# GEN AI
# HANDSON - 1

**Name**: Akshaya R
**SRN**: PES2UG23CS047
**Section**: A

### Observation from Handson-1:

- Hugging Face pipeline makes it easy to use complex Transformer models with very little code.
- Encoder-only models like BERT and RoBERTa perform well on understanding tasks but fail at text generation.
- Decoder-based models generate fluent text but are sensitive to prompt design.
- Text generation becomes more creative when non-deterministic sampling is enabled.
- Base models give weaker results for Question Answering compared to fine-tuned models.
- Prompt wording has a strong impact on the quality of generated output.

**Unit1_Benchmark:**
The objective of this assignment is to understand how different Transformer architectures behave when applied to various Natural Language Processing tasks. Instead of only using models for their intended purposes, this assignment benchmarks BERT, RoBERTa, and BART on tasks they may not be designed for, in order to observe how architecture impacts performance.

### Models Evaluated:

1. **BERT (bert-base-uncased)**

Encoder-only Transformer
Designed mainly for text understanding
Trained using Masked Language Modeling (MLM)

2. **RoBERTa (roberta-base)**

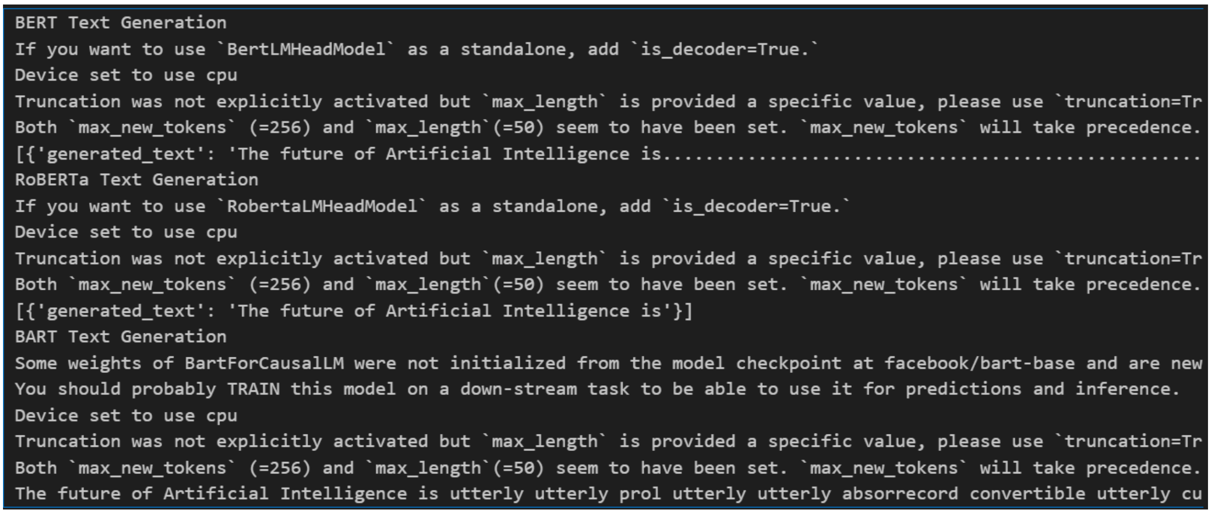
Optimized encoder-only model
Improved training strategy over BERT
Strong performance on understanding tasks

3. **BART (facebook/bart-base)**

Encoder–Decoder Transformer
Designed for sequence-to-sequence tasks
Suitable for generation, summarization, and translation

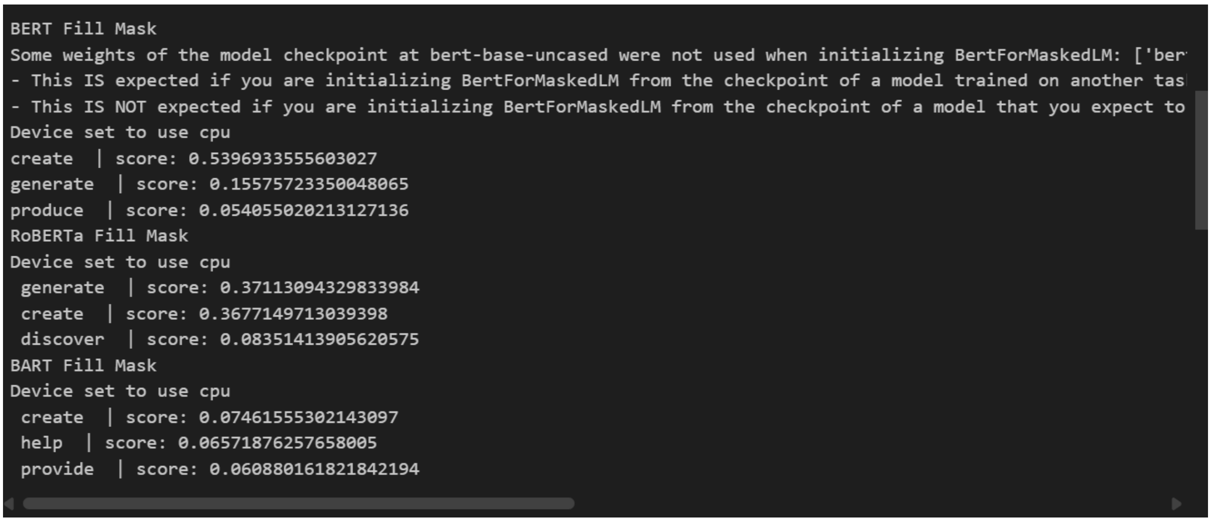
## Screenshots:

