Gen AI Assignment - 1 Observations

Name: Ananya Pandurang Prabhu
SRN: PES2UG23CS063
Section: A

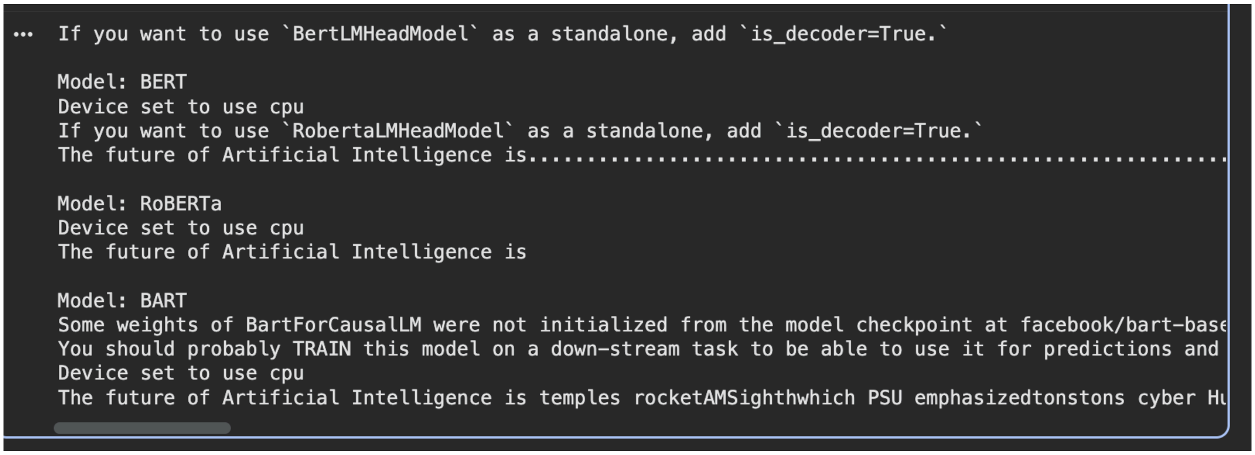
The Models given to Test:

1. BERT (bert-base-uncased): An Encoder-only model (designed for understanding, not generation).
2. RoBERTa (roberta-base): An optimized Encoder-only model.
3. BART (facebook/bart-base): An Encoder-Decoder model (designed for seq2seq tasks like translation/generation).

The first experiment was that of text generation with the given prompt of “The future of Artificial Intelligence is”.

- BERT gave an output of just …….. Since it is an encoder only architecture it is meant to only read and not write or generate texts and thus just gave an output of dots.
- Similar to BERT, the RoBERTa model is also primarily an optimized encoder only architecture model and thus cannot generate any out of the box text for a given prompt. It thus gave an empty string as output.
- Lastly, BART is an encoder-decoder model and thus can read and write text and is capable of generating text. In our case, it generated text that made no sense maybe because it wasn’t appropriately fine tuned and we are just using the base model or due to some out of vocabulary words.

BART was the best model for this task.

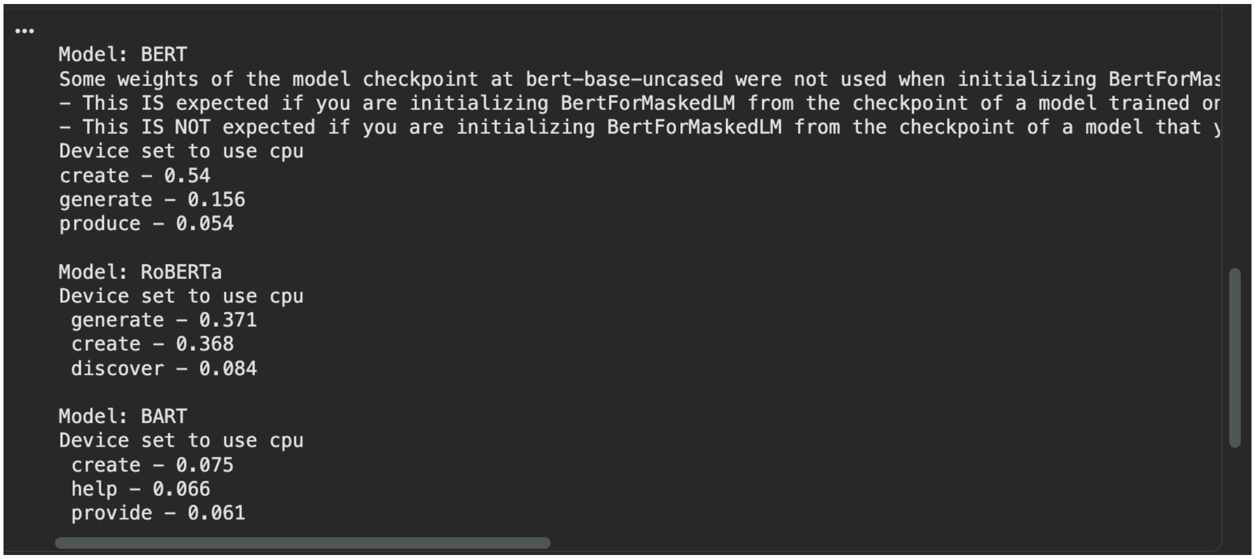


```
•••  If you want to use `BertLMHeadModel` as a standalone, add `is_decoder=True.`

     Model: BERT
     Device set to use cpu
     If you want to use `RobertaLMHeadModel` as a standalone, add `is_decoder=True.`
     The future of Artificial Intelligence is...........................................................

     Model: RoBERTa
     Device set to use cpu
     The future of Artificial Intelligence is

     Model: BART
     Some weights of BartForCausalLM were not initialized from the model checkpoint at facebook/bart-base
     You should probably TRAIN this model on a down-stream task to be able to use it for predictions and
     Device set to use cpu
     The future of Artificial Intelligence is temples rocketAMSighthwhich PSU emphasizedtonstons cyber Hu
```

