Benchmark model deliverables

NAME: ANANYAA SREE SP

SRN: PES2UG23CS066

SECTION: A

Task	Model	Classification (Success/Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Generation	BERT	Failure	Generated repetitive dots instead of meaningful text.	Encoder-only model; cannot generate next tokens.
Generation	RoBERTa	Failure	Did not continue the prompt beyond the input sentence.	Encoder-only architecture with no decoder.
Generation	BART	Success	Generated long text, but output was incoherent.	Encoder–decoder architecture supports generation.
Fill-Mask	BERT	Success	Correctly predicted words like “create” and “generate”.	Trained using masked language modeling.
Fill-Mask	RoBERTa	Success	Predicted correct masked words with high confidence.	Optimized encoder trained for MLM.
Fill-Mask	BART	Partial Success	Predicted reasonable words but with lower confidence.	Masking is not its primary training objective.
QA	BERT	Partial Success	Extracted correct answer with very low confidence.	QA head not fine-tuned for question answering.
QA	RoBERTa	Failure	Returned partial and incomplete answer.	Lacks task-specific QA fine-tuning.
QA	BART	Failure	Produced incomplete answer ending mid-phrase.	Requires fine-tuning for extractive QA.

Observations

- Encoder-only models fail at text generation.
- BERT and RoBERTa excel at masked language modeling.
- All base models perform poorly at QA without fine-tuning.

Experiment 1

BERT and RoBERTa fail to generate meaningful text. Although they run without crashing, their outputs lack semantic continuation because encoder-only models are not designed for autoregressive token generation. BART is able to generate long sequences due to its encoder–decoder architecture, demonstrating that generation is architecturally possible, even though the output quality is poor without task-specific training.

Experiment 2

BERT and RoBERTa perform exceptionally well at masked language modeling, correctly predicting words such as “create” and “generate” with high confidence. This aligns with their training objective, which explicitly involves predicting masked tokens using bidirectional context. BART performs worse in comparison, as masked word prediction is not its primary training task.

Experiment 3

All three base models show inconsistent performance in question answering. Although BERT correctly extracts the list of risks, its confidence score is extremely low. RoBERTa and BART return partial or incomplete answers. This behavior is expected because the models are not fine-tuned on a question answering dataset such as SQuAD, demonstrating that task-specific fine-tuning is critical for reliable QA performance.