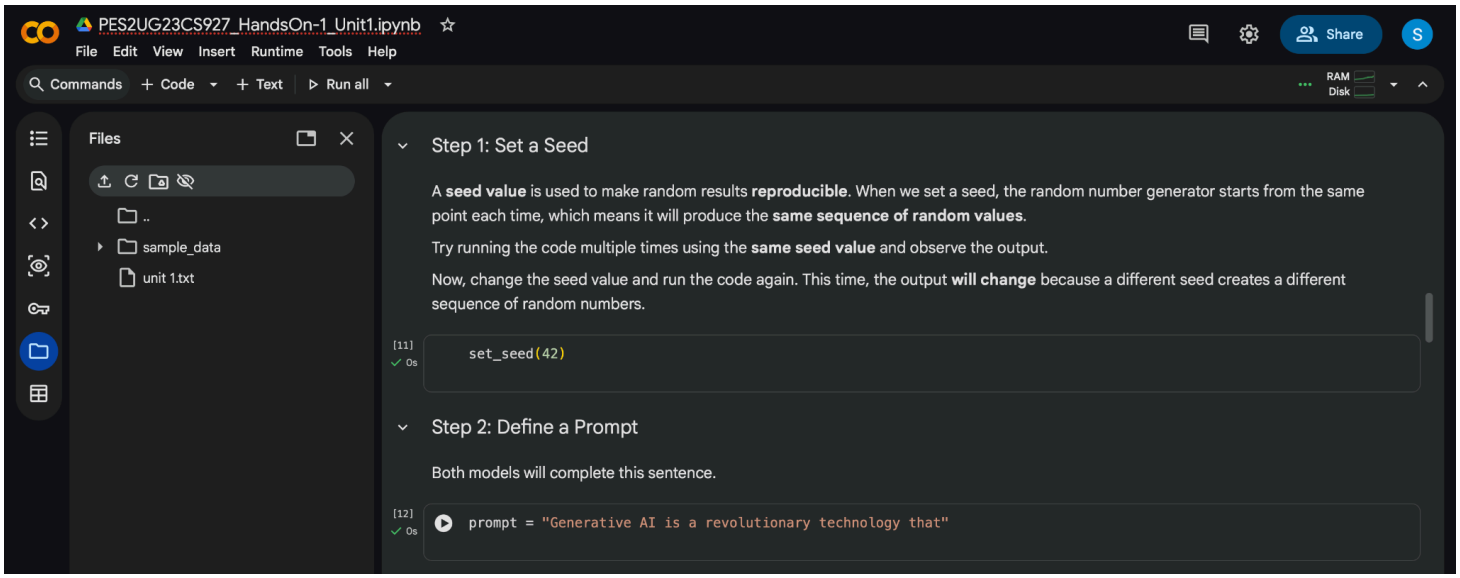
