# Assignment 6
# Modern Transformer Variants & BERT

Part B — Transformers & LLMs

18 questions

★ **Must-know** — asked very frequently in GenAI / ML interviews (6 questions). Prepare these first.

1. Give examples of encoder-only, decoder-only and encoder–decoder models. Which tasks suit each? ★ · Asked: TCS
2. Why are most modern LLMs decoder-only?
3. Explain Multi-Head Attention (MHA), Multi-Query Attention (MQA) and Grouped-Query Attention (GQA). ★
4. Which of MHA, MQA and GQA uses the least memory, and which gives the best quality? What is the trade-off?
5. What is the KV cache and why is it needed? How does it affect memory at long context lengths? ★
6. What is Flash Attention and why is it faster without changing the result?
7. What is sliding-window attention? How does it still capture global context?
8. What are the types of positional embeddings (absolute, learned, relative, RoPE, ALiBi)? Absolute vs relative? ★
9. How does RoPE help models extend to longer context windows?
10. What is a Mixture-of-Experts (MoE) model and why is it popular?
11. Advantages and disadvantages of the Transformer architecture.
12. What is a language model? Compare autoregressive vs masked language modelling.
13. What is BERT? How is it bidirectional? ★ · Asked: NLP interview rounds
14. Which tasks are used for BERT pre-training (MLM and NSP)? Why did RoBERTa drop NSP? ★
15. Cased vs uncased BERT — when would you use each?
16. Which tokenizer does BERT use? What do [PAD], [UNK], [CLS], [SEP] and [MASK] represent?
17. How do you fine-tune BERT for classification, NER and question answering?
18. NLP case study: you must classify customer complaints into 10 categories. Walk through your approach from baseline to the final model, including class imbalance. ·

Asked: NLP interview rounds