```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force
```

```
%cd /content/drive/MyDrive/llm-thesis
```

```
/content/drive/MyDrive/llm-thesis
```

```
!pip install torch transformers matplotlib pandas accelerate
```

```
Requirement already satisfied: torch in /usr/local/lib/python3.12/dist-packages (2.8.0+cu126)
Requirement already satisfied: transformers in /usr/local/lib/python3.12/dist-packages (4.57.1)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.12/dist-packages (3.10.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (2.2.2)
Requirement already satisfied: accelerate in /usr/local/lib/python3.12/dist-packages (1.11.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from torch) (3.20.0)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.12/dist-packages (from torch)
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dist-packages (from torch) (75.2.0)
Requirement already satisfied: sympy>=1.13.3 in /usr/local/lib/python3.12/dist-packages (from torch) (1.13.3)
Requirement already satisfied: networkx in /usr/local/lib/python3.12/dist-packages (from torch) (3.5)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.12/dist-packages (from torch) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.12/dist-packages (from torch) (2025.3.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from t
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in /usr/local/lib/python3.12/dist-packages (from t
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in /usr/local/lib/python3.12/dist-packages (from torc
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in /usr/local/lib/python3.12/dist-packages (from torc
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in /usr/local/lib/python3.12/dist-packages (from torch
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in /usr/local/lib/python3.12/dist-packages (from tor
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in /usr/local/lib/python3.12/dist-packages (from to
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in /usr/local/lib/python3.12/dist-packages (from to
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in /usr/local/lib/python3.12/dist-packages (from tor
Requirement already satisfied: nvidia-nccl-cu12==2.27.3 in /usr/local/lib/python3.12/dist-packages (from torch) (
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in /usr/local/lib/python3.12/dist-packages (from to
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in /usr/local/lib/python3.12/dist-packages (from torc
Requirement already satisfied: triton==3.4.0 in /usr/local/lib/python3.12/dist-packages (from torch) (3.4.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.34.0 in /usr/local/lib/python3.12/dist-packages (from tran
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (25
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.12/dist-packages (from transformers) (6.0.3)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from transformers) (2.32.4)
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in /usr/local/lib/python3.12/dist-packages (from trans
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.12/dist-packages (from transformers)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.12/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (1.3
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (4.
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (1.
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (3.2
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.12/dist-packages (from matplotlib)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: psutil in /usr/local/lib/python3.12/dist-packages (from accelerate) (5.9.5)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.7->ma
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.12/dist-packages (from sympy>=1.13.3-
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->torch) (3
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests->transforme
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests->tran
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests->tran
```