The second experiment was that of predicting the missing word in the prompt "The goal of Generative AI is to [MASK] new content."

- BERT predicted the top 3 words as ‘create’ with a score of 0.54, ‘generate’ with a score of 0.156 and ‘produce’ with a score of 0.054.
- RoBERTa predicted ‘generate’ with a score of 0.371, ‘create’ with a score of 0.368 and ‘discover’ with a score of 0.084.
- BART predicted ‘create’ with a score of 0.075, ‘help’ with a score of 0.066 and ‘provide’ with a score of 0.061.

BERT and RoBERTa were in fact the best models for this task.



```
...
    Model: BERT
    Some weights of the model checkpoint at bert-base-uncased were not used when initializing BertForMas
    - This IS expected if you are initializing BertForMaskedLM from the checkpoint of a model trained or
    - This IS NOT expected if you are initializing BertForMaskedLM from the checkpoint of a model that y
    Device set to use cpu
    create - 0.54
    generate - 0.156
    produce - 0.054

    Model: RoBERTa
    Device set to use cpu
     generate - 0.371
     create - 0.368
     discover - 0.084

    Model: BART
    Device set to use cpu
     create - 0.075
     help - 0.066
     provide - 0.061
```

The third task was that of question and answer - "What are the risks?" based on the context of the string "Generative AI poses significant risks such as hallucinations, bias, and deepfakes."

- BERT extracted ', and deepfakes' as the answer, with a score of 0.0157. This shows that it didn't really find a very confident answer. It also indicated that BertForQuestionAnswering weights were not initialized, implying it would, in fact need training.
- RoBERTa model extracted 'Generative AI poses significant risks such as hallucinations, bias,' as the answer, also with a very low score of 0.0076. It covered more of the context but still lacked confidence. It also had a warning about RobertaForQuestionAnswering weights not being initialized.
- BART model extracted 'AI poses significant' as the answer, with a low score of 0.015. It also had a warning about BartForQuestionAnswering weights not being initialized. The low scores and initialization warnings for question-answering indicate that these base models might not be well suited for this task without further fine-tuning on a question-answering dataset.

Task	Model	Classification (Success/Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Generation	BERT	Failure	………… Gave an output of a sequence of dots.	BERT is an Encoder; it isn't trained to predict the next word.
	RoBERTa	Failure	Empty string	RoBERTa is primarily an Encoder; it isn't trained to predict the next word.
	BART	Success	“The future of Artificial Intelligence is Chou apex TC Reid Reid TC TC slaughtered slaughtered slaughtered addon II Reid Reid Reid WINroup Reid Reid II taxes Reid Reidrusroup”	BART is a sequence-to-sequence model capable of generation. Without specific fine-tuning on a creative text generation task, its open-ended generation from a short prompt is often incoherent.

Fill-Mask	BERT	*Success*	Predicted 'create' (0.54), 'generate' (0.156), 'produce' (0.054).	*BERT is trained on Masked Language Modeling (MLM).*
	RoBERTa	Success	Predicted 'generate' (0.371), 'create' (0.368), 'discover' (0.084).	RoBERTa is also a Masked Language Model (MLM) and performs exceptionally well in filling in masked tokens, similar to BERT.
	BART	Success	Predicted 'create' (0.075), 'help' (0.066), 'provide' (0.061).	BART's denoising autoencoder architecture allows it to reconstruct masked tokens. While capable, its approach to masked language modeling might differ slightly from dedicated MLMs like BERT/RoBERTa.

QA	BERT	Failure	Extracted ', and deepfakes' (score of 0.0157).	The base bert-base-uncased model is not fine-tuned for Question Answering. QA models require additional fine-tuning on datasets like SQuAD to learn to extract precise answers from a given context.
	RoBERTa	Failure	Extracted 'Generative AI poses significant risks such as hallucinations, bias,' with a very low score (0.0076).	Similar to BERT, the base roberta-base model is not designed for Question Answering out-of-the-box and needs fine-tuning on a QA dataset to effectively identify answer spans. The warning confirms uninitialized QA weights.
	BART	Failure	Extracted 'AI poses	Like BERT and RoBERTa, the

			significant' with a very low score (0.015).	base bart model is not fine-tuned for Question Answering. It lacks the specific training to identify and score answer spans within a given context.