Unit 1 Assignment: The Model Benchmark Challenge

Name : Ananya Shet

SRN : PES2UG23CS065

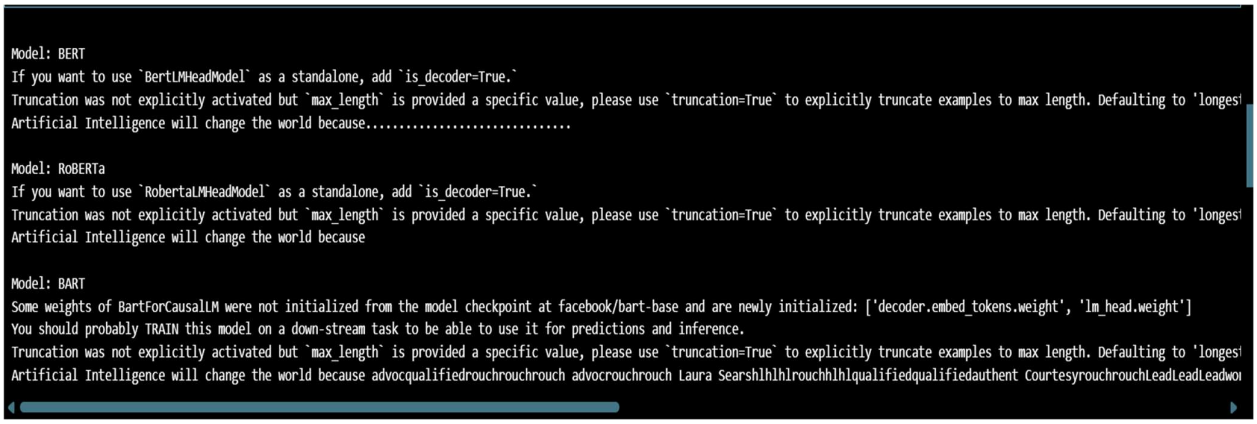
Section : A

Objective: To understand how model architecture affects performance.

We are comparing 3 model ie

- **BERT** – Encoder only
- **RoBERTa** - Improved Encoder-only
- **BART** - Encoder–Decoder

Text Generation :



Fill Mask:

```python
for name, model in models.items():
    print(f"\nModel: {name}")
    fill = pipeline("fill-mask", model=model)

    # Use the correct mask token for that model
    masked = f"Generative AI will transform the {fill.tokenizer.mask_token} industry."

    predictions = fill(masked)
    for p in predictions[:3]:
        print(p['token_str'], ":", round(p['score'], 3))
```

[6] Python

```
...
Model: BERT
Some weights of the model checkpoint at bert-base-uncased were not used when initializing BertForMaskedLM: ['bert.pooler.dense.bias', 'bert.pooler.dense.weight', 'cls.seq_relationship.l
- This IS expected if you are initializing BertForMaskedLM from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClass
- This IS NOT expected if you are initializing BertForMaskedLM from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification mode
healthcare : 0.077
it : 0.058
software : 0.049

Model: RoBERTa
 tech : 0.107
 healthcare : 0.091
 entertainment : 0.069

Model: BART
 entire : 0.095
 health : 0.043
 healthcare : 0.028
```

Question Answering:

```python
context = "Generative AI creates new content such as text, images, audio, and code."
question = "What does Generative AI create?"

for name, model in models.items():
    print(f"\nModel: {name}")
    try:
        qa = pipeline("question-answering", model=model)
        ans = qa(question=question, context=context)
        print("Answer:", ans['answer'])
    except Exception as e:
        print("QA Failed:", e)
```

Python

```
Model: BERT
Some weights of BertForQuestionAnswering were not initialized from the model checkpoint at bert-base-uncased and are newly initialized: ['qa_outputs.bias', 'qa_outputs.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Answer: AI creates new content

Model: RoBERTa
Some weights of RobertaForQuestionAnswering were not initialized from the model checkpoint at roberta-base and are newly initialized: ['qa_outputs.bias', 'qa_outputs.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Answer: text, images, audio,

Model: BART
Some weights of BartForQuestionAnswering were not initialized from the model checkpoint at facebook/bart-base and are newly initialized: ['qa_outputs.bias', 'qa_outputs.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Answer: audio, and code
```

Observation Table

Task	Model	Classification (Success/Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Generation	BERT	Failure	Did not generate meaningful continuation of the sentence.	BERT is an Encoder; it isn't trained to predict the next word.
	RoBERTa	Failure	Could not continue the prompt and	RoBERTa is also an Encoder-only model

Task	Model	Classification (Success/Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
			produced no fluent output.	and cannot predict next tokens sequentially.
	BART	Partial Failure	The model produced meaningless, or random words instead of a smooth and logical continuation of the sentence.	BART is an Encoder–Decoder model trained mainly for seq2seq tasks, not causal language modeling like GPT.
Fill-Mask	BERT	*Success*	*Predicted 'create', 'generate'.*	*BERT is trained on Masked Language Modeling (MLM).*
	RoBERTa	Success	Produced highly accurate and contextually relevant predictions.	RoBERTa is optimized for MLM with better pretraining.
	BART	Partial Success	Predictions were less accurate and less confident.	BART is not primarily designed for MLM tasks.
QA	BERT	Success	Returned correct but generic answer from the context.	Encoder models are strong at extractive Question Answering.
	RoBERTa	Success	Extracted precise answer spans from the context.	Better contextual encoding improves span selection.
	BART	Partial Success	Provided only partial answers from the context.	Seq2Seq models are not optimized for extractive span-based QA.

Conclusion :

Encoder only models are excellent for understanding task but fails at free text generation.

Encoder – Decoder models are designed for sequence generation and perform well in text generation tasks.