Submission 1 Template

Nidhi C R

PES2UG23CS382

F

Task	Model	Classification (Success / Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
Generation	BERT	Failure	Generated highly incoherent, repetitive, and meaningless text.	BERT is an encoder-only model trained for masked token prediction, not for next-token (autoregressive) generation.
Generation	RoBERTa	Failure	Returned only the input prompt and did not generate any new tokens.	RoBERTa is also encoder-only and cannot perform left-to-right text generation.
Generation	BART	Partial Success	Generated new text, but the output was noisy, repetitive, and semantically unstable.	BART is an encoder–decoder model capable of generation, but it is not optimized as a causal language model like GPT.
Fill-Mask	BERT	Success	Correctly predicted words such as “create”, “generate”, and “produce” with high confidence.	BERT is trained using Masked Language Modeling (MLM), making it well-suited for fill-mask tasks.
Fill-Mask	RoBERTa	Success	Correctly predicted “generate” and “create” with strong and balanced confidence scores.	RoBERTa improves MLM through dynamic masking, leading to better contextual understanding.
Fill-Mask	BART	Partial Success	Predicted reasonable words but with much lower confidence and weaker precision.	BART is trained for sequence denoising rather than explicit masked token prediction.
QA	BERT	Partial Success	Correctly extracted “hallucinations, bias, and deepfakes” but with very low confidence.	The model is not fine-tuned for extractive QA; its QA head was randomly initialized.

Task	Model	Classification (Success / Failure)	Observation (What actually happened?)	Why did this happen? (Architectural Reason)
QA	RoBERTa	Partial Success	Extracted a partial answer (“such as hallucinations, bias”) with very low confidence.	Lack of QA fine-tuning limits accurate answer span selection.
QA	BART	Failure	Returned an incomplete leading phrase instead of the actual risks.	Encoder–decoder models like BART are poorly suited for extractive QA without fine-tuning.