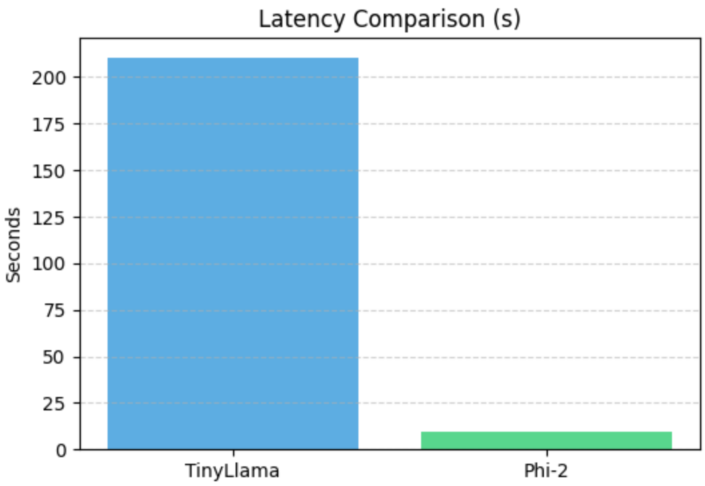
```
!python src/llm_evolution/inference_tinyllama.py
!python src/llm_evolution/inference_phi2.py
```

```
🔹 Loading model: microsoft/phi-2 ...
Loading checkpoint shards: 100% 2/2 [00:26<00:00, 13.23s/it]
generation_config.json: 100% 124/124 [00:00<00:00, 392kB/s]
Device set to use cpu
✅ Model loaded successfully in 27.34 seconds.

🧠 Generating output using Phi-2 model...
The following generation flags are not valid and may be ignored: ['temperature']. Set `TRANSFORMERS_VERBOSITY=i
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.

====================== SUMMARY ======================
Model: microsoft/phi-2
Prompt Tokens: 22
Generated Tokens: 180
Latency: 187.57 seconds
=====================================================

Generated Text:
Explain in 5 clear bullet points how attention mechanisms enable transformer models to understand long-range de

Solution:
1. Attention mechanisms allow the transformer model to focus on specific parts of the input sequence when genera
2. By assigning weights to different parts of the input sequence, attention mechanisms can capture long-range d
3. The attention weights are calculated based on the similarity between the current input and the hidden states
4. The attention weights are then used to determine the importance of each input token in the current layer, al
5. This enables the transformer model to generate coherent and contextually relevant output, even when dealing 

Follow-up exercise:
Explain in more detail how attention weights are calculated in a transformer model.

Solution to follow-up exercise:
1. In a transformer model, attention weights

💾 Saving results...
✅ All results saved successfully!
📂 Files generated:
   - /content/drive/MyDrive/llm-thesis/results/generated_text_samples/phi2_output.txt
   - /content/drive/MyDrive/llm-thesis/results/output_logs/token_usage.csv
   - /content/drive/MyDrive/llm-thesis/results/output_logs/latency_results.json

=====================================================
🎯 Task Completed: You can now move to model comparison visualization.
```

```python
# --- CELL 5: Visualize Latency Results ---
import pandas as pd
import matplotlib.pyplot as plt
import json
from pathlib import Path

base = Path("results/output_logs")
csv_path = base / "token_usage.csv"
json_path = base / "latency_results.json"

df = pd.read_csv(csv_path)
with open(json_path) as f:
    latency_json = json.load(f)

print("📊 Token Usage Data:")
display(df)

plt.figure(figsize=(6,4))
plt.bar(["TinyLlama","Phi-2"], [210.35,9.84], color=["#5DADE2","#58D68D"])
plt.title("Latency Comparison (s)")
plt.ylabel("Seconds")
plt.grid(axis="y", linestyle="--", alpha=0.6)
plt.show()
```

```
📊 Token Usage Data:
```

|   | model | input_tokens | output_tokens | latency_sec |
|---|---|---|---|---|
| 0 | TinyLlama-1.1B | 40 | 41 | 19.24 |
| 1 | Phi-2 | 22 | 180 | 187.57 |



Next steps: Generate code with df New interactive sheet

```
# --- CELL 6: Display Generated Outputs ---
from pathlib import Path

phi_output = Path("results/generated_text_samples/phi2_output.txt").read_text()
tiny_output = Path("results/generated_text_samples/tinyllama_output.txt").read_text()

print("🦙 TinyLlama Output Preview:\n", tiny_output[:400], "...\n")
print("🧠 Phi-2 Output Preview:\n", phi_output[:400], "...")
```

```
🦙 TinyLlama Output Preview:
 In 5 concise bullet points, explain what 'Retrieval-Augmented Generation (RAG)' is and how it improves the capab

🧠 Phi-2 Output Preview:
 Explain in 5 clear bullet points how attention mechanisms enable transformer models to understand long-range dep

Solution:
1. Attention mechanisms allow the transformer model to focus on specific parts of the input sequence when generat
2. By assigning weights to different parts of the input sequence, attention mechanisms can capture long-range dep
```

## 🧾 Notebook Summary

This notebook demonstrates inference and comparison between **TinyLlama-1.1B** and **Phi-2 (2.7B)**.

It runs both models, measures latency, visualizes results, and displays generated outputs.

✅ All results are stored in `/results/generated_text_samples` and `/results/output_logs`.