GenAI Unit-1 Benchmark Assignment

Name: Nithya Sri P. B.
SRN: PES2UG23CS399
Section: G

Output Table:

Task	Model	Classification (Failure/Success)	Observation	Architectural Reason
Generation	BERT	Failure	Generated random symbols and words.	BERT is an encoder-only model and is not trained to autoregressively generate tokens.
	RoBERTa	Failure	Returned only the input prompt without any continuation.	RoBERTa is also encoder-only and lacks a decoding mechanism.
	BART	Failure	Generated text but content was largely nonsensical and repetitive.	BART is encoder–decoder but `bart-base` is not trained for causal text generation by default.
Fill-Mask	BERT	Success	Correctly predicted meaningful words such as ‘create’, ‘generate’, and ‘produce’ with high confidence.	BERT is explicitly trained using Masked Language Modeling (MLM).
	RoBERTa	Failure	Model failed with error.	RoBERTa expects `<mask>` instead of `[MASK]`, causing token mismatch.
	BART	Partial Success	Predicted reasonable words like 'create', 'help', and 'provide' but with lower confidence scores.	BART supports denoising objectives but is not specialized for pure MLM tasks.
QA	BERT	Success	Extracted a nearly complete answer: 'risks such as hallucinations, bias, and deepfakes'.	Encoder architecture allows strong span extraction even without QA fine-tuning.

	RoBERTa	Partial Success	Returned a partial answer: 'deepfakes.' only.	Not fine-tuned on QA; span prediction was incomplete.
	BART	Success	Returned a relevant phrase: 'such as hallucinations, bias, and deepfakes'.	Encoder–decoder can model question–context relationships but lacks QA fine-tuning.