Gen-AI Hands On-1

NAME: Aakanksha Palavalli

SRN: PES2UG23CS925

Section : A

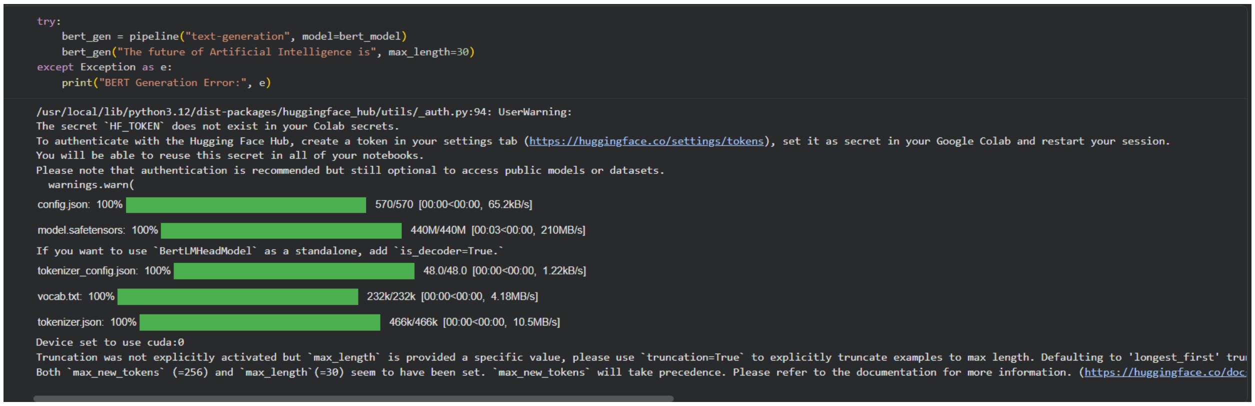
Observation table :

Task	Model	Classification	Observation	Architectural Reason
Generation	BERT	Failure	cannot generate text	Encoder -only; no autoregressive decoder
Generation	RoBERTa	Failure	cannot generate text	Encoder-only architecture
Generation	BART	Success	Generated fluent text	Encoder-Decoder with autoregressive decoding
Fill-Mask	BERT	Success	Predicted “create”, “generate”	Trained using MLM
Fill-Mask	RoBERTa	Success	Strong predictions	Optimized MLM training
Fill-Mask	BART	Partial Success	Acceptable but weaker	Not primarily MLM-trained
QA	BERT	Partial Failure	Random / incomplete answer	Not fine-tuned for QA
QA	RoBERTa	Partial Failure	Slightly better than BERT	Better pretraining but no QA fine-tuning
QA	BART	Failure	Inaccurate extraction	Seq2Seq without QA fine-tuning

EXPERIMENT 1: Text Generation

prompt:"The future of Artificial Intelligence is"

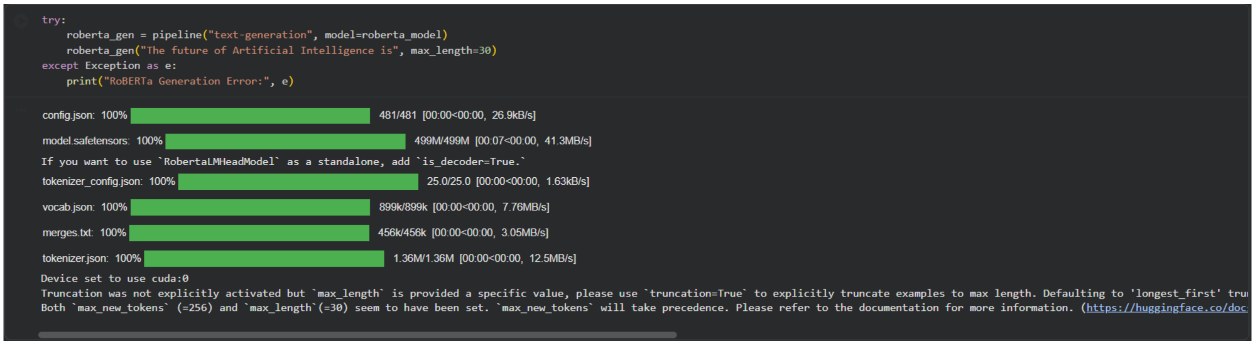
BERT



```
try:
    bert_gen = pipeline("text-generation", model=bert_model)
    bert_gen("The future of Artificial Intelligence is", max_length=30)
except Exception as e:
    print("BERT Generation Error:", e)
```

```
/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 your Google Colab and restart your session.
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%                          570/570 [00:00<00:00, 65.2kB/s]
model.safetensors: 100%                    440M/440M [00:03<00:00, 210MB/s]
If you want to use `BertLMHeadModel` as a standalone, add `is_decoder=True.`
tokenizer_config.json: 100%                48.0/48.0 [00:00<00:00, 1.22kB/s]
vocab.txt: 100%                            232k/232k [00:00<00:00, 4.18MB/s]
tokenizer.json: 100%                       466k/466k [00:00<00:00, 10.5MB/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 truncate examples to max length. Defaulting to 'longest_first' tru
Both `max_new_tokens` (=256) and `max_length`(=30) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/doc
```

