1. 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)

One issue we may encounter is interpretability. The increased complexity of our model makes it harder to interpret, especially in fields like medicine where causal inference is vital. In medicine, decisions have significant consequences on health outcomes, and transparency is required in the reasoning process. Simpler models such as decision trees or linear models are often preferred by medical professionals for their ease of interpretability and ability to support causal inferences. Another challenge we may face is computational difficulties due to the depth of our architecture. Each additional self-attention layer will create more complexity and computational overhead. This is especially true if we have a very large sequence. If the dataset we are using for training is large, this could be even more of a problem. However, self-attention layers with positional encoding can offer faster processing compared to for example LSTMs, which handle embeddings sequentially (instead of all positions simultaneously like in our system). Overfitting and lack of generalizability are additional issues that can arise in complex deep architectures, particularly with limited training data. Deep models may run the risk of memorizing training examples instead of learning meaningful generalizations. Thus, when presented with novel data or problems, the model may struggle to perform as intended. Additionally, the positional encoding scheme itself can be an issue. Traditional positional encoding methods such as sine and cosine functions are fixed and not learnable. These fixed encodings may not always capture the complex patterns in the data. Therefore, the model might struggle to generalize to unseen sequences or have limited flexibility in accommodating different types of input sequences.