Name : Mohammed Shehzaad Khan
SRN No: PES2UG23CS349
Section:`F`

Task	Model	Classification (Success/Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Generation	BERT	Failure	Output starts as fluent English but degenerates into a long sequence of repeated periods. The model does not provide a coherent continuation and stalls into punctuation-only output until the maximum length is reached.	BERT is an encoder-only model and is not trained for autoregressive next-token generation.
Generation	RoBE…	Failure	Output exactly matches the input prompt with no additional tokens generated. The model does not extend the sentence, resulting in a trivial and incomplete output.	RoBERTa is also encoder-only and lacks a decoder mechanism for sequence generation.

Generation	BART	Failure	Output begins with the prompt but quickly devolves into incoherent, noisy text with repeated and unrelated tokens, losing fluency and topical relevance.	Although BART is an encoder–decoder model, the base variant is not optimized for free-form text generation without fine-tuning.
Fill-Mask	BERT	Success	The model predicts fluent and semantically appropriate words such as “create”, “generate”, and “produce”, resulting in grammatically correct sentences.	BERT is trained using Masked Language Modeling (MLM), which directly aligns with the fill-mask task.
Fill-Mask	RoBE…	Success	The model produces fluent completions like “generate” and “create” with very close confidence scores, indicating multiple plausible options.	RoBERTa is an optimized encoder-only model trained extensively on MLM, making it well-suited for this task.

Fill-Mask	BART	Partial Success	The model fills the mask with mostly fluent words such as “create”, “provide”, and “enable”, though some completions are slightly awkward but interpretable.	BART can perform masking tasks but is not as specialized for MLM as encoder-only models.
QA	BERT	Failure	The extracted answer span (“Generative AI poses”) is fluent but incomplete and does not directly answer the question about risks.	The base BERT model is not fine-tuned for question answering and struggles to select an appropriate answer span.
QA	RoBE…	Partial Success	The model returns “deepfakes.”, which is a valid risk but only a partial answer, omitting other risks mentioned in the context.	Without QA fine-tuning, RoBERTa often selects short or narrow spans instead of comprehensive answers.
QA	BART	Success	The model returns “such as hallucinations, bias, and deepfakes”, accurately capturing the full list of risks from the context.	BART’s encoder–decoder architecture enables better span selection and contextual aggregation, even without explicit QA fine-tuning.

Objective of the Experiment:

In this lab, a controlled benchmarking experiment was conducted to study the impact of Transformer architecture on task specific performance. Three pretrained models BERT, RoBERTa, and BART were evaluated across text generation, masked language modeling, and question answering tasks. Each model was intentionally applied to tasks both aligned and misaligned with its original pretraining objective. The model outputs were systematically observed, recorded, and classified without applying any fine tuning or optimization techniques. The analysis aimed to compare encoder only and encoder decoder architectures and examine how architectural design affects model behavior, strengths, and limitations across different natural language processing tasks.