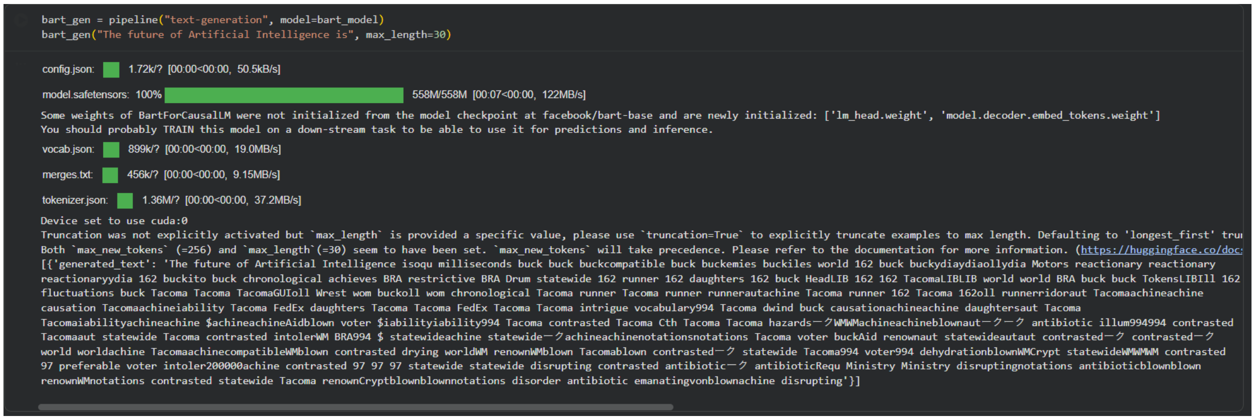
RoBERTa



```
try:
    roberta_gen = pipeline("text-generation", model=roberta_model)
    roberta_gen("The future of Artificial Intelligence is", max_length=30)
except Exception as e:
    print("RoBERTa Generation Error:", e)
```

```
config.json: 100%                          481/481 [00:00<00:00, 26.9kB/s]
model.safetensors: 100%                    499M/499M [00:07<00:00, 41.3MB/s]
If you want to use `RobertaLMHeadModel` as a standalone, add `is_decoder=True.`
tokenizer_config.json: 100%                25.0/25.0 [00:00<00:00, 1.63kB/s]
vocab.json: 100%                           899k/899k [00:00<00:00, 7.76MB/s]
merges.txt: 100%                           456k/456k [00:00<00:00, 3.05MB/s]
tokenizer.json: 100%                       1.36M/1.36M [00:00<00:00, 12.5MB/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 truncate examples to max length. Defaulting to 'longest_first' tru
Both `max_new_tokens` (=256) and `max_length`(=30) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/doc
```

