Unit 1 Assignment: The Model Benchmark Challenge

NAME : MOULYA K A
SRN : PES2UG23CS351
SECTION:6F
DATE : 27-01-2026

Task	Model	Classification	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Generation	BERT	Failure	Produced only repeated dots instead of meaningful text.	BERT is an Encoder-only model and is not trained for autoregressive next-token generation.
Generation	RoBERTa	Failure	Generation failed with a runtime error on the MPS backend.	RoBERTa is Encoder-only and not designed for text generation; backend limitations prevented execution.
Generation	BART	Success	Generated text but unstable output	Supports generation via encoder-decoder architecture but it is not optimized for free-form LM generation.
Fill-Mask	BERT	Success	Predicted “create”, “generate”, and “produce” with high confidence using [MASK].	BERT is trained using Masked Language Modeling (MLM), which exactly matches this task.
Fill-Mask	RoBERTa	Success	Predicted “generate” and “create” correctly when using <mask>.	RoBERTa is optimized for MLM and uses <mask> as its special mask token.

Fill-Mask	BART	Success	Predicted reasonable words but with lower confidence than BERT and RoBERTa.	BART supports masked prediction through denoising, but it is primarily trained for sequence-to-sequence tasks.
QA	BERT	Success	Returned “deepfakes” with very low confidence, missing other risks.	Base BERT is not fine-tuned for Question Answering; the QA head is randomly initialized.
QA	RoBERTa	Success	Extracted “risks such as hallucinations” but incomplete and very low confidence.	Without QA fine-tuning, the model cannot reliably learn correct answer spans.
QA	BART	Success	Returned a longer relevant span but still low confidence.	BART supports QA, but the base model is not trained for extractive Question Answering