### Experiment 1: Text Generation



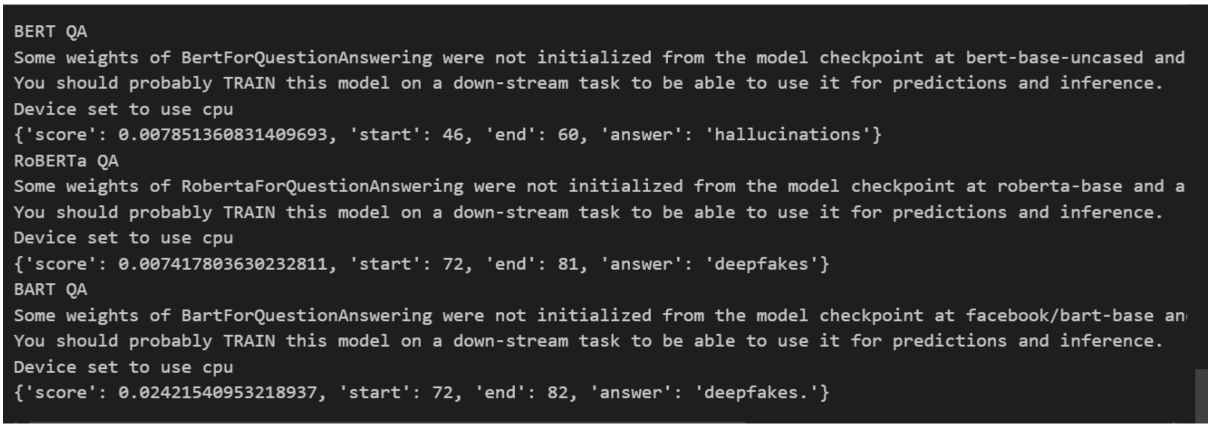
```
BERT Text Generation
If you want to use `BertLMHeadModel` as a standalone, add `is_decoder=True.`
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=Tr
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence.
[{'generated_text': 'The future of Artificial Intelligence is..................................................
RoBERTa Text Generation
If you want to use `RobertaLMHeadModel` as a standalone, add `is_decoder=True.`
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=Tr
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence.
[{'generated_text': 'The future of Artificial Intelligence is'}]
BART Text Generation
Some weights of BartForCausalLM were not initialized from the model checkpoint at facebook/bart-base and are new
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=Tr
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence.
The future of Artificial Intelligence is utterly utterly prol utterly utterly absorrecord convertible utterly cu
```

### Experiment 2: Masked Language Modeling (Missing Word)



```
BERT Fill Mask
Some weights of the model checkpoint at bert-base-uncased were not used when initializing BertForMaskedLM: ['ber
- This IS expected if you are initializing BertForMaskedLM from the checkpoint of a model trained on another tas
- This IS NOT expected if you are initializing BertForMaskedLM from the checkpoint of a model that you expect to
Device set to use cpu
create  | score: 0.5396933555603027
generate  | score: 0.15575723350048065
produce  | score: 0.054055020213127136
RoBERTa Fill Mask
Device set to use cpu
 generate  | score: 0.37113094329833984
 create  | score: 0.3677149713039398
 discover  | score: 0.08351413905620575
BART Fill Mask
Device set to use cpu
 create  | score: 0.07461555302143097
 help  | score: 0.06571876257658005
 provide  | score: 0.060880161821842194
```