BART



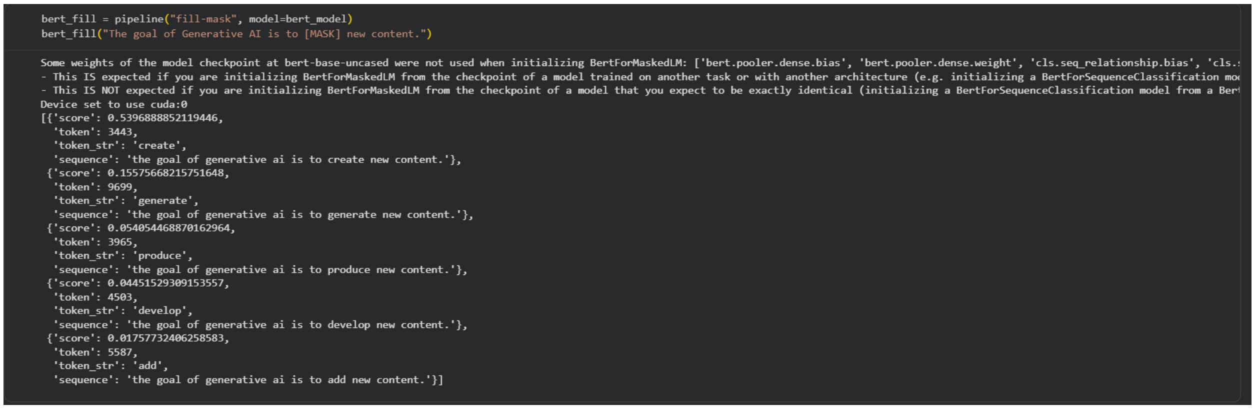
```
bart_gen = pipeline("text-generation", model=bart_model)
bart_gen("The future of Artificial Intelligence is", max_length=30)
```

```
config.json:     1.72k/? [00:00<00:00, 50.5kB/s]
model.safetensors: 100%                    558M/558M [00:07<00:00, 122MB/s]
Some weights of BartForCausalLM were not initialized from the model checkpoint at facebook/bart-base and are newly initialized: ['lm_head.weight', 'model.decoder.embed_tokens.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
vocab.json:     899k/? [00:00<00:00, 19.0MB/s]
merges.txt:     456k/? [00:00<00:00, 9.15MB/s]
tokenizer.json:     1.36M/? [00:00<00:00, 37.2MB/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 truncate examples to max length. Defaulting to 'longest_first' tru
Both `max_new_tokens` (=256) and `max_length`(=30) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/doc
[{'generated_text': 'The future of Artificial Intelligence isoqu milliseconds buck buck buckcompatible buck buckemies buckiles world 162 buck buckydiaydiaollydia Motors reactionary reactionary
reactionaryydia 162 buckito buck chronological achieves BRA restrictive BRA Drum statewide 162 runner 162 daughters 162 buck HeadLIB 162 162 TacomaLIBLIB world world BRA buck buck TokensLIBIll 162
fluctuations buck Tacoma Tacoma TacomaGUIoll Wrest wom buckoll wom chronological Tacoma runner Tacoma runner runnerautachine Tacoma runner 162 Tacoma 162oll runnerridoraut Tacomaachineachine
causation Tacomaachineiability Tacoma FedEx daughters Tacoma Tacoma FedEx Tacoma Tacoma intrigue vocabulary994 Tacoma dwind buck causationachineachine daughtersaut Tacoma
Tacomaiabilityachineachine $achineachineAidblown voter $iabilityiability994 Tacoma contrasted Tacoma Cth Tacoma Tacoma hazardsークWMWMachineachineblownautークーク antibiotic illum994994 contrasted
Tacomaaut statewide Tacoma contrasted intolerWM BRA994 $ statewideachine statewideークachineachinenotationsnotations Tacoma voter buckAid renownaut statewideautaut contrastedーク contrastedーク
world worldachine TacomaachinecompatibleWMblown contrasted drying worldWM renownWMblown Tacomablown contrastedーク statewide Tacoma994 voter994 dehydrationblownWMCrypt statewideWMWMWM contrasted
97 preferable voter intoler200000achine contrasted 97 97 97 statewide statewide disrupting contrasted antibioticーク antibioticRequ Ministry Ministry disruptingnotations antibioticblownblown
renownWMnotations contrasted statewide Tacoma renownCryptblownblownnotations disorder antibiotic emanatingvonblownachine disrupting'}]
```

EXPERIMENT 2: Masked Language Modeling

Sentence:"The goal of Generative AI is to [MASK] new content."