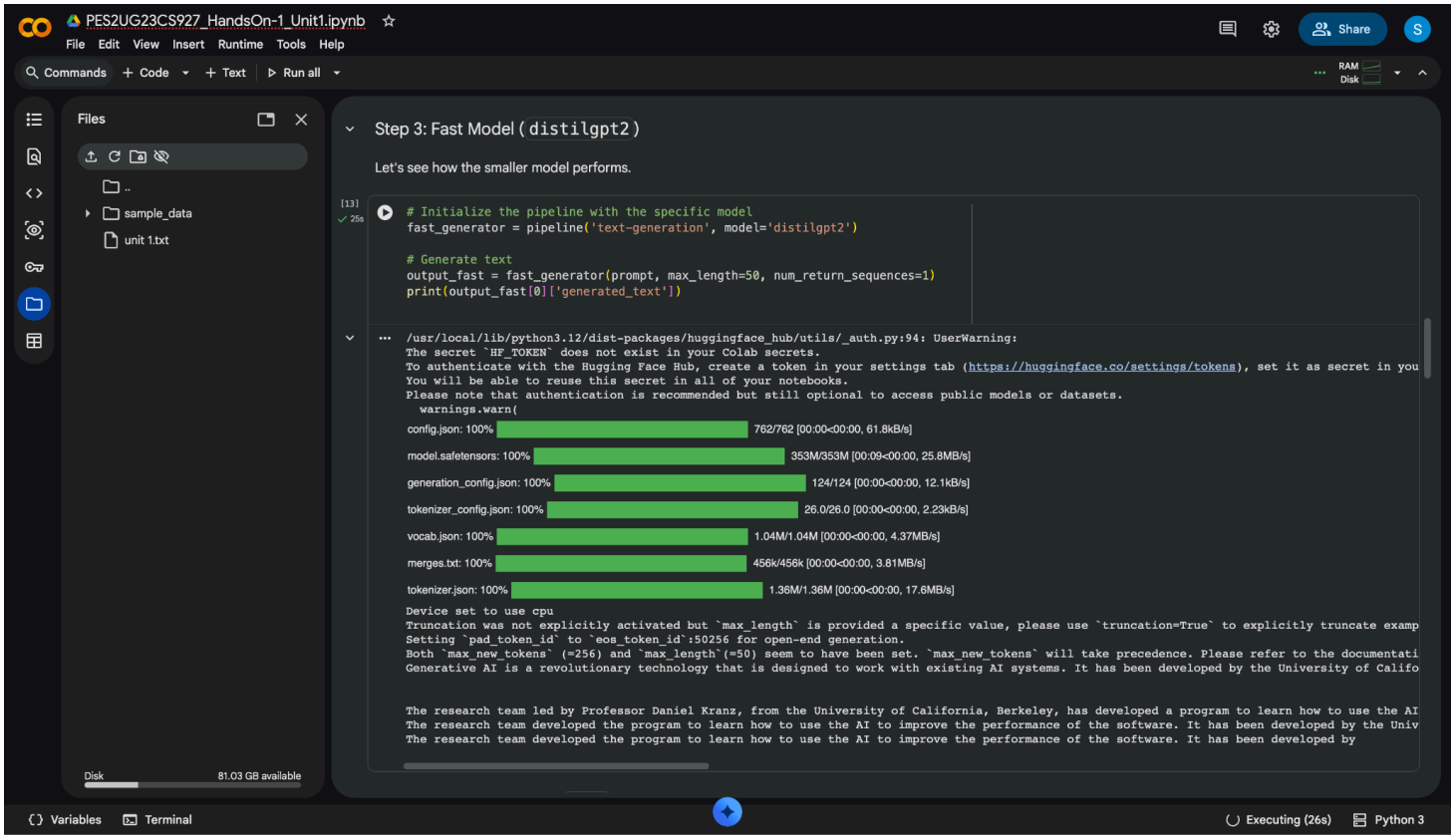