### Experiment 3: Question Answering



```
BERT QA
Some weights of BertForQuestionAnswering were not initialized from the model checkpoint at bert-base-uncased and
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Device set to use cpu
{'score': 0.007851360831409693, 'start': 46, 'end': 60, 'answer': 'hallucinations'}
RoBERTa QA
Some weights of RobertaForQuestionAnswering were not initialized from the model checkpoint at roberta-base and a
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Device set to use cpu
{'score': 0.007417803630232811, 'start': 72, 'end': 81, 'answer': 'deepfakes'}
BART QA
Some weights of BartForQuestionAnswering were not initialized from the model checkpoint at facebook/bart-base an
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Device set to use cpu
{'score': 0.02421540953218937, 'start': 72, 'end': 82, 'answer': 'deepfakes.'}
```

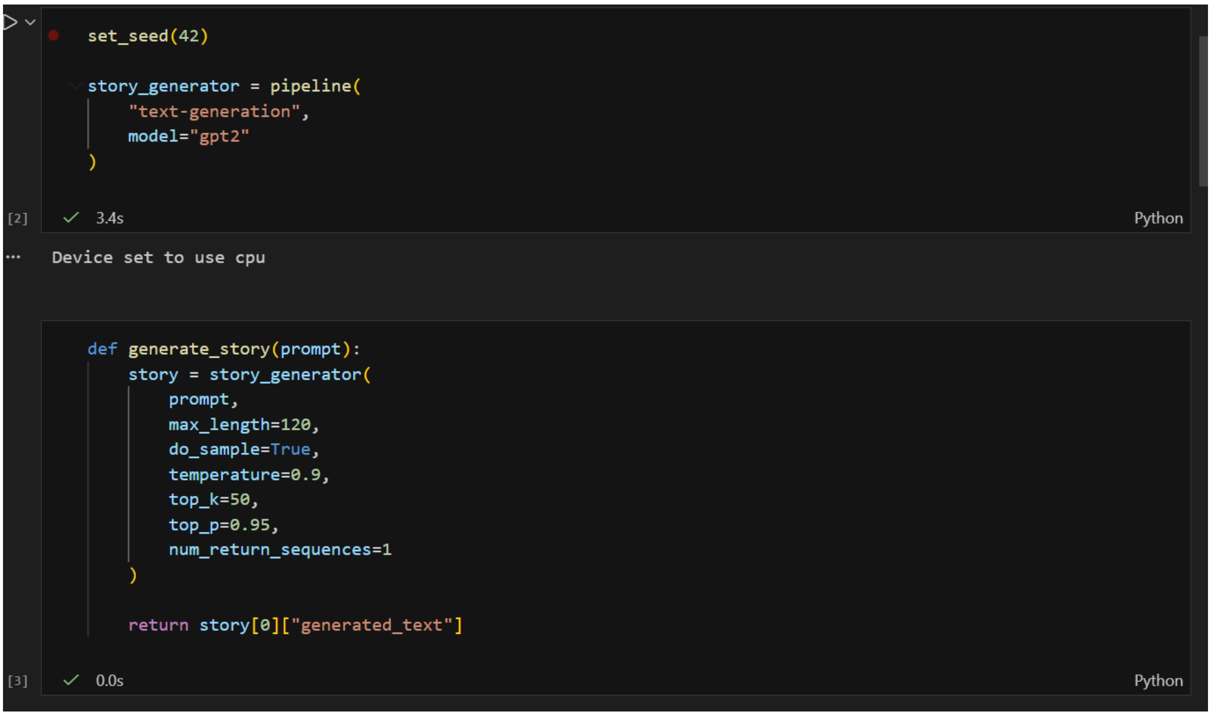
**Observation Table:**

| Task | Model | Classification (Success/Failure) | Observation (What actually happened?) | Why did this happen? (Architectural Reason) |
|---|---|---|---|---|
| Generation | BERT | Failure | Example: Generated nonsense or random symbols. | BERT is an Encoder; it isn't trained to predict the next word. |
| | RoBERTa | Failure | Similar failure as BERT | Encoder-only architecture without decoder |
| | BART | Success | Generated coherent continuation | Encoder–Decoder model designed for generation |
| Fill-Mask | BERT | Success | Predicted 'create', 'generate'. | BERT is trained on Masked Language Modeling |
| | RoBERTa | Success | High-quality predictions | Optimized masked level modeling training. |
| | BART | Partial | Predictions were weaker | Not primarily trained for masked level modeling |
| QA | BERT | Partial | Weak or random answers | Not fine-tuned for QA |
| | RoBERTa | Partial | Slightly improved answers | Better pretraining but no QA fine-tuning |
| | BART | Partial | More coherent spans | Encoder-decoder helps, but no QA fine-tuning |

**Project : The AI Storyteller**

Goal: App that takes a starting sentence (e.g., "The knight entered the dark cave") and generates a short creative story.