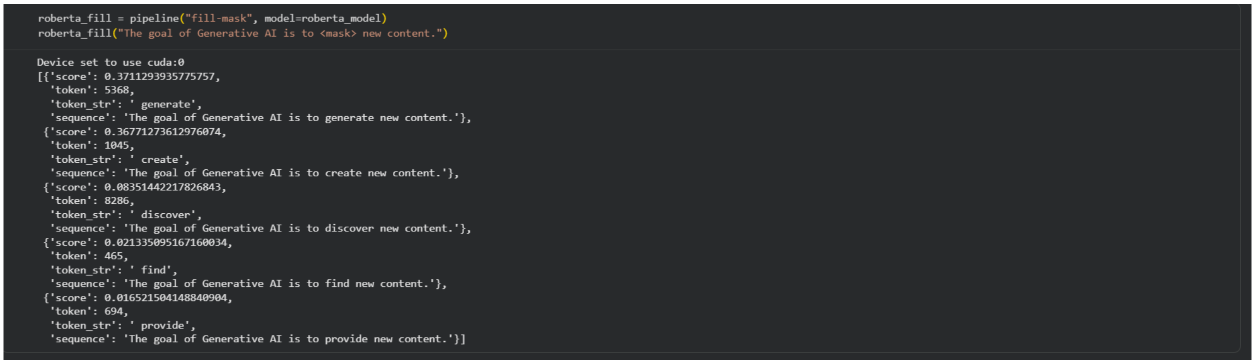
BERT



```
bert_fill = pipeline("fill-mask", model=bert_model)
bert_fill("The goal of Generative AI is to [MASK] new content.")
```

```
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.bias', 'cls.s
- 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 BertForSequenceClassification mo
- 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 model from a Ber
Device set to use cuda:0
[{'score': 0.5396888852119446,
  'token': 3443,
  'token_str': 'create',
  'sequence': 'the goal of generative ai is to create new content.'},
 {'score': 0.15575668215751648,
  'token': 9699,
  'token_str': 'generate',
  'sequence': 'the goal of generative ai is to generate new content.'},
 {'score': 0.054054468870162964,
  'token': 3965,
  'token_str': 'produce',
  'sequence': 'the goal of generative ai is to produce new content.'},
 {'score': 0.04451529309153557,
  'token': 4503,
  'token_str': 'develop',
  'sequence': 'the goal of generative ai is to develop new content.'},
 {'score': 0.01757732406258583,
  'token': 5587,
  'token_str': 'add',
  'sequence': 'the goal of generative ai is to add new content.'}]
```

