Assignment 5
Attention & Transformer Architecture

Part B — Transformers & LLMs

21 questions

★	**Must-know** — asked very frequently in GenAI / ML interviews (9 questions). Prepare these first.
[CODE]	**Hands-on task** — write and run the code (1 tasks).

1. What is a seq2seq model? What is the encoder–decoder architecture and how does it work?
2. Applications, pros and cons of the encoder–decoder architecture.
3. What is attention and why was it needed (the bottleneck problem)? ★
4. What are the different types of attention (Bahdanau/additive, Luong/multiplicative, self, cross)?
5. What is self-attention? Explain Query, Key and Value with an intuitive example. ★ · Asked: Google, NLP rounds
6. What is Scaled Dot-Product Attention? Write the formula. Why do we divide by $\sqrt{d_k}$? ★
7. What is multi-head attention and why is it better than a single head? ★
8. What is the difference between self-attention and cross-attention? How does cross-attention work in the decoder?
9. How do Transformers account for word order? Explain positional encoding using sine and cosine. ★ · Asked: NLP interview rounds
10. If the input embedding dimension is 1024, what is the dimension of the positional encoding vector? Why?
11. What are residual (skip) connections and why are they needed?
12. What is layer normalization and why is it used in Transformers? Pre-LN vs Post-LN?
13. What is the role of the feed-forward network in each Transformer block?
14. What is masked (causal) attention and why is it used in the decoder? ★
15. What is the role of the final linear + softmax layer in the decoder?
16. What is teacher forcing and why is it significant? What is exposure bias?
17. Explain the complete Transformer architecture end-to-end (draw it). ★ · Asked: TCS, Google

18. What is the time and memory complexity of self-attention with respect to sequence length? ★

19. Advantages of Transformers over traditional seq2seq (RNN) models.

20. Implement scaled dot-product attention from scratch in NumPy or PyTorch. [CODE]

21. How do LLMs understand words with multiple meanings, e.g. "bank" of a river vs a financial "bank"? ★ · Asked: EPAM