GenAI Hands-on 1

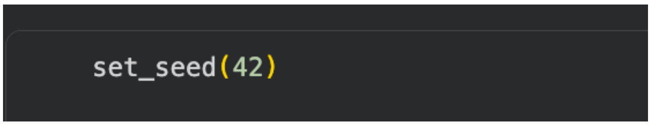
Name: Bhavana Ramkumar
SRN: PES2UG23CS905
Section: B

Class notes

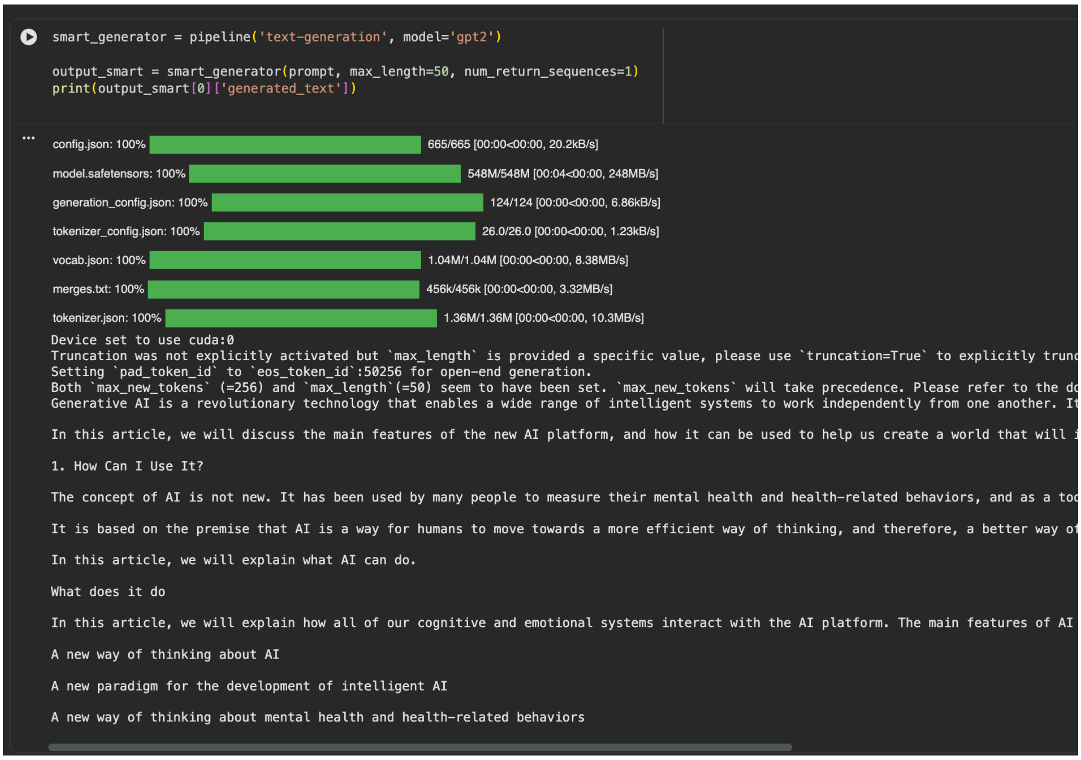
Q)SEED Value

The set_seed() function is used to initialize the pseudo-random number generator. Think of it like giving a specific starting point to a complex dice roll.
If you use the same seed value repeatedly (e.g., set_seed(42)): You will get the exact same sequence of 'random' numbers and thus the exact same output from any operations that use those random numbers (like the text generation in our example). This is crucial for reproducibility in experiments.

If you change the seed value (e.g., from 42 to 1 or 100): The random number generator will start from a different point, and consequently, it will produce a different sequence of pseudo-random numbers. This will lead to a different output for random operations like text generation. The specific numerical value of the seed itself (whether it's higher or lower) doesn't inherently make the output 'better' or 'worse'; it just ensures that the randomness unfolds in a unique, yet repeatable, way for that specific seed.