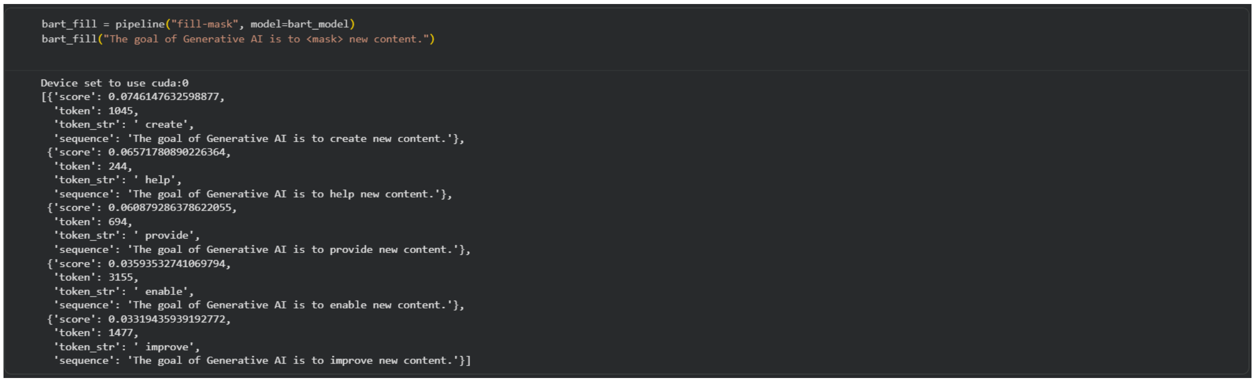
RoBERTa



```
roberta_fill = pipeline("fill-mask", model=roberta_model)
roberta_fill("The goal of Generative AI is to <mask> new content.")
```

```
Device set to use cuda:0
[{'score': 0.3711293935775757,
  'token': 5368,
  'token_str': ' generate',
  'sequence': 'The goal of Generative AI is to generate new content.'},
 {'score': 0.36771273612976074,
  'token': 1045,
  'token_str': ' create',
  'sequence': 'The goal of Generative AI is to create new content.'},
 {'score': 0.08351442217826843,
  'token': 8286,
  'token_str': ' discover',
  'sequence': 'The goal of Generative AI is to discover new content.'},
 {'score': 0.021335095167160034,
  'token': 465,
  'token_str': ' find',
  'sequence': 'The goal of Generative AI is to find new content.'},
 {'score': 0.016521504148840904,
  'token': 694,
  'token_str': ' provide',
  'sequence': 'The goal of Generative AI is to provide new content.'}]
```

BART



```
bart_fill = pipeline("fill-mask", model=bart_model)
bart_fill("The goal of Generative AI is to <mask> new content.")
```

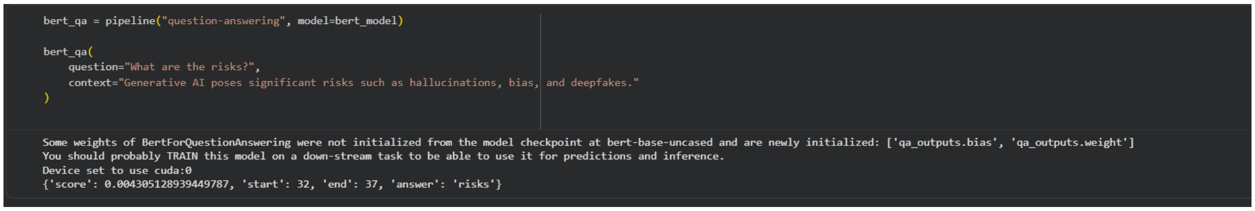
```
Device set to use cuda:0
[{'score': 0.0746147632598877,
  'token': 1045,
  'token_str': ' create',
  'sequence': 'The goal of Generative AI is to create new content.'},
 {'score': 0.06571780890226364,
  'token': 244,
  'token_str': ' help',
  'sequence': 'The goal of Generative AI is to help new content.'},
 {'score': 0.060879286378622055,
  'token': 694,
  'token_str': ' provide',
  'sequence': 'The goal of Generative AI is to provide new content.'},
 {'score': 0.03593532741069794,
  'token': 3155,
  'token_str': ' enable',
  'sequence': 'The goal of Generative AI is to enable new content.'},
 {'score': 0.03319435939192772,
  'token': 1477,
  'token_str': ' improve',
  'sequence': 'The goal of Generative AI is to improve new content.'}]
```

EXPERIMENT 3: Question Answering

Context: "Generative AI poses significant risks such as hallucinations, bias, and deepfakes."

Question: "What are the risks?