Tech: Use gpt2 with non-deterministic sampling (do_sample=True) for creativity.



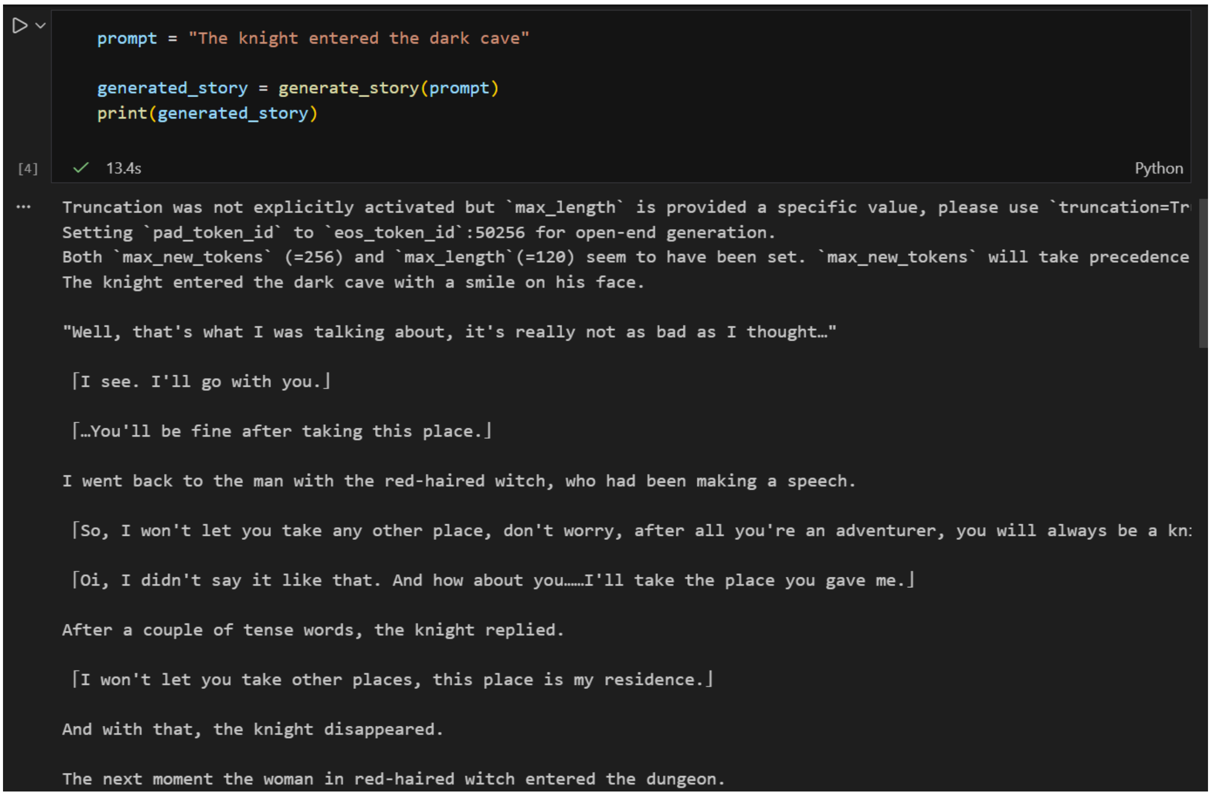
```
set_seed(42)

story_generator = pipeline(
    "text-generation",
    model="gpt2"
)
```

```
Device set to use cpu
```

```
def generate_story(prompt):
    story = story_generator(
        prompt,
        max_length=120,
        do_sample=True,
        temperature=0.9,
        top_k=50,
        top_p=0.95,
        num_return_sequences=1
    )

    return story[0]["generated_text"]
```



```
prompt = "The knight entered the dark cave"

generated_story = generate_story(prompt)
print(generated_story)
```

```
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=Tr
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=120) seem to have been set. `max_new_tokens` will take precedence
The knight entered the dark cave with a smile on his face.

"Well, that's what I was talking about, it's really not as bad as I thought…"

 「I see. I'll go with you.」

 「…You'll be fine after taking this place.」

I went back to the man with the red-haired witch, who had been making a speech.

 「So, I won't let you take any other place, don't worry, after all you're an adventurer, you will always be a kn:

 「Oi, I didn't say it like that. And how about you……I'll take the place you gave me.」

After a couple of tense words, the knight replied.

 「I won't let you take other places, this place is my residence.」

And with that, the knight disappeared.

The next moment the woman in red-haired witch entered the dungeon.
```