Shadowfox advance level task creating LM model

step 1 installing all the libraries. -->transfomers -->seaborn --> torch -->matplotlib

step 2:Loading GPT-2 and generate text

```
from transformers import GPT2Tokenizer, GPT2LMHeadModel
import torch

# Load GPT-2 tokenizer and model
tokenizer = GPT2Tokenizer.from_pretrained("gpt2")
model = GPT2LMHeadModel.from_pretrained("gpt2")
model.eval()

# Function to generate text from prompt
def generate_text(prompt, max_length=100):
    inputs = tokenizer.encode(prompt, return_tensors="pt")
    outputs = model.generate(inputs, max_length=max_length, num_return_sequences=1, do_sample=True)
    return tokenizer.decode(outputs[0], skip_special_tokens=True)

# Example usage
prompt = "The future of artificial intelligence in India is"
print(generate_text(prompt))
```

```
sr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
e secret `HF_TOKEN` does not exist in your Colab secrets.
 authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secre
u will be able to reuse this secret in all of your notebooks.
ease note that authentication is recommended but still optional to access public models or datasets.
warnings.warn(
tenizer_config.json: 100%  26.0/26.0 [00:00<00:00, 1.91kB/s]
cab.json: 100%  1.04M/1.04M [00:00<00:00, 7.28MB/s]
rges.txt: 100%  456k/456k [00:00<00:00, 15.5MB/s]
tenizer.json: 100%  1.36M/1.36M [00:00<00:00, 20.3MB/s]
nfig.json: 100%  665/665 [00:00<00:00, 54.2kB/s]
odel.safetensors: 100%  548M/548M [00:06<00:00, 130MB/s]
neration_config.json: 100%  124/124 [00:00<00:00, 5.79kB/s]
e attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `at
tting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
e attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe
e future of artificial intelligence in India is a long, dark and uncertain one… I am keenly aware that the technology is changing at a
```

step 3: Test on multiple inputs

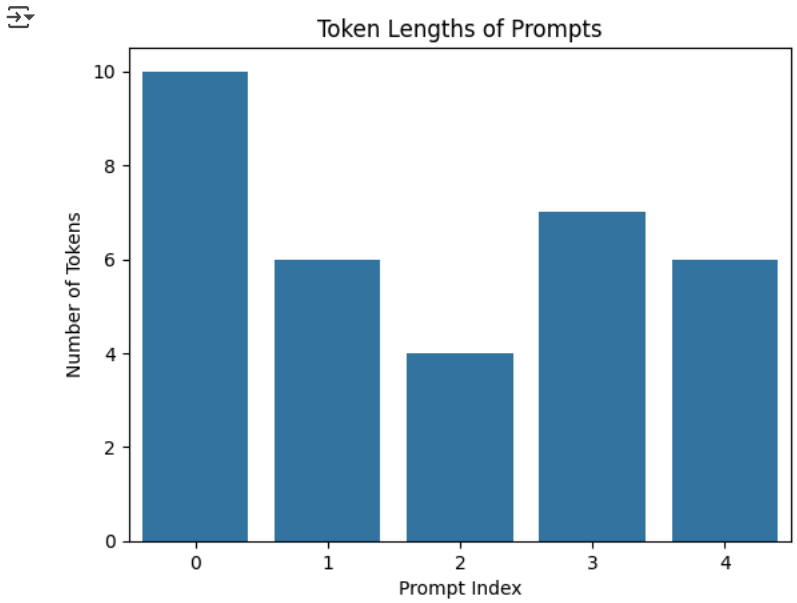
```
prompts = [
    "Once upon a time in a galaxy far away,",
    "Python is a powerful language because",
    "A healthy diet includes",
    "The economic impact of climate change is",
    "The process of machine learning involves"
]

for p in prompts:
    print(f"\nPrompt: {p}")
    print("Generated:", generate_text(p))
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `a
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.

Prompt: Once upon a time in a galaxy far away,
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `a
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Generated: Once upon a time in a galaxy far away, there was only one thing that had changed: one thing: a black hole.
```

```
To make a quick story summary, it seems that the first two movies of Star Trek are just as alien. We'll never know all the backstory, as

Prompt: Python is a powerful language because
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `a
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Generated: Python is a powerful language because of its support of multi-language support. In Scala, the Scala framework simplifies the

For that reason, Scala is designed within a much more advanced framework, including several other components:

The compiler, used for the initialisation of Scala, has already generated its own code base.



Prompt: A healthy diet includes
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `a
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Generated: A healthy diet includes ample amounts of fruit and vegetables, and plenty of refined sugars. It's only natural that you becom

However, if you want quick, easily digestible, healthy, and healthy, then it's imperative you eat the perfect diet. I'll admit I couldn'

Why is My Favorite Food Buying Guide True?

You want to eat enough fruits and veggies to fill at least 100 pounds of food stores

Prompt: The economic impact of climate change is
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `a
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Generated: The economic impact of climate change is particularly hard to estimate, especially given the sensitivity of the model to feed

According to the US Bureau of Environmental Quality, California's largest electricity producers lost 1,000 tons of sulfur dioxide, one o

Prompt: The process of machine learning involves
Generated: The process of machine learning involves learning how a character works and what characteristics of the character may be a go
```

step 4: Visualization

```
import matplotlib.pyplot as plt
import seaborn as sns

token_lengths = [len(tokenizer.encode(p)) for p in prompts]

sns.barplot(x=list(range(len(prompts))), y=token_lengths)
plt.title("Token Lengths of Prompts")
plt.xlabel("Prompt Index")
plt.ylabel("Number of Tokens")
plt.show()
```



step 5: Document your colab notebook

Title: GPT-2 Language Model Analysis using HuggingFace

** Objective**

To explore the capabilities of the GPT-2 language model in generating coherent and contextually accurate text based on various prompts. This analysis is part of the ShadowFox AI/ML internship Advanced Task.

** Tools & Libraries Used**

Python

HuggingFace Transformers

PyTorch

Matplotlib & Seaborn for visualization

** Research Questions **

Can GPT-2 generate meaningful and coherent text?

How well does GPT-2 understand different domains (e.g., science, story, tech)?

How long can GPT-2 maintain context before it starts drifting?

** Prompt Samples Tested**

“Once upon a time in a galaxy far away,”

“Python is a powerful language because”

“A healthy diet includes”

“The economic impact of climate change is”

“The process of machine learning involves”

** Observations**

GPT-2 performs well with general prompts and completes them with high fluency.

For technical or abstract topics, it may lose coherence after a few sentences.

Randomness in generation can be controlled using parameters like temperature, top_k, etc.

Visualization

Token lengths of different prompts were analyzed to understand the complexity and input size.

** Conclusion**

GPT-2 is a powerful language model that can generate creative and fluent text. It performs best with general content and has some limitations in specialized knowledge. With fine-tuning, its capabilities can be extended to domain-specific tasks.