Gen ai - assignment

Name: Delisha Riyona Dsouza
SRN: PES2G23CS166
Section: C

OBSERVATION TABLE

TASK	MODEL	CLASSIFICATION	OBSERVATION	ARCHITECTURAL REASON
GENERATION	BERT	FAILURE	Returned only the input prompt without generating new text.	BERT is an encoder-only model trained using masked language modeling and lacks a decoder for autoregressive generation.
GENERATION	RoBERTa	FAILURE	Echoed the input prompt without producing any continuation.	RoBERTa is also an encoder-only architecture and cannot generate tokens sequentially
GENERATION	BART	SUCCESS	Generated a continuation with new tokens, though the output was repetitive and noisy	BART has an encoder–decoder architecture with a decoder capable of autoregressive generation.
FILL MASK	BERT	SUCCESS	Correctly predicted words like “create” and “generate” with high confidence.	BERT is trained using masked language modeling (MLM).
FILL MASK	RoBERTa	SUCCESS	Accurately predicted context-aware words such as “generate” and “create”.	RoBERTa is optimized for MLM with improved training strategies.
FILL MASK	BART	PARTIAL SUCCESS	Produced reasonable predictions but with lower confidence.	BART is trained as a denoising autoencoder, not pure MLM.

OA	BERT	PARTIAL SUCCESS	Returned the correct answer but with very low confidence.	The model is not fine-tuned for question answering tasks.
OA	RoBERTa	PARTIAL SUCCESS	Returned only a partial answer span with low confidence.	Encoder-only model without QA-specific fine-tuning.
OA	BART	PARTIAL SUCCESS	Returned a partial span of the answer with low confidence.	Encoder–decoder model not fine-tuned for extractive question answering.

1. Introduction

The objective of this assignment was to understand how different Transformer architectures behave when applied to tasks they are not specifically designed for. Instead of focusing on accuracy, the goal was to analyze why certain models succeed or fail based on their underlying architecture.
For this purpose, three Transformer-based models—BERT, RoBERTa, and BART—were evaluated across text generation, masked language modeling, and question answering tasks.

2. What I Understood

From this assignment, I understood that model architecture plays a more critical role than model size in determining task performance.

- BERT and RoBERTa are encoder-only models trained primarily for understanding language using Masked Language Modeling (MLM).
- These models are excellent at predicting missing words but are not suitable for generating text because they lack a decoder and autoregressive mechanism.
- BART, on the other hand, follows an encoder–decoder architecture, enabling it to generate text sequences. However, since it is trained mainly for denoising and sequence-to-sequence tasks, its output can be noisy when forced into free-form generation.
- I also learned that base models without task-specific fine-tuning (e.g., for Question Answering) often produce low-confidence or partial answers, even if the answer text appears correct.

This experiment reinforced the importance of choosing the right architecture and fine-tuned model for a given NLP task.

3. What I Built

I implemented a benchmarking notebook using the HuggingFace Transformers pipeline to evaluate three models—BERT, RoBERTa, and BART—on three tasks:

1. Text Generation
2. Masked Language Modeling (Fill-Mask)
3. Question Answering

Each model was intentionally forced to perform tasks outside its optimal design to observe how architectural differences affect performance.

4. Key Observations

- **Text Generation**
    - BERT and RoBERTa failed to generate new text and only echoed the input prompt.
    - BART successfully generated text, though the output was repetitive and noisy.
- **Masked Language Modeling**
    - BERT and RoBERTa performed exceptionally well, predicting contextually appropriate words with high confidence.
    - BART produced reasonable predictions but with lower confidence.
- **Question Answering**
    - All three models returned partial or low-confidence answers because none of them were fine-tuned on QA datasets such as SQuAD.

These observations clearly demonstrate how architectural design and training objectives directly impact model behavior.

5. Conclusion

This assignment helped me understand that Transformer models are not interchangeable. Encoder-only models are best suited for understanding-based tasks, encoder–decoder models are necessary for generation, and reliable question answering requires task-specific fine-tuning. Overall, the benchmark highlighted the practical importance of aligning model architecture with the intended application.