**Q.** Suppose that we design a deep architecture to represent a sequence by stacking self-attention layers with positional encoding. What could be issues? (paragraph format)

**Ans.** There are many challenges that can arise when creating stacked self-attention layers with positional encoding. Perhaps the most significant challenge is the computational complexity of such a system, which scales with the length of the sequence and becomes very slow for longer sequences. The issue of ‘gradient vanishing’ can also occur in this setting because, in a deep network, the positional embeddings may diminish as they propagate through the layers, rendering them irrelevant. Additionally, the model might overfit to previously learned positional patterns in the training data, which could limit its ability to generalize to unseen data. Furthermore, fixed positional encodings work well only for sequences where the relationship is derived from the previous token. To handle more complex structures, a more sophisticated positional encoding mechanism would be necessary, inevitably increasing the network’s complexity. This often requires incorporating domain-specific knowledge, such as 3D spatial relationships or hierarchical dependencies, inevitably increasing the network’s complexity and computational demand.