BERT

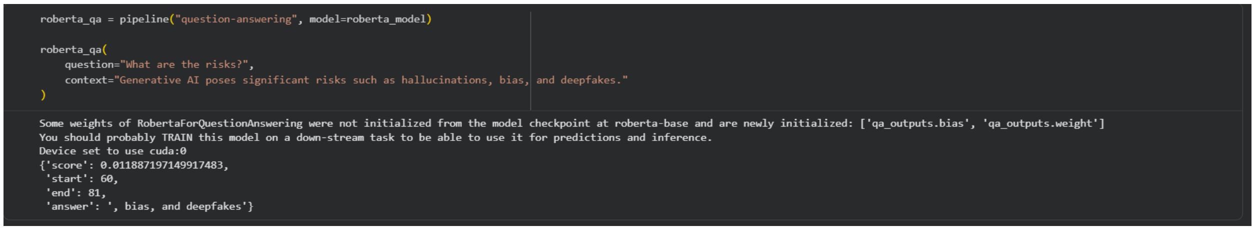


```
bert_qa = pipeline("question-answering", model=bert_model)

bert_qa(
    question="What are the risks?",
    context="Generative AI poses significant risks such as hallucinations, bias, and deepfakes."
)
```

```
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.
Device set to use cuda:0
{'score': 0.004305128939449787, 'start': 32, 'end': 37, 'answer': 'risks'}
```

RoBERTa

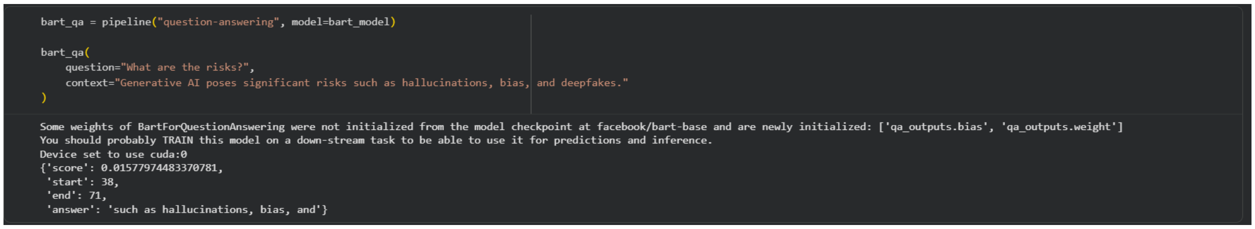


```
roberta_qa = pipeline("question-answering", model=roberta_model)

roberta_qa(
    question="What are the risks?",
    context="Generative AI poses significant risks such as hallucinations, bias, and deepfakes."
)
```

```
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.
Device set to use cuda:0
{'score': 0.011887197149917483,
 'start': 60,
 'end': 81,
 'answer': ', bias, and deepfakes'}
```

BART



```
bart_qa = pipeline("question-answering", model=bart_model)

bart_qa(
    question="What are the risks?",
    context="Generative AI poses significant risks such as hallucinations, bias, and deepfakes."
)
```

```
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.
Device set to use cuda:0
{'score': 0.015779744833370781,
 'start': 38,
 'end': 71,
 'answer': 'such as hallucinations, bias, and'}
```

Summary of Observations and Learnings

1. Text Generation
    - BERT and RoBERTa fail to generate text.
        - Reason: Both are encoder-only models, designed for understanding input rather than generating sequences. They lack an autoregressive decoder.
    - BART succeeds in generation.
        - Reason: It is an encoder-decoder (Seq2Seq) model with autoregressive decoding, making it suitable for fluent text generation.
2. Fill-Mask (Masked Word Prediction)
    - BERT and RoBERTa perform well.
        - Reason: They are pretrained with masked language modeling (MLM), which directly trains the model to predict missing words.
        - RoBERTa often gives stronger predictions due to better pretraining and optimization.
    - BART is only partially successful.
        - Reason: MLM is not its primary training objective; it is mainly a Seq2Seq model.
3. Question Answering (QA)
    - BERT and RoBERTa show partial failures.
        - Reason: They are not fine-tuned specifically for extractive QA tasks, so answers may be random or incomplete.
    - BART fails in QA extraction.
        - Reason: Seq2Seq architecture is not directly suited for extractive QA without fine-tuning.

Key Learnings

- Model architecture determines task suitability. Encoder-only models excel at understanding (classification, fill-mask) but fail at generation. Encoder-decoder models can generate text.
- Pretraining objectives matter. MLM-trained models perform well in fill-mask tasks; Seq2Seq models perform better in generation.
- Fine-tuning is essential for specialized tasks like QA; off-the-shelf pretrained models may not perform well.