Lesson 3 Practical Exercise: Experimenting With Three Transformer Variants

Your task

In this lab, you will load one small checkpoint from each Transformer family using the Hugging Face library and (a) run task-appropriate prompts, (b) measure latency, and (c) compare output style. By the end, you should be able to articulate why encoder-only models suit classification, decoder-only models suit free generation, and encoder–decoder models suit sequence-to-sequencetasks such as summarization.

You will load one small checkpoint from each family using Hugging Face Pipelines:

```
from transformers import pipeline

# encoder-only

bert_sent = pipeline("sentiment-analysis",

model="distilbert-base-uncased-finetuned-sst-2-english")

print(bert_sent("A delightful film."))

 # decoder-only

gpt_gen = pipeline("text-generation", model="gpt2", max_new_tokens=30)

print(gpt_gen("Once upon a midnight dreary,"))

 # encoder–decoder

bart_sum = pipeline("summarization", model="facebook/bart-large-cnn")

print(bart_sum("Long input paragraph …", max_length=30, min_length=10))
```

Compare latency and output style; note how each pipeline fits its typical task. You need to execute these steps in a Jupyter Colab notebook and add a cell in the end of your notebook with your results and reflections