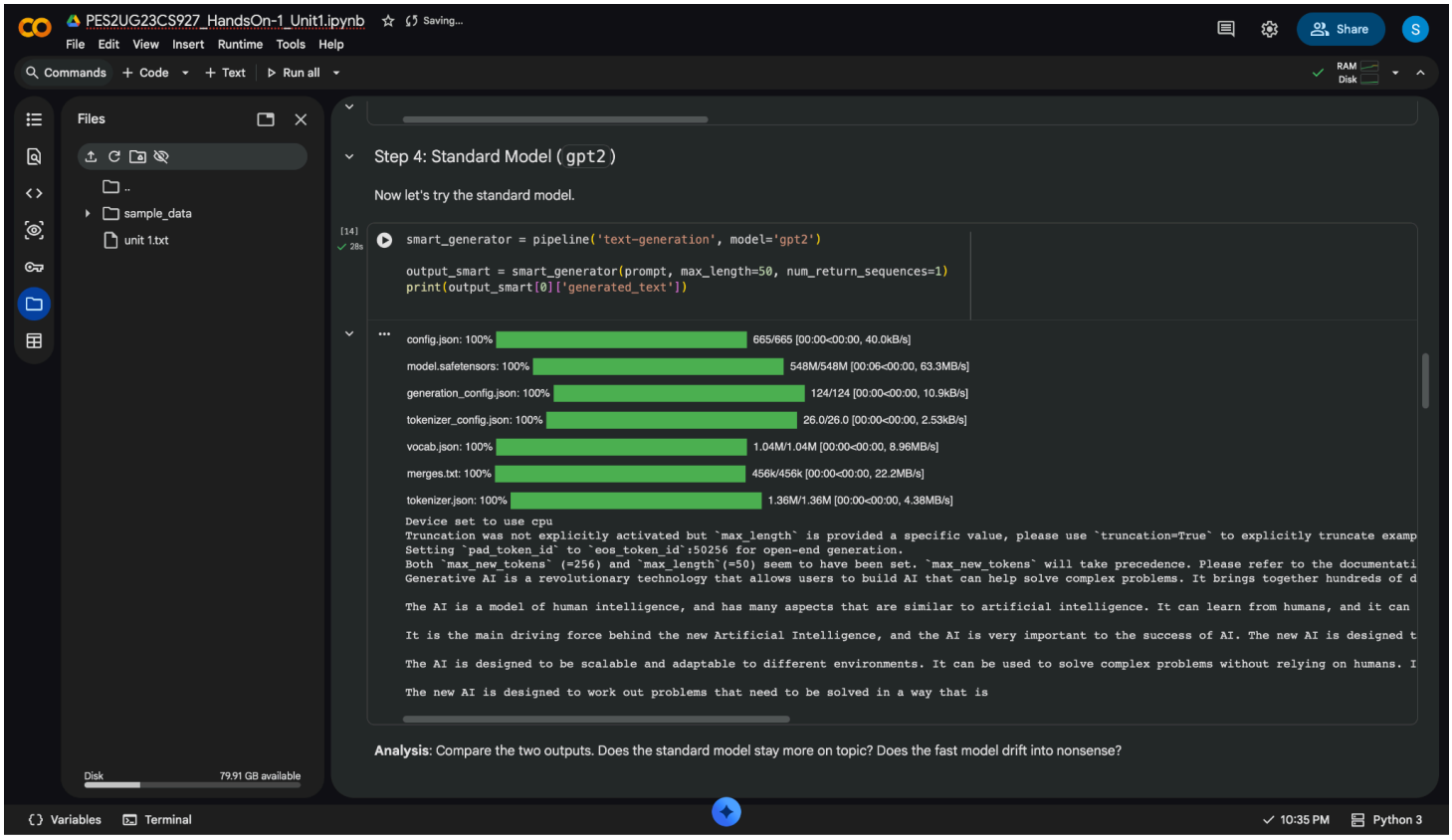
GenAI Hands On - 1

