PES UNIVERSITY

Name: Mohammed Affan Khan

SRN: PES2UG23CS343

Section: F

Date: 27/1/2026

Observation Table – Model Benchmark:

Task	Model	Classification (Success/Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Generation	BERT	Failure	Generated repeated punctuation (dots) instead of meaningful text.	BERT is an encoder-only model and is not trained for autoregressive text generation.
	RoB…	Failure	Failed to generate new text and returned the input prompt unchanged.	RoBERTa is also encoder-only and lacks a decoder for token-by-token generation.
	BART	Success	Generated new tokens, but output was long, noisy, and incoherent.	BART has an encoder–decoder architecture, enabling generation, though the base model is not tuned for causal generation.
Fill-Mask	BERT	Success	Correctly predicted words such as “generate” and “create”.	BERT is trained using Masked Language Modeling (MLM).
	RoB…	Success	Produced accurate and high-confidence predictions similar to BERT.	RoBERTa improves MLM training with more data and better optimization.
	BART	Partial Suc…	Predicted reasonable words but with weaker confidence.	BART is trained for denoising sequence-to-sequence tasks, not pure MLM.
Question Answering	BERT	Partial Failure	Returned a partial answer (“and deepfakes”) with very low confidence score.	The base BERT model is not fine-tuned on QA datasets like SQuAD.
	RoB…	Partial Failure	Produced an irrelevant answer (“Generative”) with extremely low confidence.	QA head is randomly initialized without task-specific fine-tuning.

	BART	Partial Failure	Returned an incomplete answer (“such”) with low confidence score.	Base BART is designed for generation and requires QA fine-tuning for accurate extraction.