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

**Stateful RNN vs. Stateless RNN:**

* **Stateful RNN Pros:**
  + Maintains hidden state between batches, which can be important for certain tasks like time series forecasting.
  + Allows the model to learn dependencies across sequences that span batch boundaries.
* **Stateful RNN Cons:**
  + Requires careful handling of batch size and sequence length.
  + Not suitable for all tasks; stateful RNNs are generally more complex to set up and train.

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

**Encoder-Decoder RNNs vs. Sequence-to-Sequence RNNs:**

* **Encoder-Decoder RNN Pros:**
  + Well-suited for tasks like machine translation, where the input and output sequences have different lengths.
  + Allows for the encoding of the input sequence into a fixed-length context vector, which is then used by the decoder.
* **Encoder-Decoder RNN Cons:**
  + More complex architecture compared to simple sequence-to-sequence models.
  + Requires careful design and tuning.

1. How can you deal with variable-length input sequences? What about variable-length output sequences?

**Variable-Length Sequences:**

* **Input Sequences:** You can use techniques like padding or masking to handle variable-length input sequences. Padding adds zeros to shorter sequences, making them of equal length. Masking ensures that RNN cells do not process padded elements.
* **Output Sequences:** For variable-length output sequences, you can use techniques like teacher forcing (feeding ground-truth outputs as inputs during training) and dynamic decoding during inference.

1. What is beam search and why would you use it? What tool can you use to implement it?

 **Beam Search:**

* Beam search is a search algorithm used in sequence generation tasks, such as machine translation and text generation.
* It explores multiple possible sequences in parallel, maintaining a "beam" of the most likely candidates.
* Beam search helps find more diverse and higher-quality sequences compared to greedy decoding.
* You can implement beam search using libraries like TensorFlow or PyTorch.

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

**Attention Mechanism:**

* An attention mechanism allows a model to focus on different parts of the input sequence when generating the output sequence.
* It helps improve the model's ability to capture long-range dependencies and handle variable-length sequences.
* The most common form is the "self-attention" mechanism used in the Transformer architecture.

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

**Transformer Architecture:**

* The most important layer in the Transformer architecture is the "multi-head self-attention" layer.
* Its purpose is to capture relationships between words in a sequence, enabling the model to weigh the importance of different words when making predictions.
* The self-attention layer plays a central role in the model's ability to handle sequential data.

1. When would you need to use sampled softmax?

**Sampled Softmax:**

* Sampled softmax is used in large-vocabulary scenarios in natural language processing (NLP).
* It's used when the full softmax computation is computationally expensive due to a large number of possible output classes.
* Sampled softmax approximates the full softmax by sampling a subset of output classes, making it more efficient for training.