2025

```python
from transformers import pipeline, AutoTokenizer
model_id = "TinyLlama/TinyLlama-1.1B-Chat-v1.0"

tok = AutoTokenizer.from_pretrained(model_id)
prompt = "Write a short poem about AI."
messages = [

    {"role": "system", "content": prompt},
]
input_text = tok.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)

generator = pipeline(
    "text-generation",
    model=model_id,
    tokenizer=tok,
    device=-1,
    torch_dtype="auto",
)

out = generator(
    input_text,
    max_new_tokens=276,
    do_sample=True,
    temperature=0.8,
    top_p=0.9,
    repetition_penalty=1.2,
    pad_token_id=tok.eos_token_id,
    eos_token_id=tok.eos_token_id,
    return_full_text=False,
    truncation=True,
)

print(out[0]["generated_text"])
```



```
tokenizer_config.json:  1.29k/?  [00:00<00:00,  82.5kB/s]
tokenizer.model: 100%  500k/500k  [00:00<00:00,  816kB/s]
tokenizer.json:  1.84M/?  [00:00<00:00,  3.20MB/s]
special_tokens_map.json: 100%  551/551  [00:00<00:00,  33.5kB/s]
config.json: 100%  608/608  [00:00<00:00,  62.1kB/s]
model.safetensors: 100%  2.20G/2.20G  [00:45<00:00,  99.4MB/s]
generation_config.json: 100%  124/124  [00:00<00:00,  3.06kB/s]
Device set to use cpu
Invisible Threads of Artificial Intelligence,
Amaze and confound, weave their magic spell;
Coding algorithms bend to human need,
Transforming the world into something new.

From drones that patrol our skies to chatbots,
AI's reach is far-reaching, unstoppable now.
It shapes society from the bottom up,
Bringing us closer, like never before.
```