Name:	Shakirth Anisha
SRN:	PES2UG23CS927
Section:	I
Email ID	shakirthanisha@gmail.com
Phone Number	+91 7892754353

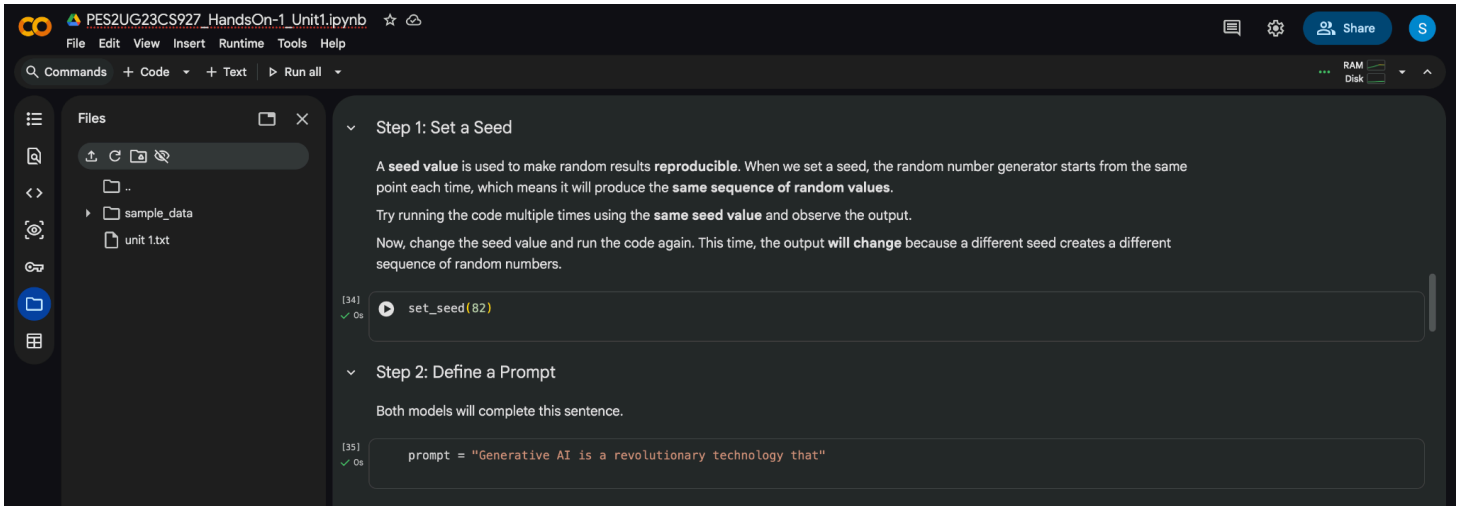
Seed - 42

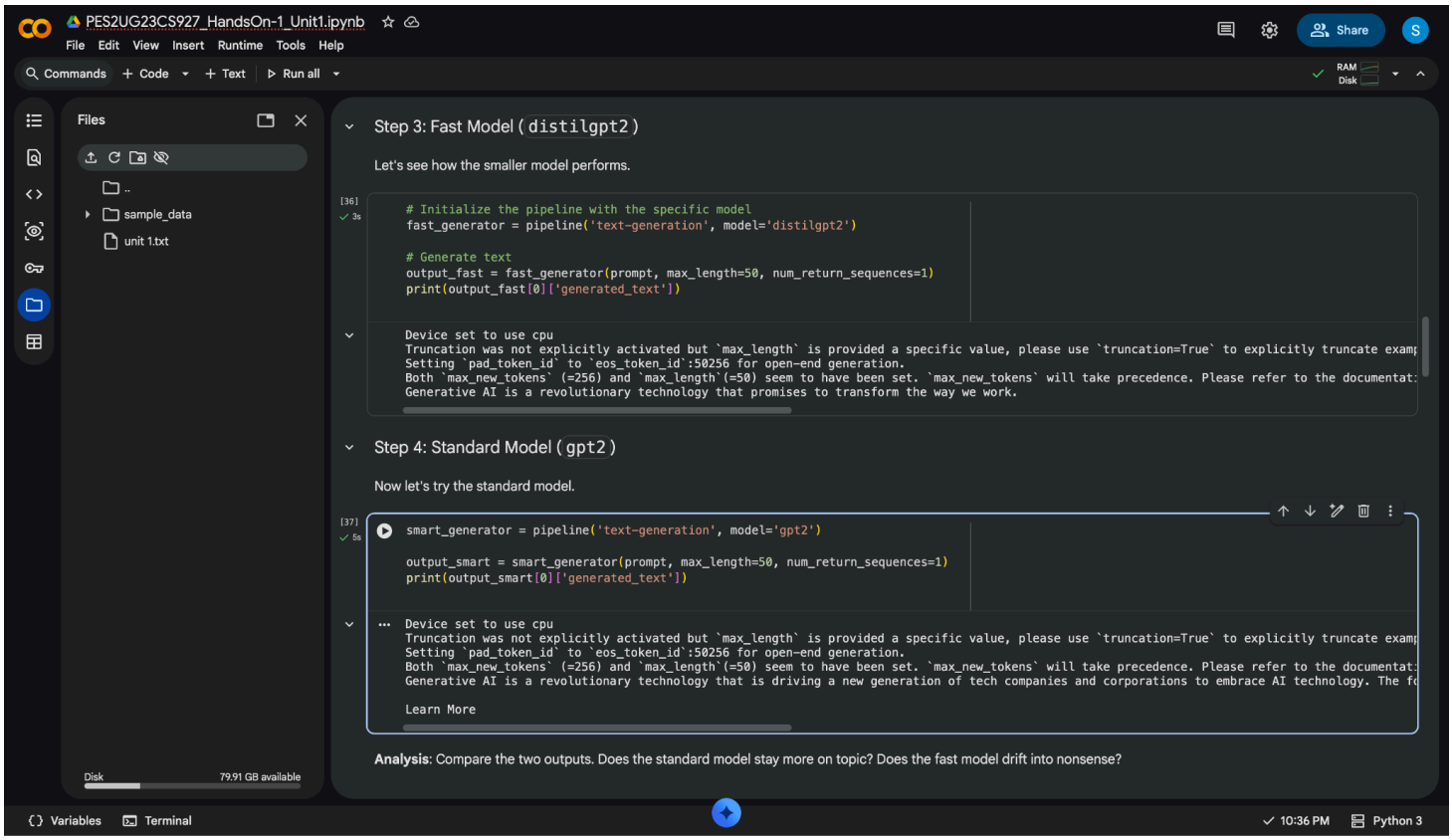



PES2UG23CS927_HandsOn-1_Unit1.ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
RAM Disk
Files
..
sample_data
unit 1.txt
Step 3: Fast Model (distilgpt2)
Let's see how the smaller model performs.
[13] 25s
# Initialize the pipeline with the specific model
fast_generator = pipeline('text-generation', model='distilgpt2')
# Generate text
output_fast = fast_generator(prompt, max_length=50, num_return_sequences=1)
print(output_fast[0]['generated_text'])
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in you
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
warnings.warn(
config.json: 100% 762/762 [00:00<00:00, 61.8kB/s]
model.safetensors: 100% 353M/353M [00:09<00:00, 25.8MB/s]
generation_config.json: 100% 124/124 [00:00<00:00, 12.1kB/s]
tokenizer_config.json: 100% 26.0/26.0 [00:00<00:00, 2.23kB/s]
vocab.json: 100% 1.04M/1.04M [00:00<00:00, 4.37MB/s]
merges.txt: 100% 456k/456k [00:00<00:00, 3.81MB/s]
tokenizer.json: 100% 1.36M/1.36M [00:00<00:00, 17.6MB/s]
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examp
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentati
Generative AI is a revolutionary technology that is designed to work with existing AI systems. It has been developed by the University of Califo
The research team led by Professor Daniel Kranz, from the University of California, Berkeley, has developed a program to learn how to use the AI
The research team developed the program to learn how to use the AI to improve the performance of the software. It has been developed by the Univ
The research team developed the program to learn how to use the AI to improve the performance of the software. It has been developed by