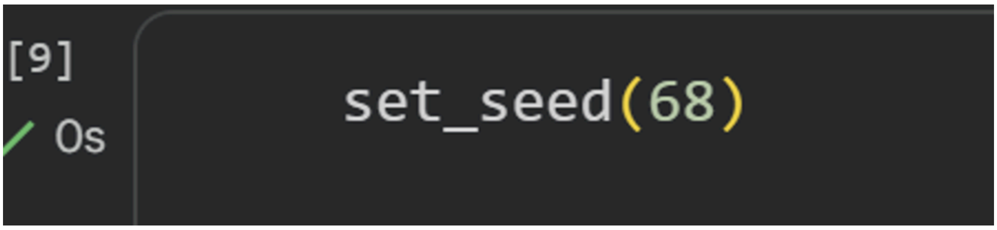


```
set_seed(42)
```

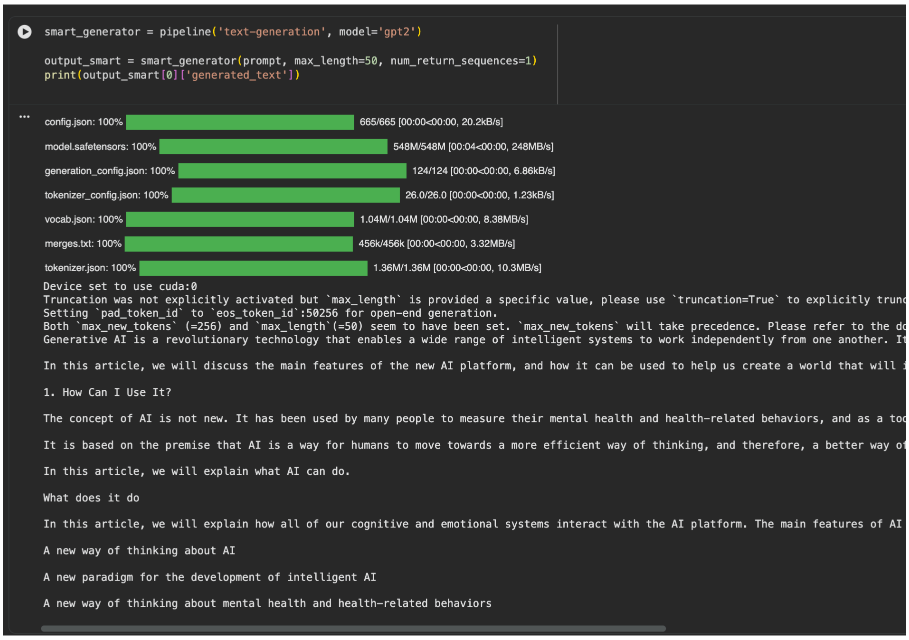


```
smart_generator = pipeline('text-generation', model='gpt2')

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, 20.2kB/s]
model.safetensors: 100%  548M/548M [00:04<00:00, 248MB/s]
generation_config.json: 100%  124/124 [00:00<00:00, 6.86kB/s]
tokenizer_config.json: 100%  26.0/26.0 [00:00<00:00, 1.23kB/s]
vocab.json: 100%  1.04M/1.04M [00:00<00:00, 8.38MB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 3.32MB/s]
tokenizer.json: 100%  1.36M/1.36M [00:00<00:00, 10.3MB/s]
Device set to use cuda:0
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly trunc
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 do
Generative AI is a revolutionary technology that enables a wide range of intelligent systems to work independently from one another. It

In this article, we will discuss the main features of the new AI platform, and how it can be used to help us create a world that will i

1. How Can I Use It?

The concept of AI is not new. It has been used by many people to measure their mental health and health-related behaviors, and as a too

It is based on the premise that AI is a way for humans to move towards a more efficient way of thinking, and therefore, a better way of

In this article, we will explain what AI can do.

What does it do

In this article, we will explain how all of our cognitive and emotional systems interact with the AI platform. The main features of AI

A new way of thinking about AI

A new paradigm for the development of intelligent AI

A new way of thinking about mental health and health-related behaviors
```



```
[9]
0s
set_seed(68)
```



```
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, 20.2kB/s]
model.safetensors: 100% 548M/548M [00:04<00:00, 248MB/s]
generation_config.json: 100% 124/124 [00:00<00:00, 6.86kB/s]
tokenizer_config.json: 100% 26.0/26.0 [00:00<00:00, 1.23kB/s]
vocab.json: 100% 1.04M/1.04M [00:00<00:00, 8.38MB/s]
merges.txt: 100% 456k/456k [00:00<00:00, 3.32MB/s]
tokenizer.json: 100% 1.36M/1.36M [00:00<00:00, 10.3MB/s]
Device set to use cuda:0
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly trunc
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 do
Generative AI is a revolutionary technology that enables a wide range of intelligent systems to work independently from one another. It

In this article, we will discuss the main features of the new AI platform, and how it can be used to help us create a world that will i

1. How Can I Use It?

The concept of AI is not new. It has been used by many people to measure their mental health and health-related behaviors, and as a too

It is based on the premise that AI is a way for humans to move towards a more efficient way of thinking, and therefore, a better way of

In this article, we will explain what AI can do.

What does it do

In this article, we will explain how all of our cognitive and emotional systems interact with the AI platform. The main features of AI

A new way of thinking about AI

A new paradigm for the development of intelligent AI

A new way of thinking about mental health and health-related behaviors
```

Q)Difference between distilled model and plain model

Aspect	Plain Model	Distilled Model
Definition	The original large model trained directly on the dataset	A smaller model trained to mimic a larger (teacher) model
Model Size	Large number of parameters	Fewer parameters (compressed)
Training Method	Trained using ground-truth labels	Trained using soft targets from the teacher model (knowledge distillation)
Accuracy	High accuracy	Slightly lower but close to the teacher model
Inference Speed	Slower	Faster
Memory Usage	High	Low
Deployment	Difficult on edge/mobile devices	Suitable for edge and mobile devices
Example	BERT-Large	DistilBERT

Tokenization

Tokenization is the process of breaking down a large string of text into smaller, manageable pieces called tokens. Every model has different tokenizer we cant use same for all models

Q)What is POS Tagging?

