1. **What are the pros and cons of using a stateful RNN versus a stateless RNN?**

* Setting an RNN to be stateful means that it can build a state across its training sequence and even maintain that state when doing predictions. The benefits of using stateful RNNs are **smaller network sizes and/or lower training times**.
* **Cons :**
* The computation of this neural network is slow.
* Training can be difficult.
* If you are using the activation functions, then it becomes very tedious to process long sequences.
* It faces issues like Exploding or Gradient Vanishing.

1. **Why do people use Encoder–Decoder RNNs rather than plain sequence-to-sequence RNNs for automatic translation?**

* This two-step model, called an Encoder–Decoder, works much better than trying to translate on the fly with a single sequence-to-sequence RNN (like the one represented on the top left), **since the last words of a sentence can affect the first words of the translation, so you need to wait until you have heard the whole.**

1. **How can you deal with variable-length input sequences?**

* The first and simplest way of handling variable length input is to **set a special mask value in the dataset, and pad out the length of each input to the standard length with this mask value set for all additional entries created**. Then, create a Masking layer in the model, placed ahead of all downstream layers.

**What about variable-length output sequences?**

* The most common way people deal with inputs of varying length is padding. You first define the desired sequence length, i.e. the input length you want your model to have. Then any sequences with a shorter length than this are padded either with zeros or with special characters so that they reach the desired length

1. **What is beam search and why would you use it?**

* Beam search is **the most popular search strategy for the sequence to sequence Deep NLP algorithms like Neural Machine Translation, Image captioning, Chatbots, etc**. Beam search considers multiple best options based on beamwidth using conditional probability, which is better than the sub-optimal Greedy search.

**What tool can you use to implement it?**

* **Beam Search Decoder**

1. **What is an attention mechanism? How does it help?**

* The idea behind the attention mechanism was to permit the decoder to utilize the most relevant parts of the input sequence in a flexible manner, by a weighted combination of all the encoded input vectors, with the most relevant vectors being attributed the highest weights.
* In the brain, attention mechanisms **allow to focus on one part of the input or memory (image, text, etc) while giving less attention to others, thus guiding the process of reasoning**. Attention mechanisms have provided and will provide a paradigm shift in machine learning

1. What is the most important layer in the Transformer architecture? What is its purpose?

* The most important part here is the “Residual Connections” around the layers. This is very important in **retaining the position related information which we are adding to the input representation/embedding across the network**. The network displayed catastrophic results on removing the Residual Connections

1. **When would you need to use sampled softmax?**

* I'd probably consider using sampled softmax **if I have over 100,000 classes, or if my final classification layer dominates overall execution time or memory use**. An obvious application is large word vocabularies, for example in language modelling.