Disk 81.03 GB available
Variables Terminal
Executing (26s) Python 3


PES2UG23CS927_HandsOn-1_Unit1.ipynb Saving...
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
RAM Disk
Files
..
sample_data
unit 1.txt
Step 4: Standard Model (gpt2)
Now let's try the standard model.
[14] 28s
smart_generator = pipeline('text-generation', model='gpt2')
output_smart = smart_generator(prompt, max_length=50, num_return_sequences=1)
print(output_smart[0]['generated_text'])
config.json: 100% 665/665 [00:00<00:00, 40.0kB/s]
model.safetensors: 100% 548M/548M [00:06<00:00, 63.3MB/s]
generation_config.json: 100% 124/124 [00:00<00:00, 10.9kB/s]
tokenizer_config.json: 100% 26.0/26.0 [00:00<00:00, 2.53kB/s]
vocab.json: 100% 1.04M/1.04M [00:00<00:00, 8.96MB/s]
merges.txt: 100% 456k/456k [00:00<00:00, 22.2MB/s]
tokenizer.json: 100% 1.36M/1.36M [00:00<00:00, 4.38MB/s]
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examp
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentati
Generative AI is a revolutionary technology that allows users to build AI that can help solve complex problems. It brings together hundreds of d
The AI is a model of human intelligence, and has many aspects that are similar to artificial intelligence. It can learn from humans, and it can
It is the main driving force behind the new Artificial Intelligence, and the AI is very important to the success of AI. The new AI is designed t
The AI is designed to be scalable and adaptable to different environments. It can be used to solve complex problems without relying on humans. I
The new AI is designed to work out problems that need to be solved in a way that is
Analysis: Compare the two outputs. Does the standard model stay more on topic? Does the fast model drift into nonsense?
Disk 79.91 GB available
Variables Terminal
10:35 PM Python 3