POS (Part-of-Speech) tagging means labeling each word in a sentence with its grammatical role (noun, verb, adjective, etc.).

Word	POS Tag	Full Form	Meaning
Transformers	NNS	Noun, Plural	A plural noun
revolutionized	VBD	Verb, Past Tense	An action that happened in the past
NLP	NNP	Proper Noun, Singular	A specific name (proper noun)

Tag	Meaning
NN	Noun (singular)
NNS	Noun (plural)
NNP	Proper noun
VB	Verb (base form)
VBD	Verb (past tense)
VBG	Verb (-ing)
JJ	Adjective
RB	Adverb

English POS tagsets have 35–45 tags.

NER

When an NER model reads a sentence, it doesn't just see words; it assigns them labels.
Eg; apple can be a fruit or a phone

Q) What is temperature?

Temperature in generative AI controls the randomness of the output. A higher temperature leads to more creative and diverse, but potentially less coherent, text. A lower temperature results in more predictable and focused, but less creative, output.

Unit1_Benchmark Observation Table

Experiment	Task	Model	Classification (Success / Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Exp 1	Text Generation	BERT (bert-base-uncased)	Failure	Generated only repeated dots after the prompt; no meaningful continuation.	BERT is an encoder-only model and is not trained for autoregressive next-token generation.
Exp 1	Text Generation	RoBERTa (roberta-base)	Failure	Output stopped at the prompt without generating new text.	RoBERTa is also encoder-only, optimized for understanding tasks, not generation.
Exp 1	Text Generation	BART (facebook/bart-base)	Failure	Generated long text but it was incoherent, repetitive, and noisy (random words).	BART supports generation, but BartForCausalLM weights were randomly initialized and not fine-tuned.
Exp 2	Masked Language Modeling	BERT (bert-base-uncased)	Success	Correctly predicted masked words such as create, generate, produce.	BERT is trained using Masked Language Modeling (MLM) with bidirectional context.
Exp 2	Masked Language Modeling	RoBERTa (roberta-base)	Success	Accurately predicted context-aware words like generate and create.	RoBERTa improves MLM training with more data and removes the NSP objective.
Exp 2	Masked Language Modeling	BART (facebook/bart-base)	Partial Success	Predicted reasonable words but with lower confidence scores than BERT/RoBERTa.	BART is trained for denoising sequence-to-sequence, not pure MLM like encoder-only models.

Exp 3	Question Answering	BERT (bert-base-uncased)	Partial Success	Extracted a mostly correct answer but with very low confidence score.	Base BERT is not fine-tuned on SQuAD, so the QA head performs weakly.
Exp 3	Question Answering	RoBERTa (roberta-base)	Partial Success	Returned incomplete answer fragments (missing full phrase).	QA layers are randomly initialized without task-specific fine-tuning.
Exp 3	Question Answering	BART (facebook/bart-base)	Success (Low Confidence)	Generated a correct full answer but with moderate confidence.	BART's encoder–decoder architecture helps QA, but it still lacks QA-specific fine-tuning.

Models

1.BERT is an encoder-only model that reads text bidirectionally, allowing it to understand context from both past and future words. It is trained using masked language modeling, making it highly effective for tasks like text classification and fill-in-the-blank problems, but it cannot generate new text.
2.RoBERTa follows the same encoder-only architecture as BERT but is trained on more data with improved strategies, leading to better performance on understanding tasks, though it still cannot generate text.
3.BART, in contrast, uses an encoder–decoder architecture where the encoder processes the input and the decoder generates output. Trained using a denoising objective, BART excels at tasks such as summarization, translation, and text generation, combining both understanding and generation capabilities.

Daily Horoscope Generator – Project Overview

Goal

The project aims to automatically generate vague but convincing daily horoscopes for any zodiac sign. By leveraging a language model (GPT-2 or DistilGPT-2), the system produces mystical, uplifting, and personalized fortunes that mimic the style of traditional astrology readings.

How It Works

1. **Input**: User provides a zodiac sign (e.g., *Taurus*).
2. **Prompt Engineering**: A carefully crafted text prompt guides the model to generate horoscope-style content.
3. **Text Generation**: GPT-2 generates a continuation of the prompt, producing a horoscope.
4. **Output**: A short, poetic horoscope that includes themes of emotions, opportunities, and subtle warnings.

Core Functionality

1. Generate horoscopes for all 12 zodiac signs

2. Reproducible results (seed control)

3. Clean, coherent output

4. Batch generation capability