Seed 82


PES2UG23CS927_HandsOn-1_Unit1.ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
RAM Disk
Files
..
sample_data
unit 1.txt
Step 1: Set a Seed
A seed value is used to make random results reproducible. When we set a seed, the random number generator starts from the same point each time, which means it will produce the same sequence of random values.
Try running the code multiple times using the same seed value and observe the output.
Now, change the seed value and run the code again. This time, the output will change because a different seed creates a different sequence of random numbers.
[34]
0s
set_seed(82)
Step 2: Define a Prompt
Both models will complete this sentence.
[35]
0s
prompt = "Generative AI is a revolutionary technology that"


PES2UG23CS927_HandsOn-1_Unit1.ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
RAM Disk
Files
..
sample_data
unit 1.txt
Step 3: Fast Model (distilgpt2)
Let's see how the smaller model performs.
[36]
3s
# Initialize the pipeline with the specific model
fast_generator = pipeline('text-generation', model='distilgpt2')
# Generate text
output_fast = fast_generator(prompt, max_length=50, num_return_sequences=1)
print(output_fast[0]['generated_text'])
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate exam
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentat
Generative AI is a revolutionary technology that promises to transform the way we work.
Step 4: Standard Model (gpt2)
Now let's try the standard model.
[37]
5s
smart_generator = pipeline('text-generation', model='gpt2')
output_smart = smart_generator(prompt, max_length=50, num_return_sequences=1)
print(output_smart[0]['generated_text'])
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate exam
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentat
Generative AI is a revolutionary technology that is driving a new generation of tech companies and corporations to embrace AI technology. The f
Learn More
Analysis: Compare the two outputs. Does the standard model stay more on topic? Does the fast model drift into nonsense?
Disk 79.91 GB available
Variables Terminal
10:36 PM Python 3

Distil Model vs Smart (GPT) Model

- **Distil model:** Smaller, faster, uses less memory, and has fewer parameters.
- **Smart (GPT) model:** Larger, computationally heavy, trained on very large datasets.
- **Distilled model:** A compressed version of a larger model that learns from a teacher model to give faster inference with minimal performance loss.

Named Entity Recognition (NER)

NER identifies and classifies entities in text such as **person, organization, location, and date**.

Transformer

A transformer is a deep learning model designed for processing sequential text using attention mechanisms.

GPT Models

GPT-2 is a **decoder-only** model for text generation. Modern GPT-based LLMs are much larger, with many layers and billions or trillions of parameters.

Transformers Library

The **Transformers** library acts as a bridge between Hugging Face models and code, enabling easy access to LLMs.

BERT vs GPT-2

- **BERT:** Designed for deep language understanding
- **GPT-2:** Designed for fluent text generation

BART

- **Full form:** Bidirectional and Auto-Regressive Transformers
- **Architecture:** Encoder–Decoder
- **Training:** Denoising autoencoder
- **Best for:** Text generation, summarization, translation

BART vs BERT:
BART supports text generation using an encoder–decoder architecture, while BERT is encoder-only and focused on language understanding.

Hugging Face

Hugging Face is an NLP platform providing pre-trained models, datasets, and tools for building and deploying machine learning applications.
